GOVERNMENT OF TAMIL NADU

HIGHER SECONDARY SECOND YEAR

STATISTICS

A publication under Free Textbook Programme of Government of Tamil Nadu

Department Of School Education

Untouchability is Inhuman and a Crime

Government of Tamil Nadu

First Edition - 2019

(Published under New Syllabus)

NOT FOR SALE

Content Creation

State Council of Educational
Research and Training

Printing & Publishing

Tamil NaduTextbook and Educational
Services Corporation

www.textbooksonline.tn.nic.in

CONTENTS

STATISTICS

3GJFMY
E-book

4QV4QA
Assessment

B273_12_STAT_EM
DIGI links

Lets use the QR code in the text books ! How ?
- Download the QR code scanner from the Google PlayStore/ Apple App Store into your smartphone
- Open the QR code scanner application
- Once the scanner button in the application is clicked, camera opens and then bring it closer to the QR code in the text book.
- Once the camera detects the QR code, a url appears in the screen.Click the url and goto the content page.

KEY FEATURES OF THE BOOK

Profile of a Statistician	Presents a brief history and contribution of a statistician
Learning Objectives	Goals to transform the classroom processes a learner centric
DO YOU KNOW?	Amazing facts, Rhetorical questions to lead students to Statistical inquiry
Note	Additional inputs to content is provided
Activity	Directions are provided to students to conduct activities in order to explore, enrich the concept
Infographics	Visual representation of the lesson to enrich learning
	To motivate the students to further explore the content digitally and take them to virtual world
Success Story	Success Stories given as a source of inspiration
Points to Remember	Summary of each lesson is given at the end
ICT	To enhance digital skills among students
Evaluation	Assess students to pause, think and check their understanding
Glossary	Explanation of scientific terms

Career in Statistics

After completion of Higher Secondary Course, the subject Statistics is an essential part of the curriculum of many undergraduate, postgraduate, professional courses and research level studies. At least one or more papers are included in the Syllabus of the following courses:

Under Graduate Courses	Post Graduate Courses	Competitive Eaminations
B.A.(Economics)	M.A.(Economics)	UPSC
B.Com	M.Com	TNPSC
B.B.A	M.B.A	Staff Selection Commission Examinations
B.C.A	M.C.A	I.A.S
B.Sc.(Maths)	M.Sc	I.F.S
B.Pharm	M.Pharm	and many more
B.Ed	M.Ed	
B.Stat	M.Stat	
B.E	M.E	
Diploma Courses	C.A	
	I.C.W.A	
	Actuarial science	

Specialized fields in Statistics : Colleges/universities, Indian Statistical Institute(ISI) offer a number of specialisations in statistics at undergraduate, postgraduate and research level. A candidate with bachelor's degree in statistics can also apply for Indian Statistical Services (ISS).

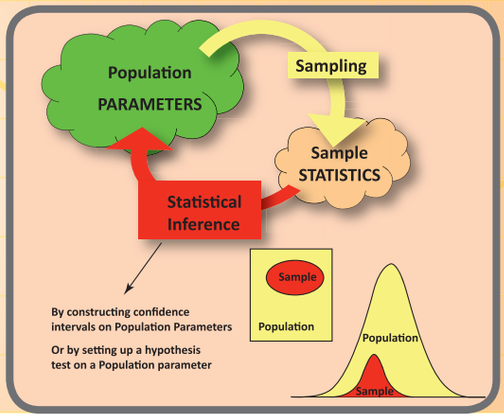
Job Titles	Job Areas
• Statisticians	• Census
• Business Analyst	• Ecology
• Mathematician	• Medicine
• Professor	• Election
• Risk Analyst	• Crime
• Data Analyst	• Economics
• Content Analyst	• Education
• Statistics Trainer	• Film
• Data Scientist	• Sports
• Consultant	• Tourism
• Biostatistician	
• Econometrician	

Skills Required for a statistician

- Strong Foundation in Mathematical Statistics
- Logical Thinking & Ability to Comprehend Key Facts
- Ability to Interact with people from various fields to understand the problems
- Strong Background in Statistical Computing
- Ability to stay updated on recent literature & statistical software
- Versatility in solving problems

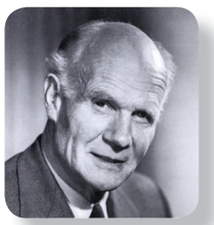
CHAPTER

1 TESTS OF SIGNIFICANCE – BASIC CONCEPTS AND LARGE SAMPLE TESTS



Jerzy Neyman

Jerzy Neyman (1894-1981) was born into a Polish family in Russia. He is one of the Principal architects of Modern Statistics. He developed the idea of confidence interval estimation during 1937. He had also contributed to other branches of Statistics, which include Design of Experiments, Theory of Sampling and Contagious Distributions. He established the Department of Statistics in University of California at Berkeley, which is one of the preeminent centres for statistical research worldwide.



Egon Sharpe Pearson

Egon Sharpe Pearson (1885-1980) was the son of Prof. Karl Pearson. He was the Editor of *Biometrika*, which is still one of the premier journals in Statistics. He was instrumental in publishing the two volumes of *Biometrika Tables for Statisticians*, which has been a significant contribution to the world of Statistical Data Analysis till the invention of modern computing facilities.

Neyman and Pearson worked together about a decade from 1928 to 1938 and developed the theory of testing statistical hypotheses. *Neyman-Pearson Fundamental Lemma* is a milestone work, which forms the basis for the present theory of testing statistical hypotheses. In spite of severe criticisms for their theory, in those days, by the leading authorities especially Prof.R.A.Fisher, their theory survived and is currently in use.

"Statistics is the servant to all sciences" – *Jerzy Neyman*

LEARNING OBJECTIVES

The students will be able to

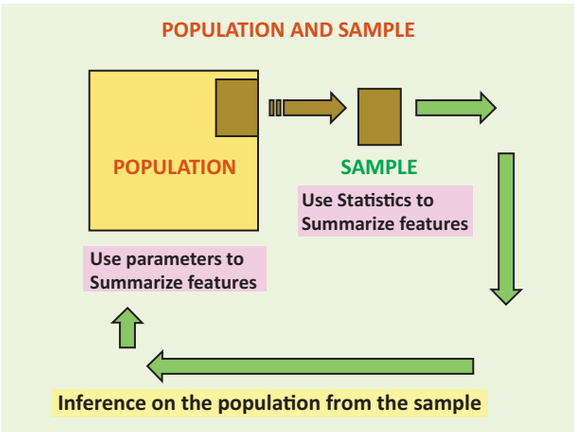
- understand the purpose of hypothesis testing;
- define parameter and statistic;
- understand sampling distribution of statistic;
- define standard error;
- understand different types of hypotheses;
- determine type I and type II errors in hypotheses testing problems;
- understand level of significance, critical region and critical values;
- categorize one-sided and two-sided tests;
- understand the procedure for tests of hypotheses based on large samples; and
- solve the problems of testing hypotheses concerning mean(s) and proportion(s) based on large samples.

6FPMTX

Introduction

In XI Standard classes, we concentrated on collection, presentation and analysis of data along with calculation of various measures of central tendency and measures of dispersion. These kinds of describing the data are popularly known as **descriptive statistics**. Now, we need to understand another dimension of statistical data analysis, which is called **inferential statistics**. Various concepts and methods related to this dimension will be discussed in the first four Chapters of this volume. Inferential Statistics may be described as follows from the statistical point of view:

One of the main objectives of any scientific investigation or any survey is to find out the unknown facts or characteristics of the population under consideration. It is practically not feasible to examine the entire population, since it will increase the time and cost involved. But one may examine a part of it, called **sample**. On the basis of this limited information, one can make decisions or draw inferences on the unknown facts or characteristics of the population.



Thus, inferential statistics refers to a collection of statistical methods in which random samples are used to draw valid inferences or to make decisions in terms of probabilistic statements about the population under study.

Before going to study in detail about Inferential Statistics, we need to understand some of the important terms and definitions related to this topic.

1.1 PARAMETER AND STATISTIC

A **population**, as described in Section 2.4 in XI Standard text book, is a collection of units/objects/numbers under study, whose elements can be considered as the values of a random variable, say, X. As mentioned in Section 9.3 in XI Standard text book, there will be a probability distribution associated with X.

Parameter: Generally, **parameter** is a quantitative characteristic, which indexes/identifies the respective distribution. In many cases, statistical quantitative characteristics calculated based on all the units in the population are the respective parameters. For example, population mean, population standard deviation, population proportion are parameters for some distributions.

Recall: The unknown constants which appear in the *probability density function or probability mass function* of the random variable X, are also called **parameters** of the corresponding distribution/population.

The parameters are commonly denoted by Greek letters. In Statistical Inference, some or all the parameters of a population are assumed to be unknown.

Random sample: Any set of reliazations $(X_1, X_2, ..., X_n)$ made on X under independent and identical conditions is called a **random sample**.

Statistic: Any statistical quantity calculated on the basis of the random sample is called a **statistic**. The sample mean, sample standard deviation, sample proportion *etc*., are called **statistics** (plural form of *statistic*).They will be denoted by Roman letters.

Let $(x_1, x_2, \ldots, x_n)$ be an observed value of $(X_1, X_2, ..., X_n)$. The collection of $(x_1, x_2, \ldots, x_n)$ is known as *sample space*, which will be denoted by '**S**'.

Note 1:

A set of n sample observations can be made on X, say, $x_1, x_2, \ldots, x_n$ for making inferences on the unknown parameters. It is to be noted that these n values may vary from sample to sample. Thus, these values can be considered as the realizations of the random variables $X_1, X_2, ..., X_n$, which are assumed to be independent and have the same distribution as that of X. These are also called independently and identically distributed (*iid)* random variables.

NOTE

The *statistic* itself is a random variable and has a probability distribution.

Note 2:

In Statistical Inference, the sample standard deviation is defined as $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It may be noted that the divisor is $n - 1$ instead of n.

Note 3:

The statistic itself is a random variable, until the numerical values of $X_1, X_2, ..., X_n$ are observed, and hence it has a probability distribution.

Notations to denote various population parameters and their corresponding sample statistics are listed in Table 1.1. The notations will be used in the first four chapters of this book with the same meaning for the sake of uniformity.

Table 1.1 Notations for Parameters and Statistics

Statistical measure	Parameter	Statistic	Value of the Statistic for a given sample
Mean	μ	$\bar{X}$	$\bar{x}$
Standard deviation	σ	S	s
Proportion	P	p	p_0

1.2 SAMPLING DISTRIBUTION

The probability distribution of a statistic is called **sampling distribution** of the statistic. In other words, it is the probability distribution of possible values of the statistic, whose values are computed from possible random samples of same size.

The following example will help to understand this concept.

Example 1.1

Suppose that a population consists of 4 elements such as 4, 8, 12 and 16. These may be considered as the values of a random variable, say, X. Let a random sample of size 2 be drawn from this population under *sampling with replacement* scheme. Then, the possible number of samples is 4^2.

It is to be noted that, if we take samples of size n each from a finite population of size N, then the number of samples will be N^n under with replacement scheme and ${}^{N}C_n$ samples under without replacement scheme.

In each of the 4^2 samples, the sample elements x_1 and x_2 can be considered as the values of the two *iid* random variables X_1 and X_2. The possible samples, which could be drawn from the above population and their respective means are presented in Table 1.2.

Table 1.2 Possible Samples and their Means

Sample Number	Sample elements (x_1, x_2)	Sample Mean $\bar{x}$
1	4,4	4
2	4,8	6
3	4,12	8
4	4,16	10
5	8,4	6
6	8,8	8
7	8,12	10
8	8,16	12
9	12,4	8
10	12,8	10
11	12,12	12
12	12,16	14
13	16,4	10
14	16,8	12
15	16,12	14
16	16,16	16

The set of pairs (x_1, x_2) listed in column 2 constitute the sample space of samples of size 2 each.

Hence, the sample space is:

$\mathbf{S}$ = {(4,4), (4,8), (4,12), (4,16), (8,4), (8,8), (8,12), (8,16), (12,4), (12,8), (12,12), (12,16), (16,4), (16,8), (16,12), (16,16)}

The sampling distribution of $\overline{X}$, the sample mean, is determined and is presented in Table 1.3.

Table 1.3 Sampling Distribution of Sample Mean

Sample mean: $\bar{x}$	4	6	8	10	12	14	16	Total
Probability: $P(\bar{X}=\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

Note 4: The sample obtained under sampling *with replacement* from a finite population satisfies the conditions for a random sample as described earlier.

Note 5: If the sample values are selected under *without replacement scheme*, independence property of $X_1, X_2, \ldots X_n$ will be violated. Hence it will not be a random sample.

Note 6: When the sample size is greater than or equal to 30, in most of the text books, the sample is termed as a **large sample**. Also, the sample of size less than 30 is termed as **small sample**. However, in practice, there is no rigidity in this number *i.e.*, 30, and that depends on the nature of the population and the sample.

Note 7: The learners may recall from XI Standard Textbook that some of the probability distributions possess the additive property. For example, if $X_1, X_2, \ldots, X_n$ are *iid* $N(\mu, \sigma^2)$ random variables, then the probability distributions of $X_1 + X_2 + \ldots + X_n$ and $\overline{X}$ are respectively the $N(n\mu, n\sigma^2)$ and $N(\mu, \sigma^2/n)$. These two distributions, in statistical inference point of view, can be considered respectively as the sampling distributions of the sample total and sample mean of a random sample drawn from the $N(\mu, \sigma^2)$ distribution. The notation $N(\mu, \sigma^2)$ refers to the normal distribution having mean μ and variance σ^2.

1.3 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is defined as the **standard error** of the statistic, which is abbreviated as *SE*.

For example, the standard deviation of the sampling distribution of the sample mean, $\bar{x}$, is known as the standard error of the sample mean, or $SE\ (\overline{X})$.

If the random variables $X_1, X_2, \ldots, X_n$ are independent and have the same distribution with mean μ *and variance* σ^2, then variance of $\overline{X}$ becomes as

$$V\left(\overline{X}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus, $SE\left(\overline{X}\right) = \dfrac{\sigma}{\sqrt{n}}$.

Also, note that mean of $\overline{X} = E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$

Example 1.2

Calculate the standard error of $\overline{X}$ for the sampling distribution obtained in *Example 1.*

Solution:

Here, the population is {4, 8, 12, 16}.

Population size (N) = 4, Sample size (n) = 2

Population mean (μ) = (4 + 8 + 12+ 16)/4 = 40/4 = 10

The population variance is calculated as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

$$= \frac{1}{4}\left[(4-10)^2 + (8-10)^2 + (12-10)^2 + (16-10)^2\right] = \frac{1}{4}\left[36+4+4+36\right] = 20.$$

Hence, $SE(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{20}{2}} = \sqrt{10}$

This can also be verified from the sampling distribution of $\bar{X}$ (*see* Table 1.3)

$$V(\bar{X}) = \sum(\bar{x} - \mu)^2 P(\bar{X} = \bar{x})$$

where the summation is taken over all values of $\bar{x}$

Thus, $V(\bar{X}) = (4-10)^2\frac{1}{16} + (6-10)^2\frac{2}{16} + (8-10)^2\frac{3}{16} + (10-10)^2\frac{4}{16}$

$$+(12-10)^2\frac{3}{16} + (14-10)^2\frac{2}{16} + (16-10)^2\frac{1}{16}$$

$$= \frac{1}{16}(36+32+12+0+12+32+36) = 10$$

Hence, the standard deviation of the sampling distribution of $\bar{X}$ is = $\sqrt{10}$.

Standard Errors of some of the frequently referred statistics are listed in Table 1.4.

Table 1.4 Statistics and their Standard Errors

Statistic	Standard error
Sample proportion: p	$\sqrt{\frac{PQ}{n}}$, where P is the population proportion and $Q = 1 - P$.
Difference between the means $\bar{X}$ and $\bar{Y}$ of two independent samples: $(\bar{X} - \bar{Y})$	$\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$ where m and n are the sizes of samples drawn from the populations whose variances are σ_X^2 and σ_Y^2 respectively.
	$\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}$, where σ^2 is the common variance of the populations.
Difference between the proportions p_X and p_Y of two independent samples: $(p_X - p_Y)$	$\sqrt{\frac{P_X Q_X}{m} + \frac{P_Y Q_Y}{n}}$, where m and n are sizes of the samples drawn from the populations whose proportions are respectively P_X and P_Y; $Q_X = 1 - P_X$, $Q_Y = 1 - P_Y$.
	$\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}$, where $\hat{p} = \frac{mp_X + np_Y}{m+n}$, $\hat{q} = 1 - \hat{p}$, m and n are sample sizes, when P_X and P_Y are unknown.

1.4 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

In many practical studies, as mentioned earlier, it is necessary to make decisions about a population or its unknown characteristics on the basis of sample observations. For example, in bio-medical studies, we may be investigating a particular theory that the recently developed medicine is much better than the conventional medicine in curing a disease. For this purpose, we propose a statement on the population or the theory. Such statements are called hypotheses.

Thus, a **hypothesis** can be defined as a statement on the population or the values of the unknown parameters associated with the respective probability distribution. All the hypotheses should be tested for their validity using statistical concepts and a representative sample drawn from the study population. '*Hypotheses*' is the plural form of '*hypothesis*'.

A **statistical test** is a procedure governed by certain determined/derived rules, which lead to take a decision about the null hypothesis for its rejection or otherwise on the basis of sample values. This process is called **statistical hypotheses testing**.

The statistical hypotheses testing plays an important role, among others, in various fields including industry, biological sciences, behavioral sciences and Economics. In each hypotheses testing problem, we will often find as there are two hypotheses to choose between *viz*., null hypothesis and alternative hypothesis.

Null Hypothesis:

A hypothesis which is to be actually tested ***for possible rejection*** based on a random sample is termed as **null hypothesis**, which will be denoted by H_0.

YOU WILL KNOW

(i) Generally, it is a hypothesis of no difference in the case of comparison.

(ii) Assigning a value to the unknown parameter in the case of single sample problems

(iii) Suggesting a suitable model to the given environment in the case of model construction.

(iv) The given two attributes are independent in the case of *Chi*-square test for independence of attributes.

Alternative Hypothesis:

A statement about the population, which contradicts the null hypothesis, depending upon the situation, is called **alternative hypothesis**, which will be denoted by H_1.

For example, if we test whether the population mean has a specified value μ_0, then the null hypothesis would be expressed as:

$$H_0: \mu = \mu_0$$

The alternative hypothesis may be formulated suitably as anyone of the following:

(i) $H_1: \mu \neq \mu_0$

(ii) $H_1: \mu > \mu_0$

(iii) $H_1: \mu < \mu_0$

The alternative hypothesis in (i) is known as two-sided alternative and the alternative hypothesis in (ii) is known as one-sided (right) alternative and (iii) is known as one-sided (left) alternative.

1.5 ERRORS IN STATISTICAL HYPOTHESES TESTING

A **statistical decision** in a hypotheses testing problem is either of rejecting or not rejecting H_0 based on a given random sample. Statistical decisions are governed by certain rules, developed applying a statistical theory, which are known as **decision rules**. The decision rule leading to rejection of H_0 is called as **rejection rule**.

The null hypothesis may be either true or false, in reality. Under this circumstance, there will arise four possible situations in each hypotheses testing or decision making problem as displayed in Table 1.5.

Table 1.5 Decision Table

	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct decision
Do not Reject H_0	Correct decision	Type II error

It must be recognized that the final decision of rejecting H_0 or not rejecting H_0 may be incorrect. The error committed by rejecting H_0, when H_0 is really true, is called **type I error**. The error committed by not rejecting H_0, when H_0 is false, is called **type II error**.

Example 1.3

A soft drink manufacturing company makes a new kind of soft drink. Daily sales of the new soft drink, in a city, is assumed to be distributed with mean sales of ₹40,000 and standard deviation of ₹2,500 per day. The Advertising Manager of the company considers placing advertisements in local TV Channels. He does this on 10 random days and tests to see whether or not sales has increased. Formulate suitable null and alternative hypotheses. What would be type I and type II errors?

Solution:

The Advertising Manager is testing whether or not sales increased more than ₹40,000. Let μ be the average amount of sales, if the advertisement does appear.

The null and alternative hypotheses can be framed based on the given information as follows:

Null hypothesis: H_o: $\mu = 40000$

i.e., The mean sales due to the advertisement is not significantly different from ₹40,000.

Alternative hypothesis: H_1: $\mu > 40000$

i.e., Increase in the mean sales due to the advertisement is significant.

(i) If type I error occurs, then it will be concluded as the advertisement has improved sales. But, really it is not.

(ii) If type II error occurs, then it will be concluded that the advertisement has not improved the sales. But, really, the advertisement has improved the sales.

The following may be the penalties due to the occurrence of these errors:

If type I error occurs, then the company may spend towards advertisement. It may increase the expenditure of the company. On the other hand, if type II error occurs, then the company will not spend towards advertisement. It may not improve the sales of the company.

1.6 LEVEL OF SIGNIFICANCE, CRITICAL REGION AND CRITICAL VALUE(S)

In a given hypotheses testing problem, the *maximum probability* with which we would be willing to tolerate the occurrence of type I error is called **level of significance** of the test. This probability is usually denoted by 'α'. Level of significance is specified before samples are drawn to test the hypothesis.

The level of significance normally chosen in every hypotheses testing problem is 0.05 (5%) or 0.01 (1%). If, for example, the level of significance is chosen as 5%, then it means that among the 100 decisions of rejecting the null hypothesis based on 100 random samples, maximum of 5 of among them would be wrong. It is emphasized that the 100 random samples are drawn under identical and independent conditions. That is, the null hypothesis H_0 is rejected wrongly based on 5% samples when H_0 is actually true. We are about 95% confident that we made the right decision of rejecting H_0.

Critical region in a hypotheses testing problem is a subset of the sample space whose elements lead to rejection of H_0. Hence, its elements have the dimension as that of the sample size, say, $n(n > 1)$. That is,

$$\text{Critical Region} = \left\{\underset{\sim}{x} = (x_1, x_2, ..., x_n) \,|\, H_0 \text{ is rejected}\right\}.$$

A subset of the sample space whose elements does not lead to rejection of H_0 may be termed as **acceptance region**, which is the complement of the critical region. Thus,

$$\mathbf{S} = \{\text{Critical Region}\} \cup \{\text{Acceptance Region}\}.$$

Test statistic, a function of statistic(s) and the known value(s) of the underlying parameter(s), is used to make decision on H_0. Consider a hypotheses testing problem, which uses a **test statistic** $t(\underset{\sim}{X})$ and a constant c for deciding on H_0. Suppose that H_0 is rejected, when $t(\underset{\sim}{x}) > c$. It is to be noted here that $t(\underset{\sim}{X})$ is a scalar and is of dimension one. Its sampling distribution is a univariate probability distribution. The values of $t(\underset{\sim}{X})$ satisfying the condition $t(\underset{\sim}{x}) > c$ will identify the samples in the sample space, which lead to rejection of H_0. It does not mean that $\left\{t \mid t(\underset{\sim}{x}) > c\right\}$ is the corresponding critical region. The value 'c', distinguishing the elements of the critical region and the acceptance region, is referred to as **critical value**. There may be one or many critical values for a hypotheses testing problem. The critical values are determined from the sampling distribution of the respective test statistic under H_0.

Example 1.4

Suppose an electrical equipment manufacturing industry receives screws in lots, as raw materials. The production engineer decides to reject a lot when the number of defective screws is one or more in a randomly selected sample of size 2.

$$\text{Define } X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ screw is defective} \\ 0, & \text{if } i^{\text{th}} \text{ screw is not defective} \end{cases}, \qquad i = 1, 2$$

Then, X_1 and X_2 are *iid* random variables and they have the *Bernoulli* (P) distribution.

Let $H_0 : P = \frac{1}{3}$ and $H_1 : P = \frac{2}{3}$

The sample space is $\mathbf{S}$ = {(0,0),(0,1),(1,0),(1,1)}

If $T(X_1, X_2)$ represents the number of defective screws, in each random sample, then the statistic $T(X_1, X_2) = X_1 + X_2$ is a random variable distributed according to the *Binomial* (2, P) distribution. The possible values of $T(X_1, X_2)$ are 0, 1 and 2. The values of $T(X_1, X_2)$ which lead to rejection of H_0 constitute the set {1,2}.

But, the critical region is defined by the elements of $\mathbf{S}$ corresponding to $T(X_1,X_2) = 1$ or 2. Thus, the critical region is {(0,1), (1,0), (1,1)} whose dimension is 2.

Note 8: When the sampling distribution is continuous, the set of values of $t(\underset{\sim}{X})$ corresponding to the rejection rule will be an interval or union of intervals depending on the alternative hypothesis. It is empahazized that **these intervals identify the elements of critical region, but they do not constitute the critical region**.

When the sampling distribution of the test statistic Z is a normal distribution, the critical values for testing H_0 against the possible alternative hypothesis at two different levels of significance, say 5% and 1% are displayed in Table 1.6.

Table 1.6 Critical values of the Z statistic

Alternative hypothesis	Level of Significance (α)	
	0.05 or 5%	0.01 or 1%
One- sided (right)	$z_\alpha = z_{0.05} = 1.645$	$z_\alpha = z_{0.01} = 2.33$
One- sided (left)	$-z_\alpha = -z_{0.05} = -1.645$	$-z_\alpha = -z_{0.01} = -2.33$
Two-sided	$z_{\alpha/2} = z_{0.025} = 1.96$	$z_{\alpha/2} = z_{0.005} = 2.58$

1.7 ONE-TAILED AND TWO-TAILED TESTS

In some hypotheses testing problem, elements of the critical region may be identified by a rejection rule of the type $t(\underset{\sim}{X}) \geq c$. In this case, $P(t(\underset{\sim}{X}) \geq c)$ will be the area, which falls at the right end (Figure1.1) under the curve representing the sampling distribution of $t(\underset{\sim}{X})$. The statistical test defined by this kind of critical region is called **right-tailed test**.

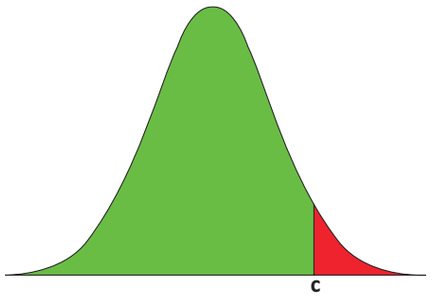


Figure 1.1. Right-tailed Test

On the other hand, suppose that the rejection rule $t(\underset{\sim}{X}) \leq c$ determines the elements of the critical region. Then, $P(t(\underset{\sim}{X}) \leq c)$ will be the area, which falls at the left end (Figure.1.2) under the curve representing the sampling distribution of $t(\underset{\sim}{X})$. The statistical test defined by this kind of critical region is called **left -tailed test**.

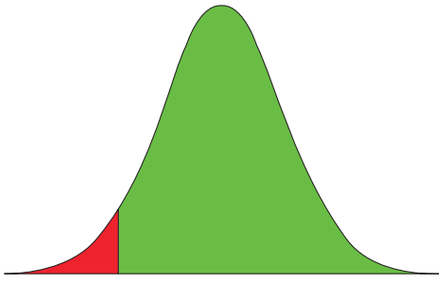


Figure 1.2 Left-tailed Test

The above two tests are commonly known as **one-tailed tests**.

Note 9: It should be noted that the sampling distribution of $t(\underset{\sim}{X})$ need not be with symmetric shape always. Sometimes, it may be positively or negatively skewed.

Example 1.5

Suppose a pizza restaurant claims its average pizza delivery time is 30 minutes. But you believe that the restaurant takes more than 30 minutes. Now, the null and the alternate hypotheses can be formulated as

H_0 : μ = 30 minutes and H_1: μ > 30 minutes

Suppose that the decision is taken based on the delivery times of 4 randomly chosen pizza deliveries of the restaurant. Let X_1, X_2, X_3, and X_4 represent the delivery times of the such four occasions. Also, let H_0 be rejected, when the sample mean exceeds 31. Then, the critical region is

$$\text{Critical Region} = \left\{(x_1, x_2, x_3, x_4) \mid \bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} > 31\right\}$$

In this case, $P(\bar{X} > 31)$ will be the area, which fall at the right end under the curve representing the sampling distribution of $\bar{X}$. Hence, this test can be categorized as a right-tailed test.

Suppose that H_0 is rejected, when either $t(\underset{\sim}{X}) \leq a$ or $t(\underset{\sim}{X}) \geq b$ holds. In this case, $P(t(\underset{\sim}{X}) \leq a)$ and $P(t(\underset{\sim}{X}) \geq b)$ will be the areas, which fall respectively at left and right ends under the curve representing the sampling distribution of $t(\underset{\sim}{X})$ (Figure 1.3). The statistical test defined with this kind of rejection rule is known as **two-tailed test**.

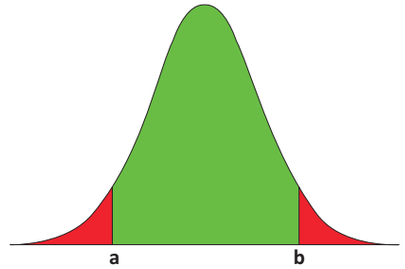


Figure 1.3 Two-tailed Test

Example 1.6

A manufacturer of ball-bearings, which are used in some machines, inspects to see whether the diameter of each ball-bearing is 5 mm. If the average diameter of ball-bearings is less than 4.75 mm or greater than 5.10 mm, then such ball-bearings will cause damages to the machine.

Here the null and the alternate hypotheses are

$$H_0 : \mu = 5 \text{ and } H_1 : \mu \neq 5.$$

Suppose that the decision on H_0 is made based on the diameter of 10 randomly selected ball-bearings. Let X_i, $i = 1, 2, \ldots, 10$ represent the diameter of the randomly chosen ball bearings. Then, the critical region is

$$\text{Critical Region} = \left\{ (x_1, x_2, \ldots, x_{10}) \mid \bar{x} = \frac{x_1 + x_2 + \ldots + x_{10}}{10} < 4.75 \text{ or } > 5.10 \right\}$$

In this case, $P\left(\bar{X} < 4.75\right)$ is the area, which will fall at the left end and $P\left(\bar{X} > 5.10\right)$ is the area, which will fall at the right end under the curve representing the sampling distribution of $\bar{X}$. This kind of test can be categorized as a two-tailed test (see Figure 1.3).

1.8 GENERAL PROCEDURE FOR TEST OF HYPOTHESES

The following steps constitute a general procedure, which can be followed for solving hypotheses testing problems based on both large and small samples.

Step 1 : Describe the population and its parameter(s). Frame the null hypothesis (H_0) and alternative hypothesis (H_1).

Step 2 : Describe the sample *i.e.*, data.

Step 3 : Specify the desired level of significance, α.

Step 4 : Specify the test statistic and its sampling distribution under H_0.

Step 5 : Calculate the value of the test statistic under H_0 for given sample.

Step 6 : Find the critical value(s) (table value(s)) from the statistical table generated from the sampling distribution of the test statistic under H_0 corresponding to α.

Step 7 : Decide on rejecting or not rejecting the null hypothesis based on the rejection rule which compares the calculated value(s) of the test statistic with the table value(s).

Now, let us see some of the large sample tests, which apply the above general procedure. As mentioned in Note-6, for large samples, the size of the sample is greater than or equal to 30. In the case of two samples considered for a hypotheses testing problem, the test is a large sample test, when the sizes of **both** the samples are greater than or equal to 30.

1.9 TEST OF HYPOTHESES FOR POPULATION MEAN
(Population variance is known)

Procedure:

Step 1 : Let μ and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is known. If μ_0 is an admissible value of μ, then frame the null hypothesis as H_0: $\mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) H_1: $\mu \neq \mu_0$ (ii) H_1: $\mu > \mu_0$ (iii) H_1: $\mu < \mu_0$

Step 2 : Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is large ($n \geq 30$).

Step 3 : Let the level of significance be α.

Step 4 : Consider the test statistic $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$ under H_0. Here, $\overline{X}$ represents the sample mean, which is defined in Note 2. The approximate sampling distribution of the test statistic under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculate the value of Z for the given sample $(x_1, x_2, \ldots, x_n)$ as

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

Step 6 : Find the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.7

A company producing LED bulbs finds that mean life span of the population of its bulbs is 2000 hours with a standard derivation of 150 hours. A sample of 100 bulbs randomly chosen is found to have the mean life span of 1950 hours. Test, at 5% level of significance, whether the mean life span of the bulbs is significantly different from 2000 hours.

Solution:

Step 1 : Let μ and σ represent respectively the mean and standard deviation of the probability distribution of the life span of the bulbs. It is given that σ = 150 hours. The null and alternative hypotheses are

Null hypothesis: H_0: μ = 2000

i.e., the mean life span of the bulbs is not significantly different from 2000 hours.

Alternative hypothesis: H_1 : $\mu \neq 2000$

i.e., the mean life span of the bulbs is significantly different from 2000 hours.

It is a two-sided alternative hypothesis.

Step 2 : **Data**

The given sample information are

Sample size (n) = 100, Sample mean ($\bar{x}$) = 1950 hours

Step 3 : **Level of significance**

α = 5%

Step 4 : **Test statistic**

The test statistic is $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, under H_0

Under the null hypothesis H_0, Z follows the $N(0,1)$ distribution.

Step 5 : **Calculation of Test Statistic**

The value of Z under H_0 is calculated from

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

as

$$z_0 = \frac{1950 - 2000}{150 / \sqrt{100}}$$

$$= -3.33$$

Thus; $|z_0|$ = 3.33

Step 6 : **Critical value**

Since H_1 is a two-sided alternative, the critical value at α = 0.05 is $z_e = z_{0.025}$ = 1.96. (*see* Table 1.6).

Step 7 : **Decision**

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq z_e$. Thus, it is a two-tailed test. For the given sample information, the rejection rule holds *i.e.*, $|z_0| = 3.33 > z_e = 1.96$. Hence, H_0 is rejected in favour of H_1: $\mu \neq 2000$. Thus, the mean life span of the LED bulbs is significantly different from 2000 hours.

Example 1.8

The mean breaking strength of cables supplied by a manufacturer is 1900 n/m^2 with a standard deviation of 120 n/m^2. The manufacturer introduced a new technique in the manufacturing process and claimed that the breaking strength of the cables has increased. In order to test the claim, a sample of 60 cables is tested. It is found that the mean breaking strength of the sampled cables is 1960 n/m^2. Can we support the claim at 1% level of significance?

Solution:

Step 1 : Let μ and σ represent respectively the mean and standard deviation of the probability distribution of the breaking strength of the cables. It is given that $\sigma = 120$ n/m^2. The null and alternative hypotheses are

Null hypothesis H_0: $\mu = 1900$

i.e., the mean breaking strength of the cables is not significantly different from 1900n/m^2.

Alternative hypothesis: H_1: $\mu > 1900$

i.e., the mean breaking strength of the cables is significantly more than 1900n/m^2.

It may be noted that it is a one-sided (right) alternative hypothesis.

Step 2 : **Data**

The given sample information are

Sample size (n) = 60. Hence, it is a large sample.

Sample mean ($\bar{x}$)= 1960

Step 3 : **Level of significance**

$\alpha = 1\%$

Step 4 : **Test statistic**

The test statistic is $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, under H_0

Since n is large, under the null hypothesis, the sampling distribution of Z is the $N(0,1)$ distribution.

Step 5 : **Calculation of test statistic**

The value of Z under H_0 is calculated from $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

$$z_0 = \frac{1960 - 1900}{120 / \sqrt{60}}$$

Thus, $z_0 = 3.87$

Step 6 : **Critical value**

Since H_1 is a one-sided (right) alternative hypothesis, the critical value at $\alpha = 0.01$ level of significance is $z_e = z_{0.01} = 2.33$ (*see* Table 1.6)

Step 7 : Decision

Since H_1 is a one-sided (right) alternative, elements of the critical region are determined by the rejection rule $z_0 > z_e$. Thus, it is a right-tailed test. For the given sample information, the observed value $z_0 = 3.87$ is greater than the critical value $z_e = 2.33$. Hence, the null hypothesis H_0 is rejected. Therefore, the mean breaking strength of the cables is significantly more than 1900 n/m^2.

Thus, the manufacturer's claim that the breaking strength of cables has increased by the new technique is found valid.

1.10 TEST OF HYPOTHESES FOR POPULATION MEAN (*POPULATION VARIANCE IS UNKNOWN*)

Procedure:

Step1 : Let μ and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ, then frame the null hypothesis as $H_0: \mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) $H_1: \mu \neq \mu_0$ (ii) $H_1: \mu > \mu_0$ (iii) $H_1: \mu < \mu_0$

Step 2 : Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is large ($n \geq 30$).

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $Z = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ under H_0, where $\bar{X}$ and S are the sample mean and sample standard deviation respectively. It may be noted that the above test statistic is obtained from Z considered in the test described in Section 1.9 by substituting S for σ.

The approximate sampling distribution of the test statistic under H_0 is the $N(0,1)$ distribution.

YOU WILL KNOW

It is important to note that the exact sampling distribution of Z is the *Student's 't'* distribution with $(n - 1)$ degrees of freedom, when n is small ($n < 30$). This hypotheses testing problem, when n is small, is discussed, in detail, in Chapter 2. When n is large, the *Student's* 't' distribution converges to the $N(0,1)$ distribution.

Step 5 : Calculate the value of Z for the given sample $(x_1, x_2, \ldots, x_n)$ as $z_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$. Here, $\bar{x}$ and s are respectively the values of $\bar{X}$ and S calculated for the given sample.

Step 6 : Find the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Rejection Rule	$\|z_0\| \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.9

A motor vehicle manufacturing company desires to introduce a new model motor vehicle. The company claims that the mean fuel consumption of its new model vehicle is lower than that of the existing model of the motor vehicle, which is 27 kms/litre. A sample of 100 vehicles of the new model vehicle is selected randomly and their fuel consumptions are observed. It is found that the mean fuel consumption of the 100 new model motor vehicles is 30 kms/litre with a standard deviation of 3 kms/litre. Test the claim of the company at 5% level of significance.

Solution:

Step 1 : Let the fuel consumption of the new model motor vehicle be assumed to be distributed according to a distribution with mean and standard deviation respectively μ and σ. The null and alternative hypotheses are

Null hypothesis H_0: $\mu = 27$

i.e., the average fuel consumption of the company's new model motor vehicle is not significantly different from that of the existing model.

Alternative hypothesis H_1: $\mu > 27$

i.e., the average fuel consumption of the company's new model motor vehicle is significantly lower than that of the existing model. In other words, the number of kms by the new model motor vehicle is significantly more than that of the existing model motor vehicle.

Step 2 : **Data:**

The given sample information are

Size of the sample (n) = 100. Hence, it is a large sample.

Sample mean ($\bar{x}$)= 30

Sample standard deviation(s) = 3

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

The test statistic under H_0 is

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

Since n is large, the sampling distribution of Z under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z for the given sample information is calculated from

$$z_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ as}$$

$$z_0 = \frac{30 - 27}{3/\sqrt{100}}$$

Thus, $z_0 = 10$.

Step 6 : Critical Value

Since H_1 is a one-sided (right) alternative hypothesis, the critical value at $\alpha = 0.05$ is $z_e = z_{0.05} = 1.645$.

Step 7 : Decision

Since H_1 is a one-sided (right) alternative, elements of the critical region are defined by the rejection rule $z_0 > z_e = z_{0.05}$. Thus, it is a right-tailed test. Since, for the given sample information, $z_0 = 10 > z_e = 1.645$, H_0 is rejected.

1.11 TEST OF HYPOTHESES FOR EQUALITY OF MEANS OF TWO POPULATIONS *(Population variances are known)*

Procedure:

Step-1 : Let μ_X and σ_X^2 be respectively the mean and the variance of Population -1. Also, let μ_Y and σ_Y^2 be respectively the mean and the variance of Population -2 under study. Here σ_X^2 and σ_Y^2 are known admissible values.

Frame the null hypothesis as $H_0: \mu_X = \mu_Y$ and choose the suitable alternative hypothesis from

(i) $H_1: \mu_X \neq \mu_Y$ (ii) $H_1: \mu_X > \mu_Y$ (iii) $H_1: \mu_X < \mu_Y$

Step 2 : Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2, where m and n are large(*i.e.*, $m \geq 30$ and $n \geq 30$). Further, these two samples are assumed to be independent.

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $Z = \dfrac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{m} + \dfrac{\sigma_Y^2}{n}}}$ under H_0, where $\bar{X}$ and $\bar{Y}$ are respectively the means of the two samples described in Step-2.

The approximate sampling distribution of the test statistic $Z = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{\sigma_X^2}{m} + \dfrac{\sigma_Y^2}{n}}}$ under H_0 (*i.e.*, $\mu_X = \mu_Y$) is the *N(0,1)* distribution.

It may be noted that the test statistic, when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, is $Z = \dfrac{(\bar{X} - \bar{Y})}{\sigma\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$.

Step 5 : Calculate the value of Z for the given samples $(x_1, x_2, ..., x_m)$ and $(y_1, y_2, ..., y_n)$ as $z_o = \dfrac{(\bar{x} - \bar{y})}{\sqrt{\dfrac{\sigma_X^2}{m} + \dfrac{\sigma_Y^2}{n}}}$.

Here, $\bar{x}$ and $\bar{y}$ are respectively the values of $\bar{X}$ and $\bar{Y}$ for the given samples.

Step 6 : Find the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.10

Performance of students of X Standard in a national level talent search examination was studied. The scores secured by randomly selected students from two districts, *viz.*, D_1 and D_2 of a State were analyzed. The number of students randomly selected from D_1 and D_2 are respectively 500 and 800. Average scores secured by the students selected from D_1 and D_2 are respectively 58 and 57. Can the samples be regarded as drawn from the identical populations having common standard deviation 2? Test at 5% level of significance.

Solution:

Step 1 : Let μ_X and μ_Y be respectively the mean scores secured in the national level talent search examination by all the students from the districts D_1 and D_2 considered for the study. It is given that the populations of the scores of the students of these districts have the common standard deviation $\sigma = 2$. The null and alternative hypotheses are

Null hypothesis: H_0: $\mu_X = \mu_Y$

i.e., average scores secured by the students from the study districts are not significantly different.

Alternative hypothesis: H_1: $\mu_X \neq \mu_Y$

i.e., average scores secured by the students from the study districts are significantly different. It is a two-sided alternative.

Step 2 : Data

The given sample information are

Size of the Sample-1 (m) = 500

Size of the Sample-2 (n) = 800. Hence, both the samples are large.

Mean of Sample-1 ($\bar{x}$) = 58

Mean of Sample-2 ($\bar{y}$) = 57

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under the null hypothesis H_0 is

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Since both m and n are large, the sampling distribution of Z under H_0 is the $N(0, 1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z is calculated for the given sample information from

$$z_0 = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ as}$$

$$z_0 = \frac{58 - 57}{2\sqrt{\frac{1}{500} + \frac{1}{800}}}$$

$$z_0 = 8.77$$

Step-6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $z_e = z_{0.025} = 1.96$.

Step-7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are defined by the rejection rule $|z_0| \geq z_e = z_{0.025}$. For the given sample information, $|z_0| = 8.77 > z_e = 1.96$. It indicates that the given sample contains sufficient evidence to reject H_0. Thus, it may be decided that H_0 is rejected. Therefore, the average performance of the students in the districts D_1 and D_2 in the national level talent search examination are significantly different. Thus the given samples are not drawn from identical populations.

1.12 TEST OF HYPOTHESES FOR EQUALITY OF MEANS OF TWO POPULATIONS (*POPULATION VARIANCES ARE UNKNOWN*)

Procedure:

Step-1 : Let μ_X and σ_X^2 be respectively the mean and the variance of Population -1. Also, let μ_Y and σ_Y^2 be respectively the mean and the variance of Population -2 under study. Here σ_X^2 and σ_Y^2 are assumed to be unknown.

Frame the null hypothesis as H_0: $\mu_X = \mu_Y$ and choose the suitable alternative hypothesis from

(i) H_1: $\mu_X \neq \mu_Y$ (ii) H_1: $\mu_X > \mu_Y$ (iii) H_1: $\mu_X < \mu_Y$

Step 2 : Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2, where m and n are large ($m \geq 30$ and $n \geq 30$). Here, these two samples are assumed to be independent.

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \text{ under } H_0 \text{ (i.e., } \mu_X = \mu_Y).$$

i.e., the above test statistic is obtained from Z considered in the test described in Section 1.11 by substituting S_X^2 and S_Y^2 respectively for σ_X^2 and σ_Y^2

The approximate sampling distribution of the test statistic $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$ under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculate the value of Z for the given samples $(x_1, x_2, \ldots, x_m)$ and $(y_1, y_2, \ldots, y_n)$ as

$$z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}$$

.

Here $\bar{x}$ and $\bar{y}$ are respectively the values of $\bar{X}$ and $\bar{Y}$ for the given samples. Also, s_X^2 *and* s_Y^2 are respectively the values of S_X^2 and S_Y^2 for the given samples.

Step 6 : Find the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.11

A Model Examination was conducted to XII Standard students in the subject of Statistics. A District Educational Officer wanted to analyze the Gender-wise performance of the students using the marks secured by randomly selected boys and girls. Sample measures were calculated and the details are presented below:

Gender	Sample Size	Sample Mean	Sample Standard deviation
Boys	100	50	4
Girls	150	51	5

Test, at 5% level of significance, whether performance of the students differ significantly with respect to their gender.

Solution:

Step 1 : Let μ_X and μ_Y denote respectively the average marks secured by boys and girls in the Model Examination conducted to the XII Standard students in the subject of Statistics. Then, the null and the alternative hypotheses are

Null hypothesis: $H_0: \mu_X = \mu_Y$

i.e., there is no significant difference in the performance of the students with respect to their gender.

Alternative hypothesis: $H_1 : \mu_X \neq \mu_Y$

i.e., performance of the students differ significantly with the respect to the gender. It is a two-sided alternative hypothesis.

Step 2 : **Data**

The given sample information are

Gender of the Students	Sample Size	Sample Mean	Sample Standard Deviation
Boys	$m = 100$	$\overline{x} = 50$	$s_X = 4$
Girls	$n = 150$	$\overline{y} = 51$	$s_Y = 5$

Since $m \geq 30$ and $n \geq 30$, both the samples are large.

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

The test statistic under H_0 is

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

The sampling distribution of Z under H_0 is the $N(0,1)$ distribution.

Step 5 : **Calculation of the Test Statistic**

The value of Z is calculated for the given sample informations from

$$z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \quad \text{as}$$

$$z_0 = \frac{50 - 51}{\sqrt{\frac{4^2}{100} + \frac{5^2}{150}}}$$

Thus, $z_0 = -1.75$

Step 6 : **Critical value**

Since H_1 is a two-sided alternative, the critical value at 5% level of significance is $z_e = z_{0.025} = 1.96$.

Step 7 : **Decision**

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq |z_0|$. Thus it is a two-tailed test. But, $|z_0| = 1.75$ is less than the critical value $z_e = 1.96$. Hence, it may inferred as the given sample information does not provide sufficient evidence to reject H_0. Therefore, it may be decided that there is no sufficient evidence in the given sample to conclude that performance of boys and girls in the Model Examination conducted in the subject of Statistics differ significantly.

1.13 TEST OF HYPOTHESES FOR POPULATION PROPORTION

Procedure:

Step 1 : Let P denote the proportion of the population possessing the qualitative characteristic (attribute) under study. If p_0 is an admissible value of P, then frame the null hypothesis as $H_0: P = p_0$ and choose the suitable alternative hypothesis from

(i) $H_1: P \neq p_0$ (ii) $H_1: P > p_0$ (iii) $H_1: P < p_0$

Step 2 : Let p be proportion of the sample observations possessing the attribute, where n is large, $np > 5$ and $n(1 - p) > 5$.

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$ under H_0. Here, $Q = 1 - P$.

The approximate sampling distribution of the test statistic under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculate the value of Z under H_0 for the given data as $z_0 = \dfrac{p - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$, $q_0 = 1 - p_0$.

Step 6 : Choose the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$P \neq p_0$	$P > p_0$	$P < p_0$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$P \neq p_0$	$P > p_0$	$P < p_0$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.12

A survey was conducted among the citizens of a city to study their preference towards consumption of tea and coffee. Among 1000 randomly selected persons, it is found that 560 are tea-drinkers and the remaining are coffee-drinkers. Can we conclude at 1% level of significance from this information that both tea and coffee are equally preferred among the citizens in the city?

Solution:

Step 1 : Let P denote the proportion of people in the city who preferred to consume tea. Then, the null and the alternative hypotheses are

Null hypothesis: $H_0 : P = 0.5$

i.e., it is significant that both tea and coffee are preferred equally in the city.

Alternative hypothesis: $H_1 : P \neq 0.5$

i.e., preference of tea and coffee are not significantly equal. It is a two-sided alternative hypothesis.

Step 2 : **Data**

The given sample information are

Sample size (n) = 1000. Hence, it is a large sample.

No. of tea-drinkers = 560

Sample proportion (p) = $\dfrac{560}{1000} = 0.56$

Step 3 : **Level of significance**

α = 1%

Step 4 : Test statistic

Since n is large, $np = 560 > 5$ and $n(1 - p) = 440 > 5$, the test statistic under the null hypothesis, is $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$.

Its sampling distribution under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z can be calculated for the sample information from

$$z_0 = \frac{p - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \text{ as}$$

$$z_0 = \frac{0.56 - 0.50}{\sqrt{\dfrac{0.5 \times 0.5}{1000}}}$$

Thus, $z_0 = 3.79$

Step 6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at 1% level of significance is $z_{\alpha/2} = z_{0.005} = 2.58$.

Step 7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq z_e$. Thus it is a two-tailed test. Since $|z_0| = 3.79 > z_e = 2.58$, reject H_0 at 1% level of significance. Therefore, there is significant evidence to conclude that the preference of tea and coffee are different.

1.14 TEST OF HYPOTHESES FOR EQUALITY OF PROPORTIONS OF TWO POPULATIONS

Procedure:

Step 1 : Let P_X and P_Y denote respectively the proportions of Population-1 and Population-2 possessing the qualitative characteristic (attribute) under study. Frame the null hypothesis as H_0: $P_X = P_Y$ and choose the suitable alternative hypothesis from

(i) H_1: $P_X \neq P_Y$ (ii) H_1: $P_X > P_Y$ (iii) H_1: $P_X < P_Y$

Step 2 : **Let p_X and p_Y denote respectively the proportions of the samples of sizes m and n** drawn from Population-1 and Population-2 possessing the attribute, where m and n are large (*i.e.*, $m \geq 30$ and $n \geq 30$). Also, $mp_X > 5$, $m(1 - p_X) > 5$, $np_Y > 5$ and $n(1 - p_Y) > 5$. Here, these two samples are assumed to be independent.

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $Z = \frac{(p_X - p_Y) - (P_X - P_Y)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$ under H_0. Here, $\hat{p} = \frac{mp_X + np_Y}{m+n}$, $\hat{q} = 1 - \hat{p}$. The approximate sampling distribution of the test statistic under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculate the value of Z for the given data as $z_0 = \frac{p_X - p_Y}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$.

Step 6 : Choose the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$P_X \neq P_Y$	$P_X > P_Y$	$P_X < P_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$P_X \neq P_Y$	$P_X > P_Y$	$P_X < P_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example 1.13

A study was conducted to investigate the interest of people living in cities towards self-employment. Among randomly selected 500 persons from City-1, 400 persons were found to be self-employed. From City-2, 800 persons were selected randomly and among them 600 persons are self-employed. Do the data indicate that the two cities are significantly different with respect to prevalence of self-employment among the persons? Choose the level of significance as $\alpha = 0.05$.

Solution:

Step1 : Let P_X and P_Y be respectively the proportions of self-employed people in City-1 and City-2. Then, the null and alternative hypotheses are

Null hypothesis: $H_0 : P_X = P_Y$

i.e., there is no significant difference between the proportions of self-employed people in City-1 and City-2.

Alternative hypothesis: $H_1 : P_X \neq P_Y$

i.e., difference between the proportions of self-employed people in City-1 and City-2 is significant. It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

City	Sample Size	Sample Proportion
City-1	$m = 500$	$p_X = \frac{400}{500} = 0.80$
City-2	$n = 800$	$p_Y = \frac{600}{800} = 0.75$

Here, $m \geq 30$, $n \geq 30$, $mp_X = 400 > 5$, $m(1-p_X) = 100 > 5$, $np_Y = 600 > 5$ and $n(1-p_Y) = 200 > 5$.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under the null hypothesis is

$$Z = \frac{p_X - p_Y}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} \text{ where } \hat{p} = \frac{mp_X + np_Y}{m+n} \text{ and } \hat{q} = 1-\hat{p}$$

The sampling distribution of Z under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z for given sample information is calculated from

$$z_0 = \frac{p_X - p_Y}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}}.$$

Now, $\hat{p} = \frac{400+600}{500+800} = \frac{1000}{1300} = 0.77$ and $\hat{q} = 0.23$

Thus, $z_0 = \frac{0.80-0.75}{\sqrt{(0.77)(0.23)\left(\frac{1}{500}+\frac{1}{800}\right)}}$

$z_0 = 2.0764$

Step 6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at 5% level of significance is $z_e = 1.96$.

Step 7 : Decision

Since H_0 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| > z_e$. Thus, it is a two-tailed test. For the given sample information, $z_e = 2.0764 > z_e = 1.96$. Hence, H_0 is rejected. We can conclude that the difference between the proportions of self-employed people in City-1 and City-2 is significant.

POINTS TO REMEMBER

- Statistic is a random variable and its probability distribution is called sampling distribution.
- Generally, the random sample used in Statistical Inference is drawn under sampling with replacement from a finite population.
- Standard error is the standard deviation of the sampling distribution.
- Hypothesis is a statement on the population or the values of the parameters.
- Null hypothesis is a hypothesis which is tested for possible rejection.
- Statistical test leads to take decision on the null hypothesis.
- In each statistical hypotheses testing problem, there is one null hypothesis and one alternative hypothesis.
- Type I error is rejecting the true null hypothesis.
- Type II error is not rejecting a false null hypothesis.
- Upper limit of the P (Type I error) is called level of significance, denoted by α.
- Critical region is a subset of the sample space defined by the rejection rule.
- Critical value identifies the elements of critical region.
- If the number of sample observations is greater than or equal to 30, it is called large sample.
- For hypotheses testing based on two samples, if the sizes of both the samples are greater than or equal to 30, they are called large samples.
- For testing population proportion, the sampling distribution of the test statistic is *N(0, 1)*, only when $n \geq 30$, $np > 5$ and $n(1 - p) > 5$.
- For testing equality of two population proportions, the sampling distribution of the test statistic is *N (0, 1)* only when $m \geq 30$, $n \geq 30$, $mp_X > 5$, $m(1 - p_X) > 5$, $np_Y > 5$ and $n(1 - p_Y) > 5$.

EXERCISE 1

7Q1BW9

I. Choose the best answer.

1. Standard error of the sample mean is

   (*a*) σ^2 (*b*) $\frac{\sigma}{n}$

   (*c*) $\frac{\sigma}{\sqrt{n}}$ (*d*) $\frac{\sqrt{n}}{\sigma}$

2. When n is large and σ^2 is unknown, σ^2 is replaced in the test statistic by

   (a) Sample mean (b) Sample variance

   (c) Sample standard deviation (d) Sample proportion

3. Critical region of a test is

(*a*) rejection region (b) acceptance region

(*c*) sample space (d) subset of the sample space

4. The critical value (table value) of the test statistic at the level of significance α for a two-tailed large sample test is

(a) $z_{\alpha/2}$ (b) z_{α}

(c) $-z_{\alpha}$ (d) $-z_{\alpha/2}$

5. In general, large sample theory is applicable when

(a) $n \geq 100$ (b) $n \geq 50$

(c) $n \geq 40$ (d) $n \geq 30$

6. When H_1 is a one-sided (right) alternative hypothesis, the critical region is determined by

(a) both right and left tails (b) neither right nor left tail

(c) right tail (d) left tail

7. Critical value at 5% level of significance for two-tailed large sample test is

(a) 1.645 (b) 2.33

(c) 2.58 (d) 1.96

8. For testing H_0: $\mu = \mu_0$ against $H_1 : \mu < \mu_0$, what is the critical value at $\alpha = 0.01$?

(a) 1.645 (b) –1.645

(c) –2.33 (d) 2.33

9. The hypotheses testing problem H_0: $\mu_0 = 45$ against H_1: $\mu_0 < 45$ be categorized as

(a) left-tailed (b) right-tailed

(c) two-tailed (d) one-tailed

10. When the alternative hypothesis is $H_1 : \mu \neq \mu_0$, the critical region will be determined by

(a) both right and left tails (b) neither right nor left tail

(c) right tail (d) left tail

11. Rejecting H_0, when it is true is called

(a) type I error (b) type II error

(c) sampling error (d) standard error

12. What is the standard error of the sample proportion under H_0?

(a) $\sqrt{\frac{PQ}{n}}$ (b) $\sqrt{\frac{pq}{n}}$

(c) $\frac{PQ}{n}$ (d) $\frac{pq}{n}$

13. The test statistic for testing the equality of two population means, when the population variances are known is

(a) $\dfrac{\overline{X}-\overline{Y}}{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}$ (b) $\dfrac{\overline{X}-\overline{Y}}{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}$

(c) $\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}$ (d) $\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}}$

14. When the population variances are known and equal, the statistic $Z=\dfrac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}$ is used to test the null hypothesis

(a) $H_0 : \mu = \mu_0$ (b) $H_0 : \mu_X = \mu_Y$

(c) $H_0 : P_X = P_Y$ (d) $H_0 : P = p_0$

15. What is the standard error of the difference between two sample proportions, $(P_X - P_Y)$?

(a) $\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}$ (b) $\sqrt{\hat{p}\hat{q}}\left(\frac{1}{m}+\frac{1}{n}\right)$

(c) $\hat{p}\hat{q}\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)}$ (d) $\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)$

16. What is level of significance?

(a) P(Type I Error) (b) P(Type II Error)

(c) upper limit of P(Type I Error) (d) upper limit of P(Type II Error)

17. Large sample test for testing population proportion is valid, when

(a) n is large, n ≥ 30, $np > 5$, $n(1 - p) > 5$ (b) n is large, n ≥ 30, $np > 5$, $n(1 - p) < 5$

(c) n is large, n ≥ 30, $np > 5$, $np\ (1 - p) > 5$ (d) n is large, n ≥ 30, $np > 5$, $np(1 - p) < 5$

18. Large sample test can be applied for two-sample problems, for testing the equality of two population means, when

(a) $m + n \geq 30$ (b) $m \geq 30$, $n \geq 30$

(c) $m \geq 30$, $m + n \geq 30$ (d) $mn \geq 30$

19. Large sample test is applicable, when the parent population is

(a) normal distribution only (b) binomial distribution

(c) Poisson distribution (d) any probability distribution

20. What is the rejection rule, based on large sample, for testing H_0 against one-sided (left) alternative hypothesis?

(a) $|z_0| \geq z_{\alpha/2}$ (b) $z_0 < -z_\alpha$

(c) $z_0 \leq -z_{\alpha/2}$ (d) $z_0 > z_\alpha$

II. Give very short answer to the following questions.

21. What is Inferential Statistics?
22. Define sample space.
23. What do you understand by random sample?
24. What are the different types of errors in hypotheses testing problem?
25. Prescribe the rejection rules for testing H_0: $\mu = \mu_0$ against all possible alternative hypotheses.
26. List the possible alternative hypotheses and the corresponding rejection rules followed in testing equality of two population means.
27. Specify the alternative hypotheses and the rejection rules prescribed for testing equality of two population proportions.
28. Define parameter.
29. What is statistic?
30. Define standard error of a statistic.
31. What do you understand by sampling distribution of a statistic?
32. What is null hypothesis?
33. Distinguish between null and alternative hypotheses.
34. For testing a null hypothesis against a two-sided alternative hypothesis, if $|z_0| > z_{\alpha/2}$ is found to hold, what will be your decision?
35. In a recent survey, conducted among 2000 randomly selected graduates in a district, 367 of them are IAS aspirants. Calculate the proportion of IAS aspirants among the graduates in the district.
36. Which type of test should be used for testing $H_0 : \mu = 27$ against $H_1 : \mu > 27$?
37. Formulate the null and alternative hypotheses for testing whether the average time required to XI standard students to complete a Chemistry laboratory exercise is less than 30 minutes.
38. If μ denotes the population mean, then find the critical value to be used for testing H_0: $\mu = 100$ against $H_1 : \mu < 100$ based on 250 observations at 5% level of significance.
39. The mean height of students studying X Standard in a school is 150 cms and the average height of 15 randomly selected students is 155 cms. Identify the parameter, statistic and their values from these information.
40. Calculate standard error of the sample mean, when sample mean is 100, sample size is 64 and population standard deviation is 24.
41. Find standard error of the sample proportion $p = 0.45$ when the population proportion is 0.5 and the sample size is 100.

III. Give short answer to the following questions.

42. Describe Decision Table.
43. What are type I and type II errors?
44. What do you mean by level of significance?
45. Explain one-tailed and two-tailed tests.
46. Explain critical value.

47. A set of 100 students is selected randomly from an Institution. The mean height of these students is 163 *cms* and the standard deviation is 10 *cms*. Calculate the value of the test statistic under $H_0 : \mu = 167$.

48. In a random sample of 50 students from School A, 35 of them preferred junk food. In another sample of 80 from School B, 40 of them preferred junk food. Find the standard error of the difference between two sample proportions.

49. If m = 35, n = 40, $\bar{x}$ = 10.8, $\bar{y}$ =11.9, s_X = 3 and s_Y = 4, then calculate standard error of $\overline{X} - \overline{Y}$.

50. In test for population proportion, if n=500 and np=383, then calculate the value of the test statistic under $H_0 : P = 0.68$.

51. In test for two population proportions, if m = 100, n = 150, mp_X = 78 and np_Y =100, then calculate the value of the test statistic under H_0: $P_X = P_Y$.

IV. Give detailed answer to the following questions.

52. Explain the general procedure to be followed for testing of hypotheses.

53. Explain the procedure for testing hypotheses for population mean, when the population variance is unknown.

54. How will you formulate the hypotheses for testing equality of means of two populations, when the population variances are known? Describe the method.

55. Describe the procedure for testing hypotheses concerning equality of means of two populations, assuming that the population variances are unknown.

56. Give a detailed account on testing hypotheses for population proportion.

57. Explain the procedure of testing hypotheses for equality of proportion of two populations.

58. A special training programme was organized by a District Educational Officer to the VIII Standard students for improving their skill in Letter Writing. Time taken by the students in a Letter Writing competition was recorded. The average time taken by 100 randomly selected students was 15 minutes. Can the Officer decide that at 1% level of significance the mean time in this kind of exercise required to VIII Students of the district is 13 minutes, assuming the population standard deviation as 8 minutes?

59. Carry out hypotheses testing exercise for testing $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X \neq \mu_Y$ with usual notations, when $\bar{x}$ = 7 and $\bar{y}$ =8 , σ_X = 3 and σ_Y = 2 and m = 40 and n = 40. Use α = 0.01.

60. Chief Educational Officer wanted to study the performance of XII Standard students in Mathematics subject. The following are the information obtained from randomly selected students from two Educational Districts.

Educational District	No. of students selected	Mean	Standard Deviation
A	45	62	15
B	53	60	17

Examine at 5% level of significance whether students in District A perform better compared to students in District B.

61. The mean yield of rice observed from randomly selected 100 plots in District A was 210 kg *per acre* with standard deviation of 10 kg *per acre*. The mean yield of rice observed from randomly selected 150 plots in District B was 220 kg *per acre* with standard deviation of 12 kg *per acre*. Assuming that the standard deviation of yield in the entire state was 11kg, test at 1% level of significance whether difference between the mean yields of rice in the two districts is significant.

62. The standard deviation of the scores secured by XI standard students in a test conducted for examining their numerical ability is known to be 15. A school implemented a new method of teaching which is supposed to increase general quantitative ability. A group of 99 students are randomly assigned to one of two classes. Fifty students in Class-I are given the new method of teaching, whereas the remaining 49 students in Class-II are taught in the standard way. At the end of the particular term, students are given the same test of ability. The average scores secured by the students studied in Class-I and Class-II are respectively 116.0 and 113.1. Does this information provide significant evidence at 5% level, to conclude that the new method improved the numerical ability of students?

63. A machine assesses the life of a ball point pen, by measuring the length of a continuous line drawn using the pen. A random sample of 80 pens of Brand A have a total writing length of 96.84 km. Random sample of 75 pens of Brand B have a total writing length of 93.75 km. Assuming that the standard deviation of the writing length of a single pen is 0.15 km for both brands, can the consumer decide to choose Brand B pens assuming that their average writing length is more than that of Brand A pens? Set level of significance as 1%.

64. A study was conducted to compare the performance of athletes of two States in Inter-State Athlete Meets. Details of the number of successes achieved by the athletes of the two States are given hereunder:

State	No. of Athletes	Mean	Standard Deviation
State-1	300	75	10
State-2	400	73	11

Does the above information ensure at 1% level of significance that the difference between the performances of the athletes of the two States is significant?

65. A District Administration conducted awareness campaign on a contagious disease utilizing the services of school students. Among 64 randomly selected households, 50 of them appreciated the involvement of students. Can the District Administration decide whether more than 90% success could be achieved in these kinds of programmes by involving the students? Fix the level of significance as 1%.

66. A coin is tossed 10, 000 times and head turned up 5,195 times. Test the hypothesis, at 5% level of significance, that the coin is unbiased.

67. A study was conducted among randomly selected families who are living in two locations of a district, and parents were asked "Whether watching TV programmes by parents affects the studies of their children?" Details are presented hereunder:

Locality	No. of Families	
	Contacted	Agreed
A	200	48
B	600	96

Test, at 5% level of significance, whether the difference between the proportions of families in the two localities agreed the statement.

68. One thousand apples kept under one type of storage were found to show rotting to the extent of 4% and 1500 apples kept under another kind of storage showed 3% rotting. Can it be reasonably concluded at 5% level of significance that the second type of storage is superior to the first?

69. Preference of school students, who participate in Sports events, to do physical exercises in Modern Gymnasium rather than doing aerobic exercises was analyzed. The number of students randomly selected from two States and their preference for Modern Gymnasium are given below.

State	No. of Students	
	Sampled	Preferred Modern Gymnasium
A	50	38
B	60	52

Test whether the difference between proportions of school students who prefer Modern Gymnasium to do their exercises in the two States is significant at 5% level of significance.

70. Interest of XII Students on Residential Schooling was investigated among randomly selected students from two regions. Among 300 students selected from Region A, 34 students expressed their interest. Among 200 students selected from Region B, 28 students expressed their interest. Does this information provide sufficient evidence to conclude at 5% level of significance that students in Region A are more interested in Residential Schooling than the students in Region B?

ACTIVITIES

1. In your institution, collect data from your Physical Education department about height/weight of the students in a particular class (Standard 9). Take a random sample of 50 students from your school and find the mean and standard deviation for their height/weight. Verify through the statistical tests of inferential statistics, Find the significant difference between sample mean and population mean. Give your comments.
2. Consider two group of students of same class (Standard 10) taking an examination in a particular subject in Tamil medium and in English medium in your school. Get samples from two medium and find their mean and standard deviation. Is there any significant difference between the performance between the students taking their examination in Tamil medium and English medium? Discuss with your friends.

Note: The teacher and students discuss and they may create their own problems similar to the above problems and expand this exercises.

ANSWERS

I. 1. (c) 2. (b) 3. (a) 4. (a) 5. (d)

6. (c) 7. (d) 8. (*c*) 9. (*a*) 10. (a)

11. (a) 12.(a) 13. (c) 14. (b) 15. (a)

16. (c) 17. (a) 18. (b) 19. (d) 20. (b)

II. 34. reject H_0 35. 0.1835 36. one-tailed test (right)

37. H_0: $\mu = 30$, H_1: $\mu < 30$ 38. −1.645 39. −1.645

39. Parameter = Mean height of X standard students in the school = 150 *cms*

Statistic = Average height of sampled students in the school = 155 *cms*

40. $SE(\bar{X}) = 3$

41. $SE(p) = 0.05$

III. 47. $z_0 = -4$

48. 0.089

49. 0.8106

50. 4.122

51. 1.937

IV. 58. $z_0 = 2.5$; do not reject H_0

59. $z_0 = -1.75$; do not reject H_0

60. $z_0 = 0.619$; do not reject H_0

61. $z_0 = -7.04$; reject H_0

62. $z_0 = 1.302$; do not reject H_0

63. $z_0 = -1.638$; do not reject H_0

64. $z_0 = 2.508$; do not reject H_0

65. $z_0 = -3.167$; do not reject H_0

66. $z_0 = 3.90$; reject H_0

67. $z_0 = 2.55$; reject H_0

68. $z_0 = 1.352$; do not reject H_0

69. $z_0 = -1.444$; do not reject H_0

70. $z_0 = -0.886$; do not reject H_0

ICT CORNER

TESTS OF SIGNIFICANCE – BASIC CONCEPTS AND LARGE SAMPLE TESTS

This activity helps to understand about Significance in Hypothesis Testing

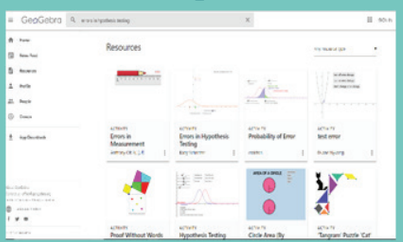
Steps:

- Open the browser and type the URL given (or) scan the QR code.GeoGebra work book called "**Tests of Significance – Basic Concepts and Large Sample Tests**"will appear.
- In that Tick the square boxes for "Show Rejection Region-Show Type I Error probability-Show Alternative Distribution"
- Now If press and drag the Orange and Blue colour buttons you will get the different results.

Step-1

Step-2

Step-3



Step-4

Pictures are indicatives only*

URL:

https://www.geogebra.org/m/VzUpY5M

B273_12_STAT_EM

CHAPTER

2 TESTS BASED ON SAMPLING DISTRIBUTIONS I

W.S. Gossett

W Gosset (1876-1937), born in England studied Chemistry and Mathematics at New College , Oxford. Upon graduating in 1899, he joined a brewery in Ireland. Gosset applied his statistical knowledge both in the brewery and on the farm to the selection of the best varieties of Barley. Gosset acquired that knowledge by study, by trial and error, and by spending two terms in 1906–1907 in the biometrical laboratory of Karl Pearson. Gosset and Pearson had a good relationship. Pearson helped Gosset with the mathematics of his research papers. The brewery where he was employed allowed publishing his work under a pseudonym ("Student"). Thus, his most noteworthy achievement is now called Student's *t*, rather than Gosset's, *t*-distribution.

B1BNMF

LEARNING OBJECTIVES

The student will be able to

- understand the purpose for using *t*-test and chi-square test .
- understand procedures for tests of hypotheses based on small samples.
- solve problems to test the hypotheses concerning mean(s) using *t*-distribution.
- solve problems to test the hypothesis whether the population has a particular variance using chi-square test.
- solve problems to test the hypotheses relating to independence of attributes and goodness of fit using chi-square test.

Introduction

In the earlier chapter, we have discussed various problems related to tests of significance based on large samples by applying the standard normal distribution. However, if the sample size is small ($n < 30$) the sampling distributions of test statistics are far from normal and the procedures discussed in Chapter-1 cannot be applied, except the general procedure (Section 1.8). But in this case, there exists a probability distribution called *t*-distribution which may be used instead of standard normal distribution to study the problems based on small samples.

2.1 STUDENT'S *t* DISTRIBUTION AND ITS APPLICATIONS

2.1.1 Student's *t*-distribution

χ^2-distribution and some of its applications are discussed in Section 2.2

If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ are independent random variables, then

$T = \dfrac{X}{\sqrt{Y/n}}$ is said to have *t*-distribution with *n* degrees of freedom. This can be denoted by t_n.

Note 1: The degrees of freedom of *t* is the same as the degrees of freedom of the corresponding chi-square random variable.

Note 2: The *t*-distribution is used as the sampling distribution(s) of the statistics(s) defined based on random sample(s) drawn from normal population(s).

i) If $X_1, X_2, \ldots, X_n$ is a random sample drawn from $N(\mu, \sigma^2)$ population then

$$X = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } Y = \frac{\sum (X_i - X)^{\sim}}{\sigma^2} \sim \chi^2_{n-1} \text{ are independent.}$$

Hence,

$$T_1 = \frac{X}{\sqrt{Y/n-1}}$$

$$= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma\sqrt{(n-1)}}{\sqrt{\sum (X_i - \overline{X})^2}}$$

$$= \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \text{where } S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{(n-1)}}.$$

ii) If $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ are independent random samples drawn from $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ populations respectively, then

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1) \text{ and } \frac{\sum_{i=1}^{m} (X_i - \overline{X})^2 + \sum_{j=1}^{n} (Y_j - \overline{Y})^2}{\sigma^2} \sim \chi^2_{m+n-2} \text{ are independent.}$$

Then,

$$T_2 = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

where $S_p^2 = \dfrac{\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 + \sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2}{m+n-2}$

3. If $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a random sample of n paired observations drawn from a bivariate normal population, then $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$ is a random sample drawn from $N(\mu_D, \sigma_D^2)$. Here $\mu_D = \mu_X - \mu_Y$.

   Hence,

$$T_3 = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

2.1.2 Properties of the Student's *t*-distribution

1. *t*–distribution is symmetrical distribution with mean zero.
2. The graph of *t*-distribution is similar to normal distribution except for the following two reasons:
   i. The normal distribution curve is higher in the middle than *t*-distribution curve.
   ii. *t*–distribution has a greater spread sideways than the normal distribution curve. It means that there is more area in the tails of *t*-distribution.
3. The *t*-distribution curve is asymptotic to *X*-axis, that is, it extends to infinity on either side.
4. The shape of *t*-distribution curve varies with the degrees of freedom. The larger is the number of degrees of freedom, closeness of its shape to standard normal distribution (fig. 2.1).
5. Sampling distribution of *t* does not depend on population parameter. It depends on degrees of freedom $(n-1)$.

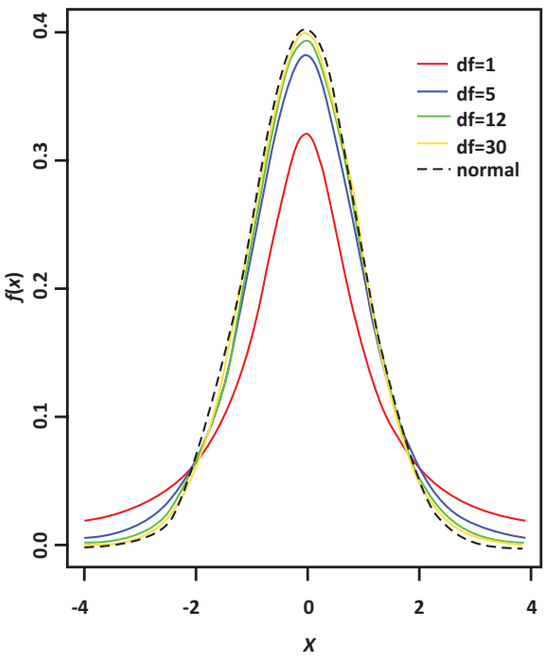


Figure 2.1. Student's *t*-distribution

2.1.3 Applications of *t*-distribution

The *t*-distribution has the following important applications in testing the hypotheses for small samples.

1. To test significance of a single population mean, when population variance is unknown, using T_1.

2. To test the equality of two population means when population variances are equal and unknown, using T_2.
3. To test the equality of two means – paired t-test, based on dependent samples, T_3.

YOU WILL KNOW

The t-distribution has few more applications but they are not considered in this Chapter. You will study these applications in higher classes.

2.1.4 Test of Hypotheses for Normal Population Mean (Population Variance is Unknown)

Procedure:

Step 1 : Let μ and σ^2 be respectively the mean and variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ, then frame the null hypothesis as

H_0: $\mu = \mu_0$ and choose the suitable alternative hypothesis from

(i) H_1: $\mu \neq \mu_0$ (ii) H_1: $\mu > \mu_0$ (iii) H_1: $\mu < \mu_0$

Step 2 : Describe the sample/data and its descriptive measures. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is small ($n < 30$).

Step 3 : Specify the level of significance, α.

Step 4 : Consider the test statistic $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ under H_0, where $\bar{X}$ and S are the sample mean and sample standard deviation respectively. The approximate sampling distribution of the test statistic under H_0 is the t-distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of t for the given sample $(x_1, x_2, \ldots x_n)$ as $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$.

here $\bar{x}$ is the sample mean and $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Critical Value (t_e)	$t_{n-1,\alpha/2}$	$t_{n-1,\alpha}$	$-t_{n-1,\alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1,\alpha/2}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example 2. 1

The average monthly sales, based on past experience of a particular brand of tooth paste in departmental stores is ₹ 200. An advertisement campaign was made by the company and then a sample of 26 departmental stores was taken at random and found that the average sales of the particular brand of tooth paste is ₹ 216 with a standard deviation of ₹ 8. Does the campaign have helped in promoting the sales of a particular brand of tooth paste?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 200$

i.e., the average monthly sales of a particular brand of tooth paste is not significantly different from ₹ 200.

Alternative Hypothesis H_1: $\mu > 200$

i.e., the average monthly sales of a particular brand of tooth paste are significantly different from ₹ 200. It is one-sided (right) alternative hypothesis.

Step 2 : Data

The given sample information are:

Size of the sample $(n) = 26$. Hence, it is a small sample.

Sample mean $(\bar{x}) = 216$, Standard deviation of the sample = 8.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t-distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ as}$$

$$t_0 = \frac{216 - 200}{8/\sqrt{26}} = 10.20$$

Step 6 : Critical value

Since H_1 is one-sided (right) alternative hypothesis, the critical value at α =0.05 is

$$t_e = t_{n-1,\,\alpha} = t_{25,0.05} = 1.708$$

Step 7 : Decision

Since it is right-tailed test, elements of critical region are defined by the rejection rule $t_0 > t_e = t_{n-1,\,\alpha} = t_{25,0.05} = 1.708$. For the given sample information $t_0 = 10.20 > t_e = 1.708$. It indicates that given sample contains sufficient evidence to reject H_0. Hence, the campaign has helped in promoting the increase in sales of a particular brand of tooth paste.

Example 2.2

A sample of 10 students from a school was selected. Their scores in a particular subject are 72, 82, 96, 85, 84, 75, 76, 93, 94 and 93. Can we support the claim that the class average scores is 90?

Solution:

Step 1 : Hypotheses

Null Hypothesis H_0: $\mu = 90$

i.e., the class average scores is not significantly different from 90.

Alternative Hypothesis $H_1 : \mu \neq 90$

i.e., the class means scores is significantly different from 90.

It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Size of the sample $(n) = 10$. Hence, it is a small sample.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Since n is small, the sampling distribution of T is the t - distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of T for the given sample information is calculated from $t_0 = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ as under:

x_i	$u_i = x_i - A$; $(A = 85)$	u_i^2
72	–13	169
82	–3	9
96	11	121
85	0	0
84	–1	1
75	–10	100
76	–9	81
93	8	64
94	9	81
93	8	64
	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 690$

Sample mean

$$\bar{x} = A + \frac{\sum_{i=1}^{10} u_i}{n} \text{ where } A \text{ is assumed mean}$$

$= 85 + 0 = 85$

Sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{10} u_i^2}$$

$$= \sqrt{\frac{1}{9} \times 690}$$

$$= \sqrt{76.67}$$

$= 8.756$

Hence,

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{85-90}{8.756/\sqrt{10}} = \frac{-5}{2.77}$$

$= -1.806$ and

$|t_0| = 1.806$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{n-1, \frac{\alpha}{2}} = t_{9,0.025} = 2.262$

Step 7 : Decision

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| > t_e = t_{n-1, \frac{\alpha}{2}} = t_{9,0.025} = 2.262$. For the given sample information $|t_0| = 1.806 < t_e = 2.262$. It indicates that given sample does not provide sufficient evidence to reject H_0. Hence, we conclude that the class average scores is 90.

2.1.5 Test of Hypotheses for Equality of Means of Two Normal Populations (Independent Random Samples)

Procedure:

Step 1 : Let μ_X and μ_Y be respectively the means of population-1 and population-2 under study. The variances of the population-1 and population-2 are assumed to be equal and unknown given by σ^2.

Frame the null hypothesis as $H_0 : \mu_X = \mu_Y$ and choose the suitable alternative hypothesis from (i) $H_1 : \mu_X \neq \mu_Y$ (ii) $H_1 : \mu_X > \mu_Y$ (iii) $H_1 : \mu_X < \mu_Y$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2, where m and n are small (*i.e.*, $m < 30$ and $n < 30$). Here, these two samples are assumed to be independent.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ under } H_0 \ (i.e., \mu_X = \mu_Y)$$

where S_p is the "pooled" standard deviation (combined standard deviation) given by

$$S_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}};$$

and

$$s_X{}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$s_Y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The approximate sampling distribution of the test statistic

$$T = \frac{(\bar{X} - \bar{Y})}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \text{under } H_0$$

is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$.

Step 5 : Calculate the value of T for the given sample $(x_1, x_2, \ldots x_m)$ and $(y_1, y_2, \ldots y_n)$ as

$$t_0 = \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Here $\bar{x}$ and $\bar{y}$ are the values of $\bar{X}$ and $\bar{Y}$ for the samples. Also $s_x{}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$, $s_y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ are the sample variances and $s_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}}$.

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \frac{\alpha}{2}}$	$-t_{(n-1), \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1,\frac{\alpha}{2}}$	$t_0 > t_{n-1,\alpha}$	$t_0 < -t_{n-1,\alpha}$

Example 2.3

The following table gives the scores (out of 15) of two batches of students in an examination.

Batch I	6	7	9	2	13	3	4	8	7	11
Batch II	5	6	5	7	1	7	2	7		

Test at 1% level of significance the average performance of the students in Batch I and Batch II are equal.

Solution:

Step 1 : **Hypotheses:** Let μ_X and μ_Y denote respectively the average performance of students in Batch I and Batch II. Then the null and alternative hypotheses are :

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are equal.

Alternative Hypothesis $H_1 : \mu_X \neq \mu_Y$

i.e., the average performance of the students in Batch I and Batch II are not equal.

Step 2 : **Data**

The given sample information are:

Sample size for Batch I : m =10

Sample size for Batch II : n = 8

Step 3 : **Level of significance**

α = 1%

Step 4 : **Test statistic**

The test statistic under H_0 is

$$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

The sampling distribution of T under H_0 is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

To find sample mean and sample standard deviation:

x_i	$u_i = x_i - \bar{x}$ $(\bar{x} = 7)$	u_i^2	y_i	$v_i = y_i - \bar{y}$ $(\bar{y} = 5)$	v_i^2
6	-1	1	5	0	0
7	0	0	6	1	1
9	2	4	5	0	0
2	-5	25	7	2	4
13	6	36	1	-4	16
3	-4	16	7	2	4
4	-3	9	2	-3	9
8	1	1	7	2	4
7	0	0			
11	4	16			
$\sum_{i=1}^{10} x_i = 70$	$\sum_{i=1}^{10} u_i = 0$	$\sum_{i=1}^{10} u_i^2 = 108$	$\sum_{i=1}^{8} y_i = 40$	$\sum_{i=1}^{8} v_i = 0$	$\sum_{i=1}^{8} v_i^2 = 38$

To find sample means:

Let $(x_1, x_2, ..., x_{10})$ and $(y_1, y_2, ..., y_8)$ denote the scores of students in Batch I and Batch II respectively.

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{70}{10} = 7$$

$$\bar{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{40}{8} = 5$$

To find combined sample standard deviation:

$$s_X^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{10} u_i^2 = \frac{108}{9} = 12$$

$$s_Y^2 = \frac{1}{7}\sum_{i=1}^{8}(y_i - \bar{y})^2 = \frac{1}{7}\sum_{i=1}^{8} v_i^2 = \frac{38}{7} = 5.4$$

Pooled standard deviation is:

$$S_p = \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}} = \sqrt{\frac{108+38}{10+8-2}} = \sqrt{9.125} = 3.021$$

The value of T is calculated for the given information as

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{7-5}{3.021\sqrt{\frac{1}{10} + \frac{1}{8}}} = 1.3957$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.01$ is

$t_e = t_{m+n-2, \frac{\alpha}{2}} = t_{16,0.005} = 2.921$

Step 7 : Decision

Since it is two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n-2, \frac{\alpha}{2}} = t_{16,0.005} = 2.921$. For the given sample information $|t_0| = 1.3957 < t_e = 2.921$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, the mean performance of the students in these batches are equal.

Example 2.4

Two types of batteries are tested for their length of life (in hours). The following data is the summary descriptive statistics.

Type	Number of batteries	Average life (in hours)	Sample standard deviation
A	14	94	16
B	13	86	20

Is there any significant difference between the average life of the two batteries at 5% level of significance?

Solution:

Step 1 : Hypotheses

Null Hypothesis $H_0 : \mu_X = \mu_Y$

i.e., there is no significant difference in average life of two types of batteries *A* and *B*.

Alternative Hypothesis $H_0 : \mu_X \neq \mu_Y$

i.e., there is significant difference in average life of two types of batteries *A* and *B*. It is a two-sided alternative hypothesis

Step 2 : Data

The given sample information are :

m = number of batteries under type A = 14

n = number of batteries under type B = 13

$\bar{x}$ = Average life (in hours) of type A battery = 94

$\bar{y}$ = Average life (in hours) of type B battery = 86

s_X = standard deviation of type A battery =16

s_Y = standard deviation of type B battery = 20

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

The sampling distribution of T under H_0 is the t-distribution with $m+n-2$ degrees of freedom *i.e.*, $t \sim t_{m+n-2}$

Step 5 : Calculation of test statistic

Under null hypotheses H_0:

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where s is the pooled standard deviation given by,

$$s_p = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}}$$

$$= \sqrt{\frac{(14-1)(16)^2 + (13-1)(20)^2}{14+13-2}} = \sqrt{325.12} = 18.03$$

The value of T is calculated for the given information as

$$t_0 = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{94-86}{18.03\sqrt{\frac{1}{14} + \frac{1}{13}}} = \frac{8}{6.944} = 1.15$$

Step 6 : Critical value

Since H_1 is two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $t_e = t_{m+n-2,\,\frac{\alpha}{2}} = t_{25,\,0.025} = 2.060$.

Step 7 : Decision

Since it is a two-tailed test, elements of critical region are defined by the rejection rule $|t_0| < t_e = t_{m+n-2,\,\frac{\alpha}{2}} = t_{25,\,0.025} = 2.060$. For the given sample information $|t_0| = 1.15 < t_e = 2.060$. It indicates that given sample contains insufficient evidence to reject H_0. Hence, there is no significant difference between the average life of the two types of batteries.

2.1.6 To test the equality of two means – paired *t*-test

Procedure:

Step 1 : Let X and Y be two correlated random variables having the distributions respectively $N(\mu_X, \sigma_X^2)$ (Population-1) and $N(\mu_Y, \sigma_Y^2)$ (Population-2). Let $D = X - Y$, then it has normal distribution $N(\mu_D = \mu_X - \mu_Y, \sigma_D^2)$.

Frame null hypothesis as

$H_0 : \mu_D = 0$

And choose alternative hypothesis from

(i) $H_1 : \mu_D \neq 0$ (ii) $H_1 : \mu_D > 0$ (iii) $H_1 : \mu_D < 0$

Step 2 : Describe the sample/data. Let $(X_1, X_2, \ldots, X_m)$ be a random sample of m observations drawn from Population-1 and $(Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations drawn from Population-2. Here, these two samples are correlated in pairs.

Step 3 : Set up level of significance (α)

Step 4 : Consider the test statistic

$$T = \frac{\bar{D}}{\frac{S}{\sqrt{n}}} \text{ under } H_0 .$$

where $\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$; $D_i = X_i - Y_i$ and $S = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}}$.

The approximate sampling distribution of the test statistic T under H_0 is t - distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of T for the given data as

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$; $d_i = x_i - y_i$ (sample mean) and

$s = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$ (sample standard deviation)

Step 6 : Choose the critical value, t_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Critical Value (t_e)	$t_{n-1, \frac{\alpha}{2}}$	$t_{n-1, \alpha}$	$-t_{n-1, \alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_D \neq 0$	$\mu_D > 0$	$\mu_D < 0$
Rejection Rule	$\lvert t_0 \rvert \geq t_{n-1, \frac{\alpha}{2}}$	$t_0 > t_{n-1, \alpha}$	$t_0 < -t_{n-1, \alpha}$

Example 2.5

A company gave an intensive training to its salesmen to increase the sales. A random sample of 10 salesmen was selected and the value (in lakhs of Rupees) of their sales per month, made before and after the training is recorded in the following table. Test whether there is any increase in mean sales at 5% level of significance.

Salesman	1	2	3	4	5	6	7	8	9	10
Before	15	22	6	17	12	20	18	14	10	16
After	17	23	16	20	14	21	18	20	10	11

Solution:

Step 1 : **Hypotheses**

Null Hypothesis $H_0 : \mu_D = 0$

i.e., there is no significant increase in the mean sales after the training.

Alternative Hypothesis $H_1 : \mu_D > 0$

i.e., there is significant increase in the mean sales after the training. It is a one-sided alternative hypothesis.

Step 2 : **Data**

Sample size $n = 10$

Step 3 : **Level of significance**

$\alpha = 5\%$

Step 4 : **Test statistic**

Test statistic under the null hypothesis is

$$T = \frac{\overline{D}}{\frac{S}{\sqrt{n}}}$$

The sampling distribution of T under H_0 is t - distribution with (10–1) = 9 degrees of freedom.

Step 5 : Calculation of test statistic

To find $\bar{d}$ and s:

Let x denote sales before training and y denote sales after training

Salesmen	x_i	y_i	$d_i = y_i - x_i$	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1	15	17	2	0	0
2	22	23	1	-1	1
3	6	16	10	8	64
4	17	20	3	1	1
5	12	14	2	0	0
6	20	21	1	-1	1
7	18	18	0	-2	4
8	14	20	6	4	16
9	10	10	0	-2	4
10	16	11	-5	-7	49
		Total	$\sum_{i=1}^{n} d_i = 20$	$\sum_{i=1}^{n} (d_i - \bar{d}) = 0$	$\sum_{i=1}^{n} (d_i - \bar{d})^2 = 140$

Here instead of $d_i = x_i - y_i$ it is assumed $d_i = y_i - x_i$ for calculations to be simpler.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{20}{10} = 2$$

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (d_i - \bar{d})^2} = \sqrt{\frac{140}{9}} = \sqrt{15.56} = 3.94$$

The calculated value of the statistic is

$$t_0 = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} = \frac{2}{\frac{3.94}{\sqrt{10}}} = 1.6052$$

Step 6 : Critical value

Since H_0 is a one-sided alternative hypothesis, the critical value at 5% level of significance is $t_e = t_{n-1,\,\alpha} = t_{9,0.05} = 1.833$

Step 7 : Decision

It is a one-tailed test. Since $|t_0| = 1.6052 < t_e = t_{n-1,\,\alpha} = t_{9,0.05} = 1.833$, H_0 is not rejected. Hence, there is no evidence that the mean sales has increased after the training.

2.2 CHI-SQUARE DISTRIBUTION AND ITS APPLICATIONS

Karl Pearson

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. He was the first to examine whether the observed data support a given specification, in a paper published in 1900. He called it 'Chi-square goodness of fit' test which motivated research in statistical inference and led to the development of statistics as separate discipline.

Karl Pearson chi-square test the dawn of Statistical Inference - *C R Rao.*

Karl Pearson's famous chi square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics - *B. Efron*

2.2.1 Chi-Square distribution

The square of standard normal variable is known as a chi-square variable with 1 degree of freedom (d.f.). Thus

If $X \sim N(\mu, \sigma^2)$, then it is known that $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Further Z^2 is said to follow χ^2 – distribution with 1 degree of freedom (χ^2 – pronounced as chi-square)

Note: i) If $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ are n *iid* random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(X_i - \mu\right)/\sigma^2 \text{ follows } \chi^2 \text{ with } n \text{ } d.f \text{ (additive property of } \chi^2)$$

ii) If μ is replaced by $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then $\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2}$ follows χ^2_{n-1}

2.2.2 Properties of χ^2 distribution

- It is a continuous distribution.
- The distribution has only one parameter *i.e. n d.f.*
- The shape of the distribution depends upon the *d.f*, *n*.
- The mean of the chi-square distribution is n and variance $2n$
- If U and V are independent random variables having χ^2 distributions with degree of freedom n_1 and n_2 respectively, then their sum $U + V$ has the same χ^2 distribution with *d.f* $n_1 + n_2$.

2.2.3 Applications of chi-square distribution

- To test the variance of the normal population, using the statistic in note (ii) (sec. 2.2.1)
- To test the independence of attributes. (sec. 2.2.5)
- To test the goodness of fit of a distribution. (sec. 2.2.6)
- The sampling distributions of the test statistics used in the last two applications are approximately chi-square distributions.

2.2.4 Test of Hypotheses for population variance of the normal population (Population mean is assumed to be unknown)

Procedure

Step 1 : Let μ and σ^2 be respectively the mean and the variance of the normal population under study, where σ^2 is known and μ unknown. If ${\sigma_0}^2$ is an admissible value of σ^2, then frame the

Null hypothesis as H_0: $\sigma^2 = {\sigma_0}^2$

and choose the suitable alternative hypothesis from

(i) H_1: $\sigma^2 \neq {\sigma_0}^2$ (ii) H_1: $\sigma^2 > {\sigma_0}^2$ (iii) H_1: $\sigma^2 < {\sigma_0}^2$

Step 2 : Describe the sample/data and its descriptive measures. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is small ($n < 30$).

Step 3 : Fix the desired level of significance α.

Step 4 : Consider the test statistic $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$ under H_0. The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with $(n-1)$ degrees of freedom.

Step 5 : Calculate the value of the of χ^2 for the given sample as ${\chi_0}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

Step 6 : Choose the critical value of χ_e^2 corresponding to α and H_1 from the following table.

Alternative Hypothesis (H_1)	$\sigma^2 \neq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sigma^2 < \sigma_0^2$
Critical value (χ_e^2)	${\chi^2}_{n-1,\frac{\alpha}{2}}$ and ${\chi_0}^2 \leq {\chi^2}_{n-1,1-\frac{\alpha}{2}}$	$\chi^2_{n-1,\alpha}$	$\chi^2_{n-1,1-\alpha}$

Step 7 : Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\sigma^2 \neq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sigma^2 < \sigma_0^2$
Rejection Rule	${\chi^2}_{n-1,\frac{\alpha}{2}}$ and ${\chi_0}^2 \leq {\chi^2}_{n-1,1-\frac{\alpha}{2}}$	${\chi_0}^2 > {\chi^2}_{n-1,\alpha}$	${\chi_0}^2 < {\chi^2}_{n-1,1-\alpha}$

NOTE

If the population mean μ is known then for testing $H_0 : \sigma^2 = \sigma_0^2$ against any of the alternatives, we use $\chi_0^2 = \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2}$ with n *d.f.*

Example 2.6

The weights (in kg.) of 8 students of class VII are 38, 42, 43, 50, 48, 45, 52 and 50. Test the hypothesis that the variance of the population is 48 kg, assuming the population is normal and μ is unknown.

Solution:

Step 1 : **Null Hypothesis** H_0: $\sigma^2 = 48$ kg.

i.e. Population variance can be regarded as 48 kg.

Alternative hypothesis H_1: $\sigma^2 \neq 48$ kg.

i.e. Population variance cannot be regarded as 48 kg.

Step 2 : The given sample information is
Sample size $(n) = 8$

Step 3 : **Level of significance**
$\alpha = 5\%$

Step 4 : **Test statistic**

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with $(n-1)$ *d.f.*

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as under:

To find $\bar{x}$ and sample variance s^2, we form the following table.

x_i	$(x_i - 46)$	$(x_i - 46)^2$
38	-8	64
42	-4	16
43	-3	9
50	4	16
48	2	4
45	-1	1
52	6	36
50	4	16
$\sum_{i=1}^{8} x_i = 368$	0	$\sum_{i=1}^{8} (x_i - \bar{x})^2 = 162$

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{368}{8} = 46$$

$$s^2 = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{(n-1)} = \frac{\sum_{i=1}^{8}(x_i - 46)^2}{(8-1)} = \frac{162}{7} = 23.143.$$

The calculated value of chi-square is $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{162}{48} = 3.375$

Step 6 : Critical values

Since H_1 is a two sided alternative, the critical values at α =0.05 are $\chi^2_{7,\,0.025} = 16.01$ and $\chi^2_{7,0.975} = 1.69$.

Step 7 : Decision

Since it is a two-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^2 \geq \chi^2_{n-1,\frac{\alpha}{2}}$ or $\chi_0^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}$.

For the given sample information, the rejection rule does not hold, since

$1.69 = \chi^2_{7,\,0.975} < \chi_0^2\,(=3.375) < \chi^2_{7,0.025} = 16.01.$

Hence, H_0 is not rejected in favour of H_1. Thus, Population variance can be regarded as 48 kg.

Example 2.7

A normal population has mean μ (unknown) and variance 9. A sample of size 9 observations has been taken and its variance is found to be 5.4. Test the null hypothesis H_0: $\sigma^2 = 9$ against H_1: $\sigma^2 > 9$ at 5% level of significance.

Solution:

Step 1 : Null Hypothesis H_0: $\sigma^2 = 9$.

i.e., Population variance regarded as 9.

Alternative hypothesis H_1: $\sigma^2 > 9$.

i.e. Population variance is regarded as greater than 9.

Step 2 : Data

Sample size (n) = 9

Sample variance (s^2) = 5.4

Step 3 : Level of significance

α = 5%

Step 4 : Test statistic

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with $(n\text{-}1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{8 \times 5.4}{9} = 4.8$$

Step 6 : Critical value

Since H_1 is a one-sided alternative, the critical values at α =0.05 is $\chi_e^2 = \chi^2_{8,\,0.05} = 15.507$.

Step 7 : Decision

Since it is a one-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^2 > \chi_e^2$.

For the given sample information, the rejection rule does not hold , since $\chi_0^2 = 4.8 < \chi^2_{8,\,0.05} = 15.507$. Hence, H_0 is not rejected in favour of H_1. Thus, the population variance can be regarded as 9.

Example 2.8

A normal population has mean μ (unknown) and variance 0.018. A random sample of size 20 observations has been taken and its variance is found to be 0.024. Test the null hypothesis H_0: $\sigma^2 = 0.018$ against H_1: $\sigma^2 < 0.018$ at 5% level of significance.

Solution:

Step 1 : Null Hypothesis H_0: $\sigma^2 = 0.018$.

i.e. Population variance regarded as 0.018.

Alternative hypothesis H_1: $\sigma^2 < 0.018$.

i.e. Population variance is regarded as lessthan 0.018.

Step 2 : Data

Sample size $(n) = 20$

Sample variance $(s^2) = 0.024$

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

Under null hypothesis H_0

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

follows chi-square distribution with $(n\text{–}1)$ degrees of freedom.

Step 5 : Calculation of test statistic

The value of chi-square under H_0 is calculated as

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19 \times 0.024}{0.018} = 25.3$$

Step 6 : **Critical value**

Since H_1 is a one-sided alternative, the critical values at α =0.05 is $\chi_e^2 = \chi^2_{19, 0.95} = 10.117$.

Step 7 : **Decision**

Since it is a one-tailed test, the elements of the critical region are determined by the rejection rule $\chi_0^2 < \chi_e^2$

For the given sample information, the rejection rule does not hold, since

$\chi_0^2 = 25.3 > \chi_e^2 = \chi^2_{19, 0.95} = 10.117$.

Hence, H_0 is not rejected in favour of H_1. Thus, the population variance can be regarded as 0.018.

2.2.5 Test of Hypotheses for independence of Attributes

Another important application of χ^2 test is the testing of independence of attributes.

Attributes: Attributes are qualitative characteristic such as levels of literacy, employment status, *etc*., which are quantified in terms of levels/scores.

Contigency table: Independence of two attributes is an important statistical application in which the data pertaining to the attributes are cross classified in the form of a two – dimensional table. The levels of one attribute are arranged in rows and of the other in columns. Such an arrangement in the form of a table is called as a contingency table.

Computational steps for testing the independence of attributes:

Step 1 : **Framing the hypotheses**

Null hypothesis H_0: The two attributes are independent

Alternative hypothesis H_1: The two attributes are not independent.

Step 2 : **Data**

The data set is given in the form of a contigency as under. Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N};\ i = 1, 2, ...m;\ j = 1, 2, ..n$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

NOTE

The contingency table consisting of m rows and n columns.
The observed data is presented in the form of a contingency table :

		Attribute B						Total
		B_1	B_2	…	B_j	…	B_n	
Attribute A	A_1	O_{11}	O_{12}	…	O_{1j}	…	O_{1n}	R_1
	A_2	O_{21}	O_{22}	…	O_{2j}	…	O_{2n}	R_2
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{in}	R_i
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_m	O_{m1}	O_{m2}	…	O_{mj}	…	O_{mn}	R_m
Total	C_1	C_2	…	C_j	…	C_n	$N = m \times n$	

Step 3 : Level of significance

Fix the desired level of significance α

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Step 5 : Critical value

The critical value is obtained from the table of χ^2 with $(m-1)(n-1)$ degrees of freedom at given level of significance, α as $\chi^2_{(m-1)(n-1),\alpha}$.

Step 6 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value. If $\chi_0^2 \geq \chi^2_{(m-1)(n-1),\alpha}$ reject H_0.

Note:

- N, the total frequency should be reasonably large, say greater than 50.
- No theoretical cell-frequency should be less than 5. If cell frequencies are less than 5, then it should be grouped such that the total frequency is made greater than 5 with the preceding or succeeding cell.

Example 2.9

The following table gives the performance of 500 students classified according to age in a computer test. Test whether the attributes age and performance are independent at 5% of significance.

Performance	Below 20	21-30	Above 30	Total
Average	138	83	64	285
Good	64	67	84	215
Total	202	150	148	500

Solution:

Step 1 : **Null hypothesis** H_0: The attributes age and performance are independent.

Alternative hypothesis H_1: The attributes age and performance are not independent.

Step 2 : **Data**

Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N} \quad i = 1, 2; j = 1, 2, 3$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

Performance	Below average	Average	Above average	Total
Average	$\frac{285\times202}{500}=115.14$	$\frac{285\times150}{500}=85.5$	$\frac{285\times148}{500}=84.36$	285
Good	$\frac{215\times202}{500}=86.86$	$\frac{215\times150}{500}=64.5$	$\frac{215\times148}{500}=63.64$	215
Total	202	150	148	500

Step 3 : **Level of significance** α = 5%

Step 4 : **Calculation**

Calculate the value of the test statistic as

$$\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

This chi-square test statistic is calculated as follows:

$$\chi_0^2 = \frac{(138-115.14)^2}{115.14} + \frac{(83-85.50)^2}{88.50} + \frac{(64-84.36)^2}{84.36} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64}$$

= 22.152 with degrees of freedom (3–1)(2–1) = 2

Step 5 : **Critical value**

From the chi-square table the critical value at 5% level of significance is $\chi^2_{(2-1)(3-1),0.05} = \chi^2_{2,0.05} = 5.991$.

Step 6 : Decision

As the calculated value $\chi_0^2 = 22.152$ is greater than the critical value $\chi^2_{2,0.05} = 5.991$, the null hypothesis H_0 is rejected. Hence, the performance and age of students are not independent.

NOTE

If the contigency table is 2 x 2 then the value of χ^2 can be calculated as given below:

	A	not A	Total
B	a	b	$a+b$
not B	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_\alpha^2(1d.f)$$

The following example will illustrate the procedure

Example 2.10

A survey was conducted with 500 female students of which 60% were intelligent, 40% had uneducated fathers, while 30 % of the not intelligent female students had educated fathers. Test the hypothesis that the education of fathers and intelligence of female students are independent.

Solution:

Step 1 : Null hypothesis H_0: The attributes are independent *i.e.* No association between education fathers and intelligence of female students

Alternative hypothesis H_1: The attributes are not independent *i.e* there is association between education of fathers and intelligence of female students

Step 2 : Data

The observed frequencies (O) has been computed from the given information as under.

	Intelligent females	Not intelligent females	Row total
Educated fathers	$300-120=180$	$\frac{30}{100}\times 200 = 60$	240
Uneducated fathers	$\frac{40}{100}\times 300 = 120$	$200-60=140$	260
Total	$\frac{60}{100}\times 500 = 300$	$500-300=200$	N= 500

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where, a= 620, b = 380, c = 550, d = 450 and N = 2000

$$\chi_0^2 = \frac{2000(620\times 450 - 380\times 550)^2}{(620+380)(550+450)(620+550)(380+450)} = 10.092$$

Step 5 : Critical value

From chi-square table the critical value at 5% level of significance is $\chi^2_{1,0.05} = 3.841$

Step 6 : Decision

The calculated value $\chi_0^2 = 10.092$ is greater than the critical value $\chi^2_{1,0.05} = 3.841$, the null hypothesis H_0 is rejected. Hence, education of fathers and intelligence of female students are not independent.

2.2.6 Tests for Goodness of Fit

Another important application of chi-square distribution is testing goodness of a pattern or distribution fitted to given data. This application was regarded as one of the most important inventions in mathematical sciences during 20th century. Goodness of fit indicates the closeness of observed frequency with that of the expected frequency. If the curves of these two distributions do not coincide or appear to diverge much, it is noted that the fit is poor. If two curves do not diverge much, the fit is fair.

Computational steps for testing the significance of goodness of fit:

Step 1 : Framing of hypothesis

Null hypothesis H_0 : The goodness of fit is appropriate for the given data set

Alternative hypothesis H_1 : The goodness of fit is not appropriate for the given data set

Step 2 : Data

Calculate the expected frequencies (E_i) using appropriate theoretical distribution such as Binomial or Poisson.

Step 3 : Select the desired level of significance α

Step 4 : Test statistic

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where k = number of classes

O_i and E_i are respectively the observed and expected frequency of i^{th} class such that $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$.

If any of E_i is found less than 5, the corresponding class frequency may be pooled with preceding or succeeding classes such that E_i's of all classes are greater than or equal to 5. It may be noted that the value of k may be determined after pooling the classes.

The approximate sampling distribution of the test statistic under H_0 is the chi-square distribution with k-1-s *d.f* , s being the number of parametres to be estimated.

Step 5 : Calculation

Calculate the value of chi-square as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The above steps in calculating the chi-square can be summarized in the form of the table as follows:

Step 6 : Critical value

The critical value is obtained from the table of χ^2 for a given level of significance α.

Step 7 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value, at the desired level of significance.

Example 2.11

Five coins are tossed 640 times and the following results were obtained.

Number of heads	0	1	2	3	4	5
Frequency	19	99	197	198	105	22

Fit binomial distribution to the above data.

Solution:

Step 1 : Null hypothesis H_0: Fitting of binomial distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of binomial distribution is not appropriate to the given data.

Step 2 : Data

Compute the expected frequencies:

n = number of coins tossed at a time = 5

Let X denote the number of heads (success) in n tosses

N = number of times experiment is repeated = 640

To find mean of the distribution

x	f	fx
0	19	0
1	99	99
2	197	394
3	198	594
4	105	420
5	22	110
Total	640	1617

Mean : $\bar{x} = \frac{\sum fx}{\sum f} = \frac{1617}{640} = 2.526$

The probability mass function of binomial distribution is :

$p(x) = {}^nC_x\, p^x\, q^{n-x},\ x = 0,1,\ldots, n$ (2.1)

Mean of the binomial distribution is $\bar{x} = np$.

Hence, $\hat{p} = \frac{\bar{x}}{n} = \frac{2.526}{5} \approx 0.5$

$\hat{q} = 1 - \hat{p} \approx 0.5$

For $x = 0$, the equation (2.1) becomes

$P(X = 0) = P(0) = 5c_0\ (0.5)^5 = 0.03125$

The expected frequency at $x = N\,P(x)$

The expected frequency at x =0 : $N \times P(0)$

$= 640 \times 0.03125 = 20$

We use recurrence formula to find the other expected frequencies.

The expected frequency at x+1 is

$\frac{n-x}{x+1}\left(\frac{p}{q}\right)$ × Expected frequency at x

x	$\frac{n-x}{x+1}$	$\frac{p}{q}$	$\frac{n-x}{x+1}\left(\frac{p}{q}\right)$	Expected frequency at $x = N\,P(x)$
0	5	1	5	20
1	2	1	2	100
2	1	1	1	200
3	0.5	1	0.5	200
4	0.2	1	0.2	100
5	0	1	0	20

Table of expected frequencies:

Number of heads	0	1	2	3	4	5	Total
Expected frequencies	20	100	200	200	100	20	640

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Step 5 : Calculation

The test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
19	20	–1	1	0.050
99	100	–1	1	0.010
197	200	–3	9	0.045
198	200	–2	4	0.020
105	100	5	25	0.250
22	20	2	4	0.200
			Total	0.575

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 0.575$$

Step 6 : Critical value

Degrees of freedom = $k - 1 - s = 6 - 1 - 1 = 4$

Critical value for *d.f* 4 at 5% level of significance is 9.488 *i.e.*, $\chi^2_{4,0.05} = 9.488$

Step 7 : Decision

As the calculated $\chi_0^2 (=0.575)$ is less than the critical value $\chi^2_{4,0.05} = 9.488$, we do not reject the null hypothesis. Hence, the fitting of binomial distribution is appropriate.

Example 2.12

A packet consists of 100 ball pens. The distribution of the number of defective ball pens in each packet is given below:

x	0	1	2	3	4	5
f	61	14	10	7	5	3

Examine whether Poisson distribution is appropriate for the above data at 5% level of significance.

Solution:

Step 1 : **Null hypothesis** H_0: Fitting of Poisson distribution is appropriate for the given data.

Alternative hypothesis H_1: Fitting of Poisson distribution is not appropriate for the given data.

Step 2 : **Data**

The expected frequencies are computed as under:

To find the mean of the distribution.

x	f	fx
0	61	0
1	14	14
2	10	20
3	7	21
4	5	20
5	3	15
Total	100	90

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{90}{100} = 0.9$$

Probability mass function of Poisson distribution is:

$$p(x) = \frac{e^{-m} m^x}{x!};\ x = 0, 1, \ldots \qquad (2.2)$$

In the case of Poisson distribution mean $(m) = \bar{x} = 0.9$.

At $x = 0$, equation (2.2) becomes

$$p(0) = \frac{e^{-m}\ m^0}{0!} = e^m = e^{0.9} = 0.4066 .$$

The expected frequency at x is $N\,P(x)$

Therefore, The expected frequency at 0 is

$$N \times P(0)$$
$$= 100 \times 0.4066$$
$$= 40.66$$

We use recurrence formula to find the other expected frequencies.

The expected frequency at $x+1$ is

$$\frac{m}{x+1} \times \text{Expected frequency at } x$$

x	$\frac{m}{x+1}$	Expected frequency at $x = N\,P(x)$
0	0.9	40.66
1	$\frac{0.9}{2}$	36.594
2	$\frac{0.9}{3}$	16.4673
3	$\frac{0.9}{4}$	4.94019
4	$\frac{0.9}{5}$	1.1115
5	$\frac{0.9}{6}$	0.20007

Table of expected frequency distribution (on rounding to the nearest integer)

x	0	1	2	3	4	5
Expected frequency	41	37	16	5	1	0

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Step 5 : Calculation

Test statistic is computed as under:

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
61	41	20	400	9.756
14	37	-23	529	14.297
10	16	-6	36	2.250
7	5			
5 } 15	1 } 6	9	81	13.5
3	0			
			Total	51.253

Note: In the above table, we find the cell frequencies 0,1 in the expected frequency column (E) is less than 5, Hence, we combine (pool) with either succeeding or preceding one such that the total is made greater than 5. Here we have pooled with preceding frequency 5 such that the total frequency is made greater than 5. Correspondingly, cell frequencies in observed frequencies are pooled.

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 51.253$$

Step 6 : Critical value

Degrees of freedom = $(k - 1 - s) = 4 - 1 - 1 = 2$

Critical value for 2 *d.f* at 5% level of significance is 5.991 *i.e.*, $\chi^2_{2,0.05} = 5.991$

Step 7 : Decision

The calculated $\chi_0^2 (=51.253)$ is greater than the critical value (5.991) at 5% level of significance. Hence, we reject H_0. i.e., fitting of Poisson distribution is not appropriate for the given data.

Example 2.13

A sample 800 students appeared for a competitive examination. It was found that 320 students have failed, 270 have secured a third grade, 190 have secured a second grade and the remaining students qualified in first grade. The general opinion that the above grades are in the ratio 4:3:2:1 respectively. Test the hypothesis the general opinion about the grades is appropriate at 5% level of significance.

Step 1 : Null hypothesis H_0: The result in four grades follows the ratio 4:3:2:1

Alternative hypothesis H_1: The result in four grades does not follows the ratio 4:3:2:1

Step 2 : Data

Compute expected frequencies:

Under the assumption on H_0, the expected frequencies of the four grades are:

$$\frac{4}{10}\times 800 = 320\ ; \frac{3}{10}\times 800 = 240; \frac{2}{10}\times 800 = 160; \frac{1}{10}\times 800 = 80$$

Step 3 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
320	320	0	0	0
270	240	30	900	3.75
190	160	30	900	5.625
20	80	-60	3600	45
			Total	54.375

The test statistic is calculated as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
$$= 54.375$$

Step 4 : Critical value

The critical value of χ^2 for 3 d.f. at 5% level of significance is 7.81 *i.e.*, $\chi^2_{3,0.05} = 7.81$

Step 5 : Decision

As the calculated value of χ_0^2(=54.375) is greater than the critical value $\chi^2_{3,0.05} = 7.81$, reject H_0. Hence, the results of the four grades do not follow the ratio 4:3:2:1.

Example 2.14

The following table shows the distribution of digits in numbers chosen at random from a telephone directory.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the occurence of the digits in the directory are equal at 5% level of significance.

Step 1 : Null hypothesis H_0: The occurrence of the digits are equal in the directory.

Alternative hypothesis H_1: The occurrence of the digits are not equal in the directory.

Step 2 : Data

The expected frequency for each digit = $\frac{10000}{10} = 1000$

Step 3 : Level of significance $\alpha = 5\%$

Step 3 : Test statistic

The test statistic is computed using the following table.

Observed frequency (O_i)	Expected frequency (E_i)	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1026	1000	26	676	0.676
1107	1000	107	11449	11.449
997	1000	3	9	0.009
966	1000	34	1156	1.156
1075	1000	75	5625	5.625
933	1000	67	4489	4.489
1107	1000	107	11449	11.449
972	1000	28	784	0.784
964	1000	36	1296	1.296
853	1000	147	21609	21.609
			Total	58.542

The test statistic is calculated as

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= 58.542$$

Step 4 : Critical value

Critical value for 9 df at 5% level of significance is 16.919 i.e., $\chi^2_{9,0.05} = 16.919$

Step 5 : Decision

Since the calculated χ_0^2 (58.542) is greater than the critical value $\chi^2_{9,0.05} = 16.919$, reject H_0. Hence, the digits are not uniformly distributed in the directory.

POINTS TO REMEMBER

- If the number of elements in the sample is less than 30, it is called a small sample.
- For conducting t-tests the parent population(s) should be normal and the samples(s) should be small.
- In case of two sample problems based on t-distribution the sizes of both samples must be less than 30.
- The t-distribution is symmetrical about its mean(zero)
- When the degrees of freedom is large the t-distribution converges to $N(0,1)$ distribution.

- The degree of freedom represents the number of independent observation in the sample.
- The sampling distribution of the test statistic for testing hypothesis about normal population mean is t_{n-1}, when n is small and σ is unknown.
- The sampling distribution of the test statistic for testing equality of two normal population mean is t_{m+n-2} when $m, n < 30$ and the common population variance σ^2 is unknown.
- If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2$ with 1 d.f.
- The uses of χ^2 – distribution are (i) testing the specified variance of a normal population (ii) testing goodness of fit and (iii) testing independence of attributes.
- When expected frequency for a cell is less than 5, it is should be clubbed with the adjacent cells such that the expected frequency in the resultant cell is greater than 5.
- The degrees of freedom for the χ^2 – statistic used for the independence of attributes is $(m-1) \times (n-1)$, where m and n are respectively the number of rows and columns in a contegency table.
- The expected cell frequency testing independence of attributes is calculated as
$$\frac{\text{Row total} \times \text{Column total}}{\text{Sample Size}}$$
- The expected cell frequency in testing goodness of fit is calculated as sample size × {probability for the corresponding cell}

EXERCISE 2

I. Choose the best answer.

278RKM

1. Student's 't' distribution was found by
   a) Karl Pearson b) Laplace
   c) R.A. Fisher d) William S.Gosset
2. Support of student's t random variable is
   a) $-\infty < t \leq 0$ b) $0 \leq t < \infty$
   c) $-\infty < t < \infty$ d) $0 \leq t \leq 1$
3. Paired t-test is applicable when the observations in both the samples are
   a) Paired b) Correlated
   c) Equal in number d) all the above
4. The number of degrees of freedom for the test statistic $t = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}$ is
   a) $n-1$ b) n c) $n-2$ d) $n+1$
5. Standard error of difference between two sample means in the case of small samples is
   a) $\sigma^2\sqrt{\frac{1}{m}+\frac{1}{n}}$ b) $\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}$ c) $s_p\sqrt{\frac{1}{m}+\frac{1}{n}}$ d) $\sqrt{\frac{\sigma_1}{m}+\frac{\sigma_2}{n}}$

6. If the size of sample is larger than 30, the t-distribution tends to
   a) normal distribution b) F-distribution
   c) chi-square distribution d) Poisson distribution

7. If a random sample of 10 observations has variance 324 then standard error is
   a) $18/\sqrt{10}$ b)18/10
   c) 10/18 d) $2/\sqrt{5}$

8. A sample of 16 units was taken for testing an hypothesis concerning the mean of a normal population. Then the degrees of freedom of the appropriate test statictic is
   a) 14 b) 15
   c) 16 d) 8

9. If s_1^2 and s_2^2 are respectively the variance of two independent random samples of sizes 'm' and 'n'. Then standard deviation of the combined sample is
   a) $\sqrt{\frac{ms_1^2 + ns_2^2}{m+n}}$ b) $\sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n}}$
   c) $\sqrt{\frac{ms_1^2 + ns_1^2}{m+n+2}}$ d) $\sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}$

10. A company gave an intensive training to its salesman to increase the sales. A random sample of 6 salesmen was selected and the value of their sales made before and after the training is recorded. Which test will be more appropriate to test whether there is an increase in mean sales
    a) normal test b) paired t-test c) χ^2-test d) F-test

11. If the order of the contigency table is (5×4). Then the degree of freedom of the corresponding chi-square test statistic is
    a) 18 b) 17 c) 12 d) 25

12. For testing the hypothesis concerning variance of a normal population _____ is used.
    a) t-test b) F-test c) Z-test d) χ^2-test

13. If σ^2 is the variance of normal population, then the degrees of freedom of the sampling distribution of the test statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ is:
    a) $n-1$ b) $n+1$ c) n d) $n-2$

14. If n is the degree of freedom of chi-square distribution then its variance is
    a) n b) $n-1$ c) $2n$ d) $n+1$

15. If chi-square is performed for testing goodness of fit to a data with k classes on estimating 's' parameters then degrees of fredom of test statistic is.
    a) $k-s$ b) $(k-1)(s-1)$
    c) $k-1-s$ d) $k-1$

16. The statistic χ^2, with usual notations, in case of contingency table of order $(m \times n)$ is given by
    a) $\chi_0^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ b) $\chi_0^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)}{E_i}\right]^2$
    c) $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$ d) $\chi_c^2 = \sum_{i=1}^{k} \frac{O_i}{E_i}$

II. Give very short answer to the following questions.

17. Define student's t-statistic.
18. Define: degrees of freedom.
19. Define the paired t-statistic.
20. When paired t-test can be applied?
21. Write the test statistic to test the difference between normal population means.
22. Write the standard error of the difference between sample means.
23. Define chi-square statistic.
24. Write the applications of chi-square distribution.
25. What are the minimum requirements of chi-square test?
26. Define an attribute.
27. Give the recurrence formula for binomial distribution.

III. Give short answer to the following questions.

28. List out the properties of t-distribution.
29. Write down the applications of t-distribution.
30. Explain the testing procedure to test the normal population mean, when population variance is unknown.
31. Write down the procedure to test significance for equality of means of two normal populations based on small samples.
32. A random sample of ten students is taken and their marks in a particular subject are recorded. The average mark is 60 with standard deviation 6.5. Test the hypothesis that the average mark of students is 55.
33. State the properties of χ^2 distribution.
34. What is a contigency table?
35. Write the procedure to test the population variance.
36. Write the test procedure for testing the independence of attributes.
37. Write down the computational steps for testing the significance of goodness of fit.
38. Give the test statistic for 2×2 contingency tables.

IV. Give detailed answer to the following questions.

39. A random sample of 10 packets containing cashew nuts weigh (in grams) 70,120,110,101, 88,83,95,98,107,100 each. Test whether the population mean weight of 100 grams?
40. The average run of cricket player from the past records is 80. The recent scores of the player in 6 test matches are 84, 82, 83, 79, 83 and 85. Test whether the average run is more than 80?
41. The heights (in feet) of 6 rain trees in a town A are 30, 28, 29, 32, 31, 36 and that of 8 rain trees in another town B are 35, 36, 37, 30, 32, 29, 35, 30. Is there any significant difference in mean heights of rain trees?

42. Samples of two types of electric bulbs were tested for life (in hours) and the following data were obtained.

	TYPE-I	TYPE-II
Number of units	8	7
Mean of the samples (in hrs.)	1134	1024
Standard deviation of the samples (in hrs)	35	40

Test the hypothesis that the population means are equal at 5% level of significance.

43. The number of pages typed by 5 DTP – operators for 1 hour in the morning sessions are 10, 12, 13, 8, 9 and the number of pages typed by them in the afternoon are 11, 15, 12, 10, 8. Is there any significant difference in the mean number of pages typed?

44. An IQ test was conducted to 5 persons before and after they were trained. The results are given below:

Candidates	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether any change in IQ at 1% level of significance.

45. The marks secured by 9 students in Statistics and that of 12 students in Business Mathematics are given below:

Marks in Statistics	65	74	64	58	60	67	71	69	75			
Marks in Business Mathematics	52	45	59	47	53	64	58	62	54	61	57	48

Test whether the mean marks obtained by the students in Statistics and Business mathematics differ significantly at 1% level of significance.

46. A test was conducted with 6 students before and after the training programme. Their marks were recorded and tabulated as shown below. Test whether the training was helpful in improving their scores.

Before training	100	160	113	122	120	105
After training	120	155	120	128	115	100

47. An experiment was conducted 144 times with tossing of four coins and the number of heads appeared at each throw are recorded.

No. of heads	0	1	2	3	4
frequency	10	34	56	36	8

Fit binomial distribution to the above data.

48. The distribution of the number of defective blades produced in a single shift in a factory over 100 shifts is given below.

Number of defective blades	0	1	2	3	4
Number of shifts	12	14	23	18	33

Test whether the number of defective blades follows a Poisson distribution with mean = 0.44. Use $\alpha = 0.05$.

49. The quality grade of electric components produced in two factories is given in the table given below.

Factory	Quality of grade				Total
	Poor	Medium	Good	Excellent	
A	136	165	151	148	600
B	31	58	55	36	180
Total	167	223	206	184	780

Test whether there is any association between factories and quality of grades.

50. The eyesight was tested among 2000 randomly selected patients from a city and the following details are obtained.

Gender	Eye-sight		Total
	Poor	Good	
Male	620	380	1000
Female	550	450	1000
Total	1170	830	2000

Can we conclude that there is an association between gender and quality of eye-sight at 5% level of significance?

51. The weights (in kg) of 10 students from a school are 38,40,45,53,47,43,55,48,52,49. Can we say that variance of the distribution of weights of all students from the above school is equal to 20 kg?

52. In a sample of 200 households in a colony, two brands of hair oils A and B are applied by 90 females. Further, 60 females and 70 males are using brand A. To test whether there is any association between the gender and brand of hair oil used, by constructing a contigency table.

ANSWERS

I. 1. d 2. c 3. d 4. a 5. c
6. a 7. a 8. c 9. d 10. b
11. c 12.d 13. a 14.c 15. c
16 a

III. 32. $t = 2.43$, reject H_0

IV. 39. $t = 10.15$, reject H_0

40. $t = 3.16$, reject H_0

41. $|t| = -1.23443$, we do not reject H_0

42. $t = 5.683$, reject H_0

43. $|t| = 1$, we do not reject H_0

44. $t = 0.8164$, we do not reject H_0

45. $t = 4.4898$,reject H_0

46. $|t| = 0.7304$, we do not reject H_0

47. $\chi_0^2 = 0.407407$, we do not reject H_0 at 5% level with 4 d.f.

48. $\chi_0^2 = 35.10855$, reject H_0 at 5% level with 4 d.f.

49. $\chi_0^2 = 370.4034$, reject H_0 at 5% level with 3 d.f.

50. $\chi_0^2 = 10.09165$, reject H_0 at 5% level with 2 d.f.

51. $\chi_0^2 = 14$, we do not reject H_0 at 5% level with 9 d.f.

52.

Gender	Hair Oil Brands		TOTAL
	A	B	
Male	70	40	110
Female	60	30	90
Total	130	70	200

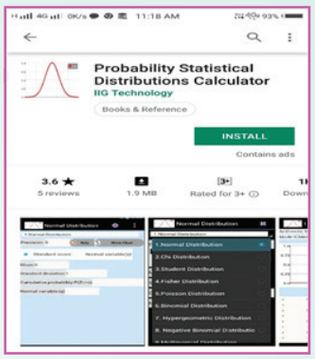
$\chi_0^2 = 0.1998$, we do not reject H_0 at 5% level with 1 d.f.

ICT CORNER

TESTS BASED ON SAMPLING DISTRIBUTIONS I

STATS IN YOUR PALM

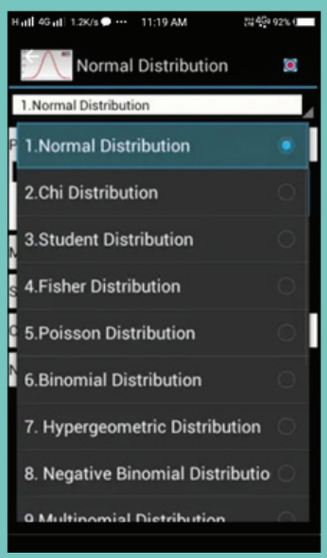
Th is activity is to calculate Chi distribution, Binomial Distribution, Students Distribution



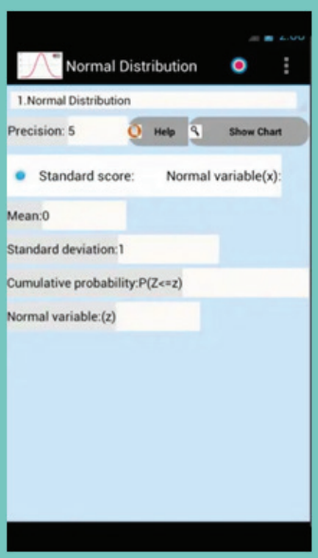
Steps:

- This is an android app activity. Open the browser and type the URL given (or) scan the QR code. (Or) search for Probability Statistical Distributions Calculator in google play store.
- (i) Install the app and open the app, (ii) click "**Menu**", (iii) In the menu page click "**Students Distribution**" menu.
- Input freedom degree and t-store, cumulative probability to get the output.

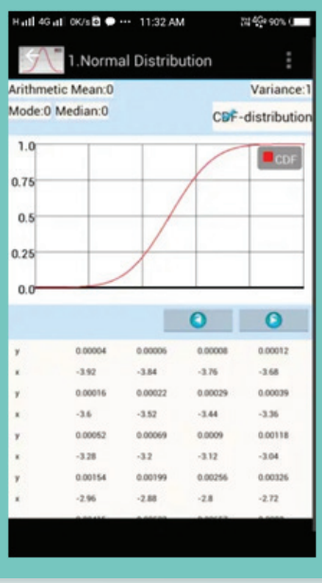
Step-1



Step-2



Step-3



Pictures are indicatives only*

URL:

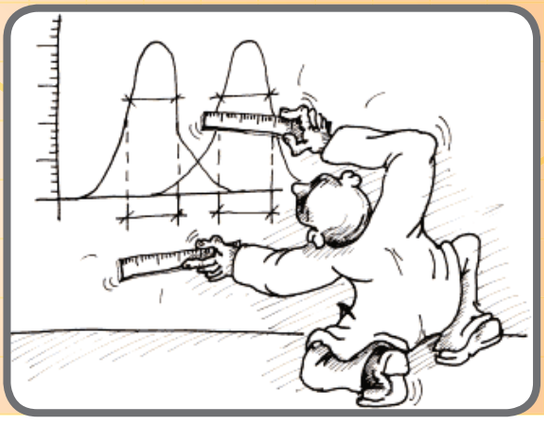
URL:http://play.google.com/store/apps/details?id=net.eaglepeak.distributions_calculator

https://www.geogebra.org/m/wfencemf

B273_12_STAT_EM

CHAPTER

3 TESTS BASED ON SAMPLING DISTRIBUTIONS – II



R. A. Fisher

Sir Ronald Aylmer Fisher (1890–1962) was a British statistician and geneticist. His work in statistics, made him popularly known as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century Statistics". In Genetics, his work used Mathematics to combine Mendelian Genetics and natural selection and this contributed to the revival of Darwinism in the early 20th century revision of the Theory of Evolution.

"Natural selection is a mechanism for generating an exceedingly high degree of improbability"

"The Best time to plan an experiment is after you have done it"

"The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic"

LEARNING OBJECTIVES

The students will be able to

- compare variances of two populations
- understand the testing of hypothesis for comparing three or more population means.
- differentiate Treatments and Blocks.
- differentiate one-way and two-way Analysis of Variance.
- calculate *F*-ratio for Treatments and Blocks.
- infer by comparing the estimated and critical values.

AP6J38

Introduction

In the previous chapters, we have discussed various concepts used in testing of hypotheses and problems relating to means of the populations. Although many practical problems involve inferences about population means or proportions, the inference about population variances is important and needs to be studied. In this chapter we will study (i) testing equality of two population variances (ii) one-way ANOVA and (iii) two-way ANOVA, using *F*-distribution.

3.1 *F*-DISTRIBUTION AND ITS APPLICATIONS

F-statistic is the ratio of two sums of the squares of deviations of observations from respective means. The sampling distribution of the statistic is *F*-distribution.

Definition: *F*-Distribution

Let X and Y be two independent χ^2 random variates with m and n degrees of freedom respectively. Then $F = \frac{X/m}{Y/n}$ is said to follow *F*-distribution with (m, n) degrees of freedom. This *F*-distribution is named after the famous statistician R.A. Fisher (1890 to 1962).

Definition: *F*-Statistic

Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be two independent random samples drawn from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ populations respectively.

Then,

$$\frac{1}{\sigma_X^2}\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2 \sim \chi^2_{m-1} \text{ and } \frac{1}{\sigma_Y^2}\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2 \sim \chi^2_{n-1}$$

are independent

(1) Hence, *F*-Statistic is defined as

$$F=\frac{(m-1)S_X^2}{\sigma_X^2}\Big/\frac{(n-1)S_Y^2}{\sigma_Y^2} \sim F_{m-1,n-1}$$

where

$$S_X^2=\frac{1}{m-1}\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2 \text{ and } S_Y^2=\frac{1}{n-1}\sum_{j=1}^{n}\left(Y_j-\overline{Y}\right)^2$$

(2) *F*-Statistic is also defined as the ratio of two mean square errors.

CARE

If the populations are not normal, *F* – test may not be used.

Assumptions for testing the ratio of two normal population variances

i) The population from which the samples were obtained must be normally distributed.

ii) The two samples must be independent of each other.

Applications of *F*-distribution

The following are some of the important applications where the sampling distribution of the respective statistic under H_0 is *F*–distribution.

(i) Testing the equality of variances of two normal populations. [Using (1)]

(ii) Testing the equality of means of k (>2) normal populations. [Using (2)]

(iii) Carrying out analysis of variance for two-way classified data. [Using (2)]

3.2 TEST OF SIGNIFICANCE FOR TWO NORMAL POPULATION VARIANCES

Test procedure:

This test compares the variances of two independent normal populations, *viz.*, $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

Step 1 : **Null Hypothesis** $H_0 : \sigma_X^2 = \sigma_Y^2$

That is, there is no significant difference between the variances of the two normal populations.

The alternative hypothesis can be chosen suitably from any one of the following

(i) $H_1 : \sigma_X^2 < \sigma_Y^2$ (ii) $H_1 : \sigma_X^2 > \sigma_Y^2$ (iii) $H_1 : \sigma_X^2 \neq \sigma_Y^2$

Step 2 : **Data**

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent samples drawn from two normal populations respectively.

Step 3 : **Level of significance** α

Step 4 : **The test Statistic**

$F = \dfrac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \dfrac{S_1^2}{S_2^2}$ under H_0 and its sampling distribution under H_0 is $F_{(m-1,\, n-1)}$.

Step 5 : **Calculation of the Test Statistic**

The test statistic $F_0 = \dfrac{s_X^2}{s_Y^2}$

Step 6 : **Critical values**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Critical value(s) f_e	$f_{(m-1,\, n-1),1-\alpha}$	$f_{(m-1,\, n-1),\, \alpha}$	$f_{(m-1,\, n-1),1-\alpha/2}$ and $f_{(m-1,\, n-1),\, \alpha/2}$

Step 7 : **Decision**

H_1	$\sigma_X^2 < \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$
Rejection Rule	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha}$	$F_0 \geq f_{(m-1,\, n-1),\, \alpha}$	$F_0 \leq f_{(m-1,\, n-1),\, 1-\alpha/2}$ or $F_0 \geq f_{(m-1,\, n-1),\, \alpha/2}$

Note 1: Since $f_{(m-1,\, n-1),\, 1-\alpha}$ is not avilable in the given F-table, it is computed as the reciprocal of $f_{(n-1,\, m-1),\alpha}$.

i.e., $f_{(m-1,\, n-1),\, 1-\alpha} = \dfrac{1}{f_{(n-1,\, m-1),\, \alpha}}$

Note 2: A F-test is based on the ratio of variances, it is also known as Variance Ratio Test.

Note 3: When μ_X and μ_Y are known, for testing the equality of variances of two normal populations, the test statistic is

$$F = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2} \text{ and follows } F_{m,n}\text{-distribution under } H_0$$

Example 3.1

Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means as 160 inches square and 91 inches square respectively. Test the hypothesis that the variances of the two populations from which the samples are drawn are equal at 10% level of significance.

Solution:

Step 1 : **Null Hypothesis:** $H_0 : \sigma_X^2 = \sigma_Y^2$

That is there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X^2 \neq \sigma_Y^2$

That is there is significant difference between the two population variances.

Step 2 : **Data**

$m = 9, n = 8$

$$\sum_{i=1}^{9}(x_i - \bar{x})^2 = 160 \qquad \sum_{j=1}^{8}(y_j - \bar{y})^2 = 91$$

Step 3 : **Level of significance**

$\alpha = 10\%$

Step 4 : **Test Statistic** $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$, under H_0.

Step 5 : **Calculation**

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 \text{ and } s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

$$s_X^2 = \frac{160}{8} = 20 \qquad s_Y^2 = \frac{91}{7} = 13$$

$$F_0 = \frac{s_X^2}{s_Y^2} = \frac{20}{13} = 1.54$$

Step 6 : **Critical values**

Since H_1 is a two-sided alternative hypothesis the corresponding critical values are:

$$f_{(8,7),0.05} = 3.73 \text{ and } f_{(8,7),0.95} = \frac{1}{f_{(7,8),0.05}} = \frac{1}{3.5} = 0.286$$

Step 7 : **Decision**

Since $f_{(8,7),0.95} = 0.286 < F_0 = 1.54 < f_{(8,7),0.05} = 3.73$, the null hypothesis is not rejected and we conclude that there is no significant difference between the two population variances.

Note 4: The critical values of F corresponding to $\alpha = 0.05$ requires table values at 0.025 and 0.975 which are not provided. Hence α is taken as 0.1 in this example.

Example 3.2

A medical researcher claims that the variance of the heart rates (in beats per minute) of smokers is greater than the variance of heart rates of people who do not smoke. Samples from two groups are selected and the data is given below. Using = 0.05, test whether there is enough evidence to support the claim.

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Solution:

Step 1 : **Null Hypothesis:** $H_0 : \sigma_1^2 = \sigma_2^2$

That is there is no significant difference between the two population variances.

$H_1 : \sigma_1^2 > \sigma_2^2$

That is, the variance of heart rates of smokers is greater than that of non-smokers.

Step 2 : **Data**

Smokers	Non Smokers
$m = 25$	$n = 18$
$s_1^2 = 36$	$s_2^2 = 10$

Step 3 : **Level of significance** $\alpha = 5\%$

Step 4 : **Test statistic**

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2}$$

Step 5 : **Calculation**

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 6 : **Critical value**

$$f_{(m\text{-}1,n\text{-}1),0.05} = f_{(24,17),0.05} = 2.19$$

Step 7 : **Decision**

Since $F_0 = 3.6 > f_{(24,17),0.05} = 2.19$, the null hypothesis is rejected and we conclude that the variance of heart beats for smokers seems to be considerably higher compared to that of the non-smokers.

Example 3.3

The following table gives the random sample of marks scored by students in two schools, A and B.

School A	63	72	80	60	85	83	70	72	81
School B	86	93	64	82	81	75	86	63	63

Is the variance of the marks of students in school A is less than that of those in school B? Test at 5% level of significance.

Solution:

Let $X_1, X_2, \ldots, X_m$ represent sample values for school A and let $Y_1, Y_2, \ldots, Y_n$ represent sample values for school B.

Step 1 : **Null Hypothesis:** $H_1 : \sigma_X^2 = \sigma_Y^2$

That is, there is no significant difference between the two population variances.

Alternative Hypothesis: $H_1 : \sigma_X^2 < \sigma_Y^2$

That is, the variance of marks in school A is significantly less than that of school B.

Step 2 : **Data**

$X_1, X_2, \ldots, X_m$ are sample from school A

$Y_1, Y_2, \ldots, Y_n$ are sample from school B

Step 3 : **Test statistic**

$$F = \frac{s_X^2}{s_Y^2}$$

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

$$s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

Step 4 : **Calculations**

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
63	-11	121	86	9	81
72	-2	4	93	16	256
80	6	36	64	-13	169
60	-14	196	82	5	25
85	11	121	81	4	16
83	9	81	75	-2	4
70	-4	16	86	9	81
72	-2	4	63	-14	196
81	7	49	63	-14	196
666		628	693		1024

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m} = \frac{666}{9} = 74$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{693}{9} = 77$$

$$s_X^2 = \frac{1}{9-1} \times 628 = \frac{1}{8} \times 628 = 78.5$$

$$s_Y^2 = \frac{1}{9-1} \times 1024 = \frac{1}{8} \times 1024 = 128$$

$$F_0 = \frac{78.5}{128.8} = 0.609$$

Step 5 : Level of significance

$\alpha = 5\%$

Step 6 : Critical value

$$f_{(9\text{-}1,9\text{-}1),0.95} = \frac{1}{f_{(8,8),0.05}} = \frac{1}{3.44} = 0.291$$

Step 7 : Decision

Since $F_0 = 0.609 > f_{(8,8),0.95} = 0.291$, the null hypothesis is not rejected and we conclude that in school B there seems to be more variance present than in school A.

3.3 ANALYSIS OF VARIANCE (ANOVA)

In chapter 2, testing equality means of two normal populations based on independent small samples was discussed. When the number of populations is more than 2, those methods cannot be applied.

ANOVA is used when we want to test the equality of means of more than two populations. For example, through ANOVA, one may compare the average yield of several varieties of a crop or average mileages of different brands of cars.

ANOVA cannot be used in all situations and for all types of variables. It is based on certain assumptions, and they are listed below:

1. The observations follow normal distribution.
2. The samples are independent.
3. The population variances are equal and unknown.

According to R.A. Fisher ANOVA is the "Separation of variance, ascribable to one group of causes from the variance ascribable to other groups".

The data may be classified with respect to different levels of a single factor/or different levels of two factors.

The former is called one-way classified data and the latter is called two-way classified data. Applications of ANOVA technique to these kinds of data are discussed in the following sections.

3.3.1 One-way ANOVA

ANOVA is a statistical technique used to determine whether differences exist among three or more population means.

In one-way ANOVA the effect of one factor on the mean is tested. It is based on independent random samples drawn from k – different levels of a factor, also called treatments.

The following notations are used in one-way ANOVA. The data can be represented in the following tabular structure.

Data representation for one-way ANOVA

Treatments					Total
Treatment 1	x_{11}	x_{12}	…	x_{1n_1}	$x_{1.}$
Treatment 2	x_{21}	x_{22}	…	x_{2n_2}	$x_{2.}$
⋮	⋮	⋮	…	⋮	⋮
Treatment k	x_{k1}	x_{k2}	…	x_{kn_k}	$x_{k.}$

x_{ij} - the j^{th} sample value from the ith treatment, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$

k - number of treatments compared.

$x_{i.}$ - the sample total of i^{th} treatment.

n_i - the number of observations in the i^{th} treatment.

$$\sum_{i=1}^{k} n_i = n$$

The total variation in the observations x_{ij} can be split into the following two components

i) variation between the levels or the variation due to different bases of classification, commonly known as treatments.

ii) The variation within the treatments *i.e.* inherent variation among the observations within levels.

Causes involved in the first type of variation are known as assignable causes. The causes leading to the second type of variation are known as chance or random causes.

The first type of variation that is due to assignable causes, can be detected and controlled by human endeavor and the second type of variation that is due to chance causes, is beyond the human control.

3.3.2 Test Procedure

Let the observations x_{ij}, $j = 1, 2, \ldots, n_i$ for treatment i, be assumed to come from $N(\mu_i, \sigma^2)$ population, $i = 1, 2, \ldots, k$ where σ^2 is unknown.

Step 1 : Framing Hypotheses

Null Hypothesis $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

That is, there is no significant difference among the population means of k treatments.

Alternative Hypothesis

$H_1: \mu_i \neq \mu_j$ for atleast one pair (i,j); $i, j = 1,2,\ldots,k$; $i \neq j$

That is, at least one pair of means differ significantly.

Step 2 : Data

Data is presented in the tabular form as described in the previous section

Step 3 : Level of significance : α

Step 4 : Test Statistic

$F = \dfrac{MST}{MSE}$ which follows $F_{(k\text{-}1,\, n\text{-}k)}$, under H_0

To evaluate the test statistic we compute the following:

(i) Correction factor: $C.F = \dfrac{G^2}{n}$ where $G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$

(ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C.F$

(iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \dfrac{x_{i.}^2}{n_i} - C.F$,

where $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$, $i = 1,2,\ldots,k$

(iv) Sum of Squares due to Error: $SSE = TSS - SST$

Degrees of Freedom (d.f)

Degrees of freedom (d.f.)		d.f.
Total Sum of Squares	Total no. of observations –1	n–1
Treatment Sum of Squares	Total no. of observations –1	k–1
Error of Sum Squares	Total no. of observations –1	n–k

Mean Sum of Squares

Mean Sum of Squares due to treatment: $MST = \dfrac{SST}{k-1}$

Mean Sum of Squares due to Error:

$$MSE = \frac{SSE}{n-k}$$

Step 5 : Calculation of Test statistic

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	k-1	$MST = \frac{SST}{k-1}$	$F_0 = \frac{MST}{MSE}$
Error	SSE	n-k	$MSE = \frac{SSE}{n-k}$	
Total	TSS	n-1		

Step 6 : Critical value

$f_e = f_{(k\text{-}1,\, n\text{-}k),\alpha}$.

Step 7 : Decision

If $F_0 < f_{(k\text{-}1,\, n\text{-}k),\alpha}$ then reject H_0.

3.3.3 Merits and Demerits of One-Way ANOVA

Merits

- Layout is very simple and easy to understand.
- Gives maximum degrees of freedom for error.

Demerits

- Population variances of experimental units for different treatments need to be equal.
- Verification of normality assumption may be difficult.

Example 3.4

Three different techniques namely medication, exercises and special diet are randomly assigned to (individuals diagnosed with high blood pressure) lower the blood pressure. After four weeks the reduction in each person's blood pressure is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among the three techniques.

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Solution:

Step 1 : Hypotheses

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$

That is, there is no significant difference among the three groups on the average reduction in blood pressure.

Alternative Hypothesis: H_1: $\mu_i \neq \mu_j$ for atleast one pair (i, j); $i, j = 1, 2, 3$; $i \neq j$.

That is, there is significant difference in the average reduction in blood pressure in atleast one pair of treatments.

Step 2 : Data

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Step 3 : Level of significance $\alpha = 0.05$

Step 4 : Test statistic

$$F_0 = \frac{MST}{MSE}$$

Step 5 : Calculation of Test statistic

						Total	**Square**
Medication	10	12	9	15	13	59	3481
Exercise	6	8	3	0	2	19	361
Diet	5	9	12	8	4	38	1444
						G = 116	5286

Individual squares

Medication	100	144	81	225	169
Exercise	36	64	9	0	4
Diet	25	81	144	64	16

$$\sum\sum x_{ij}^2 = 1162$$

1. Correction Factor: $CF = \frac{G^2}{n} = \frac{(116)^2}{15} = \frac{13456}{15} = 897.06$

2. Total Sum of Squares: $TSS = = \sum\sum x_{ij}^2 - C.F$

$= 1162 - 897.06 = 264.94$

3. Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \frac{5286}{5} - 897.06$$
$$= 1057.2 - 897.06$$
$$= 160.14$$

4. Sum of Squares due to Error: $SSE = TSS - SST$

$= 264.94 - 160.14 = 104.8$

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments	160.14	3 – 1 = 2	80.07	$F_o = \frac{80.07}{8.73} = 9.17$
Error	104.8	12	8.73	
Total	264.94	$n - 1 = 15 - 1 = 14$		

Step 6 : Critical value

$f_{(2, 12),0.05} = 3.8853$.

Step 7 : Decision

As $F_0 = 9.17 > f_{(2, 12),0.05} = 3.8853$, the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in atleast one pair of techniques.

Example 3.5

Three composition instructors recorded the number of spelling errors which their students made on a research paper. At 1% level of significance test whether there is significant difference in the average number of errors in the three classes of students.

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Solution:

Step 1 : Hypotheses

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

That is there is no significant difference among the mean number of errors in the three classes of students.

Alternative Hypothesis

$H_1 : \mu_i \neq \mu_j$ for at one pair (i, j); $i,j = 1,2,3$; $i \neq j$.

That is, atleast one pair of groups differ significantly on the mean number of errors.

Step 2 : Data

Instructor 1	2	3	5	0	8		
Instructor 2	4	6	8	4	9	0	2
Instructor 3	5	2	3	2	3	3	

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_0 = \frac{MST}{MSE}$$

Step 5 : Calculation of Test statistic

								Total	Square
Instructor 1	2	3	5	0	8			18	324
Instructor 2	4	6	8	4	9	0	2	33	1089
Instructor 3	5	2	3	2	3	3		18	324
								69	

Individual squares

Instructor 1	4	9	25	0	64		
Instructor 2	16	36	64	16	81	0	4
Instructor 3	25	4	9	4	9	9	

$$\sum\sum x_{ij}^2 = 379$$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(69)^2}{18} = \frac{4761}{18} = 264.5$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$= 379 - 264.5 = 114.5$

Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$= \left(\frac{324}{5} + \frac{1089}{7} + \frac{324}{6}\right) - 264.5$

$= (64.8 + 155.6 + 54) - 264.5$

$= (274.4) - 264.5$

$= 9.9$

Sum of Squares due to Error: $SSE = TSS - SST$

$= 114.5 - 9.9$

$= 104.6$

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments	9.9	$3 - 1 = 2$	$\frac{9.9}{2} = 4.95$	$F_0 = \frac{4.95}{6.97} = 0.710$
Error	104.6	15	$\frac{104.6}{15} = 6.97$	
Total		$n - 1 = 18 - 1 = 17$		

Step 6 : Critical value

The critical value = $f_{(15, 2),0.05} = 3.6823$.

Step 7 : Decision

As $F_0 = 0.710 < f_{(15, 2),0.05} = 3.6823$, null hypothesis is not rejected. There is no enough evidence to reject the null hypothesis and hence we conclude that the mean number of errors made by these three classes of students are not equal.

3.4 TWO-WAY ANOVA

In two-way ANOVA a study variable is compared over three or more groups, controlling for another variable. The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block. The accuracy of the test in two-way ANOVA is considerably higher than that of the one-way ANOVA, as the additional factor, block is used to reduce the error variance.

In two-way ANOVA, the data can be represented in the following tabular form.

Blocks

Groups or Treatments	1	2	3	…	m	$x_{i.}$
1	x_{11}	x_{12}	$x_{13.}$	⋯	x_{1m}	$x_{1.}$
2	x_{21}	x_{22}	$x_{2.}$	⋯	x_{2m}	$x_{2.}$
3	x_{31}	x_{32}	$x_{3.}$	⋯	x_{3m}	$x_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	x_{k1}	x_{k2}	x_{k3}	⋯	x_{km}	$x_{k.}$
$x_{.j}$	$x_{.1}$	$x_{.2}$	$x_{.3}$	⋯	$x_{.m}$	G

We use the following notations.

x_{ij} - i^{th} treatment value from the j^{th} block, $i = 1,2, ..., k; j = 1,2, ..., m$.

The i^{th} treatment total - $x_{i.} = \sum_{j=1}^{m} x_{ij}$, $i = 1, 2, ..., k$

The j^{th} block total - $x_{.j} = \sum_{i=1}^{k} x_{ij}$, $j = 1, 2, ..., m$

Note that, $k \times m = n$, where m = number of blocks, and k = number of treatments (groups) and n is the total number of observations in the study.

The total variation present in the observations x_{ij} can be split into the following three components:

i) The variation between treatments (groups)

ii) The variation between blocks.

iii) The variation inherent within a particular setting or combination of treatment and block.

3.4.1 Test Procedure

Steps involved in two-way ANOVA are:

Step 1 : In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.

Framing Hypotheses

Null Hypotheses

H_{01}: There is no significant difference among the population means of different groups (Treatments)

H_{02}: There is no significant difference among the population means of different Blocks

Alternative Hypotheses

H_{11}: Atleast one pair of treatment means differs significantly

H_{12}: Atleast one pair of block means differs significantly

Step 2 : **Data** is presented in a rectangular table form as described in the previous section.

Step 3 : **Level of significance** α.

Step 4 : **Test Statistic**

$$F_{0t}(\text{treatments}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{j=1}^{m}\sum_{i=1}^{k} x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 - C.F$

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \frac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^{m} \frac{x_{.j}^2}{k} - C.F$

v) Sum of Squares due to Error: $SSE = TSS\text{-}SST\text{-}SSB$

vi) Degrees of freedom

Degrees of freedom (d.f.)	d.f.
Total Sum of Squares	$n-1$
Treatment Sum of Squares	$k-1$
Block Sum of Squares	$m-1$
Error of Sum Squares	$(m-1)(k-1)$

vii) Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \frac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \frac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \frac{SSE}{(k-1)(m-1)}$

Step 5 : Calculation of the Test Statistic

ANOVA Table (two-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	k-1	MST	$F_{0t} = \frac{MST}{MSE}$
Blocks	SSB	m-1	MSB	$F_{0b} = \frac{MSB}{MSE}$
Error	SSE	$(k\text{-}1)(m\text{-}1)$	MSE	
Total	TSS	n-1		

Step 6 : Critical values

Critical value for treatments $= f_{(k\text{-}1,(m\text{-}1)(k\text{-}1)),\alpha}$

Critical value for blocks $= f_{(m\text{-}1,\,(m\text{-}1)(k\text{-}1)),\alpha}$

Step 7 : Decision

For Treatments: If the calculated F_{0t} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the treatment means, in atleast one pair.

For Blocks: If the calculated F_{0b} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the block means, in at least one pair.

3.4.2 Merits and Demerits of two-way ANOVA

Merits

- Any number of blocks and treatments can be used.
- Number of units in each block should be equal.
- It is the most used design in view of the smaller total sample size since we are studying two variable at a time.

Demerits

- If the number of treatments is large enough, then it becomes difficult to maintain the homogeneity of the blocks.
- If there is a missing value, it cannot be ignored. It has to be replaced with some function of the existing values and certain adjustments have to be made in the analysis. This makes the analysis slightly complex.

Comparison between one-way ANOVA and two-way ANOVA

Basis of comparison	ANOVA	
	One-way	Two-way
Independent variable	One	Two
Compares	Three or more levels of one factor	Three or more levels of two factors, simultaneously
Number of observations	Need not be same in each treatment group	Need to be equal in each treatment group

Example 3.6

A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table.

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Test whether there is significant difference among methods and among salesmen.

Solution:

Step 1 : Hypotheses

Null Hypotheses: $H_{01}: \mu_{M_1} = \mu_{M_2} = \mu_{M_3}$ (for treatments)

That is, there is no significant difference among the three programs in their mean sales.

$$H_{02}: \mu_{S_1} = \mu_{S_2} = \mu_{S_3} = \mu_{S_4} \text{ (for blocks)}$$

Alternative Hypotheses:

H_{11} : At least one average is different from the other, among the three programs.

H_{12} : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

Step-5 : Calculation of the Test Statistic

	Methods			Total $x_{i.}$	$x_{i.}^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	256
4	7	5	4	16	256
x_i	22	28	16	66	1140
$x_{i.}^2$	484	784	256	1524	

Squares

16	36	4
36	100	36
25	49	16
49	25	16
		$\sum\sum x_{ij}^2 = 408$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$= 408 - 363 = 45$

Sum of Squares due to Treatments: $SST = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \frac{1140}{3} - 363$

$= 380 - 363 = 17$

Sum of Squares due to Blocks: $SSB = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \frac{1524}{4} - 363$

$= 381 - 363$

$= 18$

Sum of Squares due to Error: $SSE = TSS - SST - SSB$

$= 45 - 17 - 18 = 10$

Mean sum of squares: $MST = \frac{SST}{k-1} = \frac{17}{3} = 5.67$

$MSB = \frac{SSB}{m-1} = \frac{18}{2} = 9$

$MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between treatments (Programs)	17	3	5.67	$F_{ot} = \frac{5.67}{1.67} = 3.40$
Between blocks (Salesmen)	18	2	9	$F_{ob} = \frac{9}{1.67} = 5.39$
Error	10	6	1.67	
Total		11		

Step 6 : Critical values

(i) $f_{(3, 6),0.05} = 4.7571$ (for treatments)

(ii) $f_{(2, 6),0.05} = 5.1456$ (for blocks)

Step 7 : Decision

i) Calculated $F_{0t} = 3.40 < f_{(3,\ 6),0.05} = 4.7571$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

ii) Calculate $F_{0b} = 5.39 > f_{(2,\ 6),0.05} = 5.1456$, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

Example 3.7

The illness caused by a virus in a city concerning some restaurant inspectors is not consistent with their evaluations of cleanliness of restaurants. In order to investigate this possibility, the director has five restaurant inspectors to grade the cleanliness of three restaurants. The results are shown below.

Inspectors	Restaurants		
	I	II	III
1	71	55	84
2	65	57	86
3	70	65	77
4	72	69	70
5	76	64	85

Carry out two-way ANOVA at 5% level of significance.

Solution:

Step 1 :

Null hypotheses

$H_{0I} : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (For inspectors - Treatments)

That is, there is no significant difference among the five inspectors over their mean cleanliness scores

$H_{0R} : \mu_I = \mu_{II} = \mu_{III}$ (For restaurants - Blocks)

That is, there is no significant difference among the three restaurants over their mean cleanliness scores

Alternative Hypotheses

H_{1I}: At least one mean is different from the other among the Inspectors

H_{1R}: At least one mean is different from the other among the Restaurants.

Step 2 : Data

Inspectors	Restaurants		
	I	II	III
1	71	55	84
2	65	57	86
3	70	65	77
4	72	69	70
5	76	64	85

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

For inspectors: F_{0a} (treatments) = $= \frac{MST}{MSE}$

For restaurants: F_{0b} (blocks) $= \frac{MSB}{MSE}$

Step-5 : Calculation of the Test Statistic

Inspectors	Restaurants			Total $x_{i.}$	$x_{i.}^2$
	I	II	III		
1	71	55	84	210	44100
2	65	57	86	208	43264
3	70	65	77	212	44944
4	72	69	70	211	44521
5	76	64	85	225	50625
$x_{.j}$	354	310	402	1066	
$x_{.j}^2$	125316	96100	161604		

Squares

5041	3025	7056
4225	3249	7396
4900	4225	5929
5184	4761	4900
5776	4096	7225
		$\sum\sum x_{ij}^2 = 76988$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(1066)^2}{15} = \frac{1136356}{15} = 75757.07$

Total Sum of Squares: $TSS = \sum\sum x_{ij}^2 - C.F$

$= 76988 - 75757.07 = 1230.93$

Sum of Squares due to Treatments: $SST = \dfrac{\sum_{j=1}^{l} x_{i.}^2}{l} - C.F$

$= \dfrac{227454}{3} - 75757.07$

$= 75818 - 75757.07$

$= 60.93$

Sum of Squares due to Blocks: $SSB = \dfrac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$

$= \dfrac{383020}{5} - 75757.07$

$= 76604 - 75757.07$

$= 846.93$

Sum of squares due to error: $SSE = TSS - SST - SSB$

$= 1230.93 - 60.93 - 846.93$

$= 323.07$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	*F*-ratio
Between inspectors	60.93	4	15.23	$F_{0I} = \dfrac{15.23}{40.38} = 0.377$
Between restaurants	846.93	2	423.47	$F_{0R} = \dfrac{423.47}{40.38} = 10.49$
Error	323.07	8	40.38	
Total	1230.93	14		

Step 6 : Critical values

(i) $f_{(4, 8),0.05} = 3.838$ (for inspectors)

(ii) $f_{(2, 8),0.05} = 4.459$ (for restaurants)

Step 7 : Decision

i) As $F_{0I} = 0.377 < f_{(4, 8),0.05} = 3.838$, the null hypothesis is not rejected and we conclude that there is no significant difference among the mean cleanliness scores of inspectors.

ii) As $F_{0R} = 10.49 > f_{(2, 8),0.05} = 4.459$, the null hypothesis is rejected and we conclude that there exists significant difference in atleast one pair of restaurants over their mean cleanliness scores.

POINTS TO REMEMBER

- F-statistic is the ratio of two independent sample variances
- If X and Y are two independent χ^2 variates with m and n degrees of freedom respectively, then $F = \frac{X/m}{Y/n}$ is said to follow F distribution with (m, n) degrees of freedom.
- Two independent random samples of size m and n are taken from Normal populations. Then the statistic $F = \frac{s_X^2}{s_Y^2}$ is a random variable following the F-distribution with $m-1$ and $n-1$ degrees of freedom.
- According to R.A. Fisher, ANOVA is the "Separation of variance, ascribable to one group of causes from the variance ascribable to other groups".
- One-way ANOVA is used to compare means in more than two groups.
- Two-way ANOVA is used to compare means in more than two groups, controlling another variable.
- Assumptions required for ANOVA are:
  - The observations follow normal distribution.
  - Experimental units assigned to treatments are random.
  - The sample observations are independent.
  - The population variances of the groups are unknown but are assumed to be equal.

EXERCISE 3

CE136V

I. Choose the best Answer.

1. ANOVA was developed by

   (a) S.D. Poisson (b) Karl Pearson

   (c) R.A. Fisher (d) W.S. Gosset

2. ANOVA technique originated in the field of

   (a) Industry (b) Agriculture

   (c) Medicine (d) Genetics

3. One of the assumptions of ANOVA is that the population from which the samples are drawn is

   (a) Binomial (b) Poisson (c) Chi-square (d) Normal

4. In one-way classification the total variation can be split into

   (a) Two components (b) Three components

   (c) Four components (d) Only one components

5. The null hypothesis in the ANOVA is that all the population means are

(a) Equal (b) Variable

(c) Unequal (d) none of the above

6. In one-way classification with 30 observation and 5 treatments the degrees of freedom for error is

(a) 29 (b) 4 (c) 25 (d) 150

7. In two-way classification the total variation *TSS* is

(a) *SST*+*SSB*+*SSE* (b) *SST*-*SSB*+*SSE*

(c) *SST*+*SSB*-*SSE* (d) *SST*+*SSB*

8. In two-way classification if *TSS* = 210, *SST* = 32, *SSB* = 42 then *SSE* =

(a) 126 (b) 74 (c) 136 (d) 178

9. In two-way classification with 5 treatments and 4 blocks the degrees of freedom due to error is

(a) 12 (b) 19 (c) 16 (d) 15

10. The formula for comparing three or more means in one-way analysis of variance is

(a) $F = \frac{MST}{MSE}$ (b) $F = \frac{TSS}{SST}$

(c) $F = \frac{MSB}{MST}$ (d) $F = \frac{MST}{MSB}$

11. _____ test is used to compare three or more means.

(a) t (b) χ^2 (c) F (d) Z

12. When there is no significant difference among three or more means the value of F will be close to

(a) 0 (b) -1 (c) 1 (d) ∞

13. *F*-test is also called as

(a) mean ratio test (b) variance ratio test

(c) variance test (d) standard deviation ratio test

14. The Analysis of Variance procedure is appropriate for testing the equivalence of three or more population

(a) variances (b) proportions

(c) means (d) observations

15. In two-way classification with 'm' treatments and 'n' blocks the degrees of freedom due to error is

(a) mn-1 (b) m-1 (c) n-1 (d) (m-1)(n-1)

16. If the calculated value of F is greater than the critical value at the given level of significance then the H_0 is

(a) Rejected (b) Not rejected

(c) Always true (d) Sometimes true

17. _____ and _____ causes are present in Analysis of Variance techniques

(a) Chance, error (b) Fixed, block

(c) Assignable, chance (d) Assignable, fixed

18. In ANOVA, the sample observations are

(a) dependent (b) independent

(c) equal (d) unequal

19. The correction factor is _____ in ANOVA (with the usual notations).

(a) $\frac{\sum T_{ij}^2}{n}$ (b) $\frac{\sum T_{i.}^2}{n}$

(c) $\frac{G^2}{n}$ (d) $\frac{\sum T_{i.}}{n}$

20. Mean Sum of Squares is the ratio of Sum of Squarea to its

(a) number of blocks (b) number of treatments

(c) degrees of freedom (d) total sum of squares

II. Give very short answers to the following questions.

21. What is Analysis of Variance?
22. Write the applications of F-statistic.
23. What are the assumptions of ANOVA?
24. Define: Between group variance and within group variance.
25. State the hypotheses used in one-way ANOVA.
26. What are the components in a two-way ANOVA?
27. Name the causes of variation?

III. Give short answer to the following questions.

28. What are the merits and demerits of one-way classification?
29. Write the model ANOVA table for one-way classification.
30. What are the values to be found for finding the test statistic in one-way classification?
31. What are the merits and demerits of two-way classification?
32. Write the model ANOVA table for two-way classification.
33. What are the components in two-way ANOVA?
34. What are the values to be found for finding the test statistic in two way classification?
35. Compare one-way and two-way ANOVA.

IV. Give detailed answer to the following questions.

36. In a sample of 8 observations, the sum of the squares of deviations of the sample values from its sample mean was 84.4. In another sample of 10 observations it was 102.6. Test whether the two population variances are equal at 5% level.
37. Two random samples gave the following results

Sample	Size	Sample mean	Sum of squares of deviations from the mean
I	10	15	90
II	12	14	108

Test whether the populations have same variances at 5% level of significance.

38. The following data refer to the yield of wheat in quintals on plots of equal area in two agricultural blocks A and B.

	Number of plots	Mean yield	Sample variance
Block A	8	60	50
Block B	6	51	40

Is the variance of yield for block A is greater than that of block B at 5% level of significance.

39. The calories contained in 1/2 cup servings of ice-creams selected randomly from two national brands are listed here. At 5% level of significance, is there sufficient evidence to conclude that the variance of calories is less for brand A than brand B?

Brand A	330	310	300	310	300	350	380	300	300
Brand B	300	300	270	290	310	370	300	310	250

40. The carbohydrates contained in servings of some randomly selected chocolate and non-chocolate candies are listed below. Is there sufficient evidence to conclude that the variance in carbohydrates varies between chocolate and non-chocolate candies? Use = 2%.

Chocolate	29	25	18	40	41	25	32	30	38	34	25	28
Non-chocolate	39	39	37	29	30	38	39	10	29	55	29	

41. A home gardener wishes to determine the effects of four fertilizers on the average number of tomatoes produced. Test at 5% level of significance the hypothesis that the fertilizers A, B, C and D have equal average yields.

A	14	10	12	16	17
B	9	11	12	8	10
C	16	15	14	10	18
D	10	11	11	13	8

42. Three processes *X*, *Y* and *Z* are tested to see whether their outputs are equivalent. The following observations on outputs were made.

X	10	13	12	11	10	14	15	13
Y	9	11	10	12	13			
Z	11	10	15	14	12	13		

Carry out the one-way analysis of variance and state your conclusion.

43. A test was given to five students taken at random from XII class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the one-way ANOVA.

44. A farmer applies three types of fertilizers on four separate plots. The figures on yield per acre are tabulated below.

Fertilizer	Plots			
	A	B	C	D
Nitrogen	6	4	8	6
Potash	7	6	6	9
Phosphate	8	5	10	9

Test whether there is any significant difference among mean yields of different plots and among different fertilizers.

45. Operators are tested for their efficiency in terms of number of units produced per day by five different types of machines. Test at 5% level of significance whether the operators and machines differ in terms of their efficiency?

Operators	Machine types				
	A	B	C	D	E
I	8	10	7	12	6
II	12	13	8	9	12
III	7	8	6	8	8
IV	5	5	3	5	14

ANSWERS

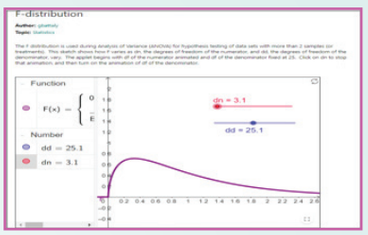
I. 1. (c) 2. (b) 3. (d) 4. (a) 5. (a)
6. (c) 7. (a) 8. (c) 9. (a) 10. (a)
11. (c) 12. (c) 13. (b) 14. (c) 15. (d)
16. (a) 17. (c) 18. (b) 19. (c) 20. (c)

III. 36. $F_0 = 1.06$, H_0 is not rejected
37. $F_0 = 1.02$, H_0 is not rejected
38. $F_0 = 1.25$, H_0 is not rejected
39. $F_0 = 1.34$, H_0 is not rejected
40. $F_0 = 2.52$, H_0 is not rejected
41. $F_0 = 4.59$, H_0 is rejected
42. $F_0 = 1.097$, H_0 is not rejected
43. $F_0 = 3.33$, H_0 is not rejected
44. F_0 =2.39, H_0 is not rejected $F_0 = 3.59$, H_0 is not rejected
45. $F_0 = 2.53$, H_0 is not rejected $F_0 = 1.24$, H_0 is not rejected

ICT CORNER

TESTS BASED ON SAMPLING DISTRIBUTIONS – II

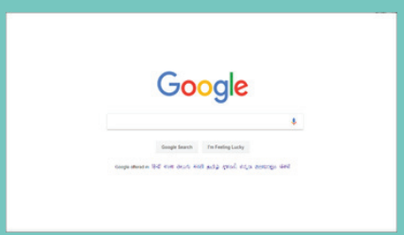
This activity helps to understand about *F*-DISTRIBUTION



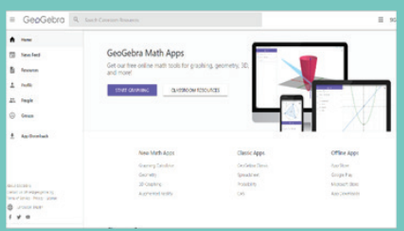
Steps:

- Open the browser and type the URL given (or) scan the QR code. GeoGebra work book called "***F*-Distribution**" will appear.
- In this several work sheets for statisticsare given, open the worksheet named "*F*-Distribution"
- Drag and move the Red colour and Blue colour button or type the values in the left side box for result

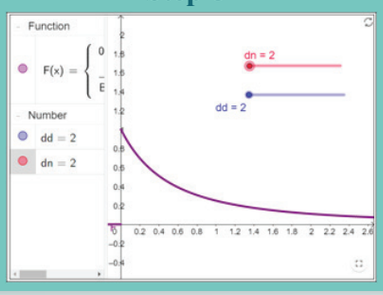
Step-1



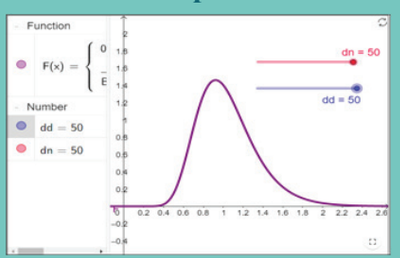
Step-2



Step-3



Step-4



Pictures are indicatives only*

URL:

https://www.geogebra.org/m/A45YdMfJ

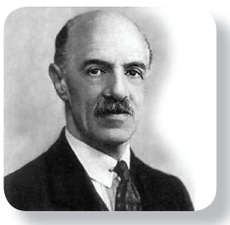
B273_12_STAT_EM

CHAPTER

4 CORRELATION ANALYSIS

Karl Pearson

Karl Pearson (1857-1936) was a English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. The linear correlation coefficient is also called Pearson product moment correlation coefficent. It was developed by Karl Pearson with a related idea by Francis Galton (see Regression analysis - for Galton's contribution). It is the first of the correlation measures developed and commonly used.



Charles Spearman

Charles Edward Spearman (1863-1945) was an English psychologist and ,after serving 15 years in Army he joined to study PhD in Experimental Psychology and obtained his degree in 1906. Spearman was strongly influenced by the work of Galton and developed rank correlation in 1904.He also pioneered factor analysis in statistics.

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering the existence of relation and measuring the intensity of relationship is known as correlation"

—CROXTON AND COWDEN

LEARNING OBJECTIVES

DNBR97

The student will be able to

- learn the meaning, definition and the uses of correlation.
- identify the types of correlation.
- understand correlation coefficient for different types of measurement scales.
- differentiate different types of correlation using scatter diagram.
- calculate Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient and Yule's coefficient of association.
- interpret the given data with the help of coefficient of correlation.

Introduction

"Figure as far as you can, then add judgment"

The statistical techniques discussed so far are for ***only one variable***. In many research situations one has to consider two variables simultaneously to know whether these ***two variables*** are related linearly. If so, what type of relationship that exists between them. This leads to bivariate (two variables) data analysis namely correlation analysis. If two quantities vary in such a way that movements (upward or downward) in one are accompanied by the movements(upward or downward) in the other, these quantities are said to be co-related or correlated.

The correlation concept will help to answer the following types of questions.

- Whether study time in hours is related with marks scored in the examination?
- Is it worth spending on advertisement for the promotion of sales?
- Whether a woman's age and her systolic blood pressure are related?
- Is age of husband and age of wife related?
- Whether price of a commodity and demand related?
- Is there any relationship between rainfall and production of rice?

4.1 DEFINITION OF CORRELATION

Correlation is a statistical measure which helps in analyzing the interdependence of two or more variables. In this chapter the dependence between only two variables are considered.

1. **A.M. Tuttle** defines correlation as:

   "An analysis of the co-variation of two or more variables is usually called correlation"

2. **Ya-kun-chou** defines correlation as:

   "The attempts to determine the degree of relationship between variables".

Correlation analysis is the process of studying the strength of the relationship between two related variables. High correlation means that variables have a strong linear relationship with each other while a low correlation means that the variables are hardly related. The type and intensity of correlation is measured through the correlation analysis. The measure of correlation is the correlation coefficient or correlation index. It is an absolute measure.

Uses of correlation

- Investigates the type and strength of the relationship that exists between the two variables.
- Progressive development in the methods of science and philosophy has been characterized by the rich knowledge of relationship.

4.2 TYPES OF CORRELATION

1. ***Simple (Linear) correlation*** (2 variables only) : The correlation between the given two variables. It is denoted by r_{xy}
2. ***Partial correlation (more than 2 variables):*** The correlation between any two variables while removing the effect of other variables. It is denoted by $r_{xy.z\ldots}$

3. ***Multiple correlation (more than 2 variables) :*** The correlation between a group of variables and a variable which is not included in that group. It is denoted by $R_{y.(xz...)}$

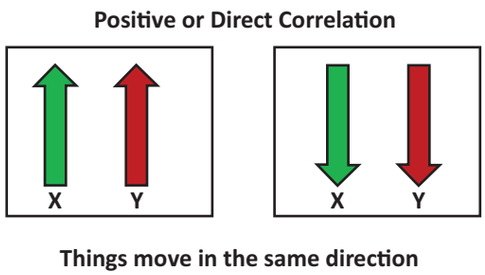
In this chapter, we study simple correlation only, multiple correlation and partial correlation involving three or more variables will be studied in higher classes .

4.2.1 Simple correlation or Linear correlation

Here, we are dealing with data involving two related variables and generally we assign a symbol 'x ' to scores of one variable and symbol 'y ' to scores of the other variable. There are five types in simple correlation. They are

1. Positive correlation (Direct correlation)
2. Negative correlation (Inverse correlation)
3. Uncorrelated
4. Perfect positive correlation
5. Perfect negative correlation

1) Positive correlation: (Direct correlation)

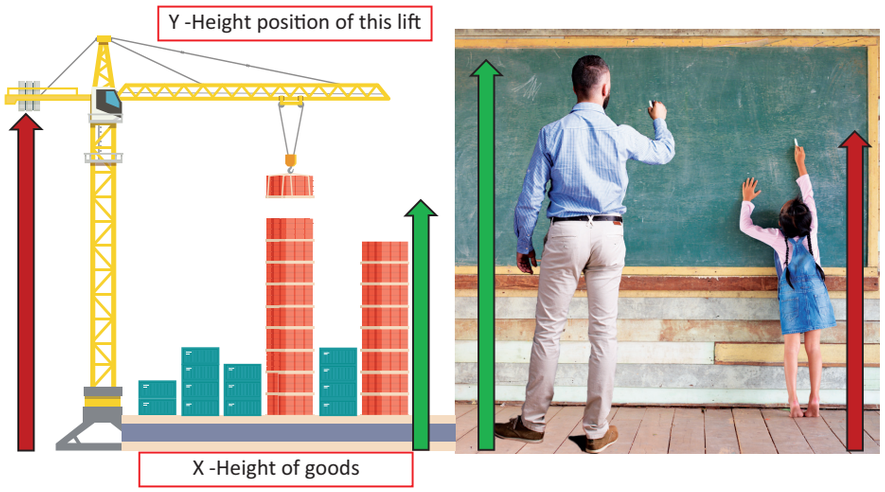


The variables are said to be positively correlated if larger values of x are associated with larger values of y and smaller values of x are associated with smaller values of y. In other words, if both the variables are varying in the *same direction* then the correlation is said to be positive. In other words, if one variable increases, the other variable (on an average) also increases or if one variable decreases, the other (on an average)variable also decreases.

For example,

i) Income and savings

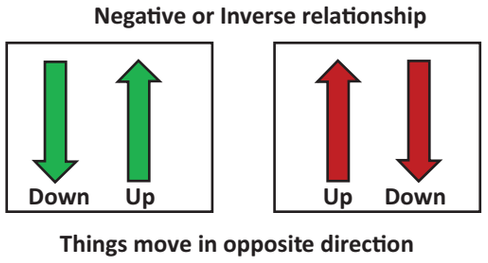
ii) Marks in Mathematics and Marks in Statistics. (*i.e.*,Direct relationship pattern exists).



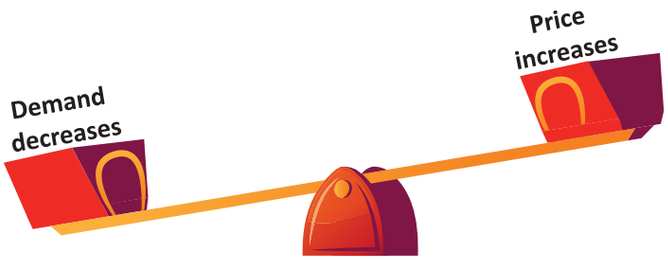
Height of the Lift increases / decreases according to the Height of goods increases / decreases.

The starting position of writing depends on the height of the writer.

2) Negative correlation: (Inverse correlation)



The variables are said to be negatively correlated if smaller values of x are associated with larger values of y or larger values x are associated with smaller values of y. That is the variables varying in the ***opposite directions*** is said to be negatively correlated. In other words, if one variable increases the other variable decreases and vice versa.



For example,

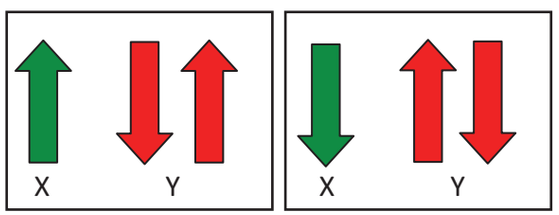
i) Price and demand

ii) Unemployment and purchasing power

3) Uncorrelated:

The variables are said to be uncorrelated if smaller values of x are associated with smaller or larger values of y and larger values of x are associated with larger or smaller values of y. If the two variables do not associate linearly, they are said to be uncorrelated. Here $r = 0$.

Important note: Uncorrelated does not imply independence. This means "do not interpret as the two variables are independent instead interpret as there is no specific linear pattern exists but there may be non linear relationship".



4) Perfect Positive Correlation

If the values of x and y increase or decrease ***proportionately*** then they are said to have perfect positive correlation.

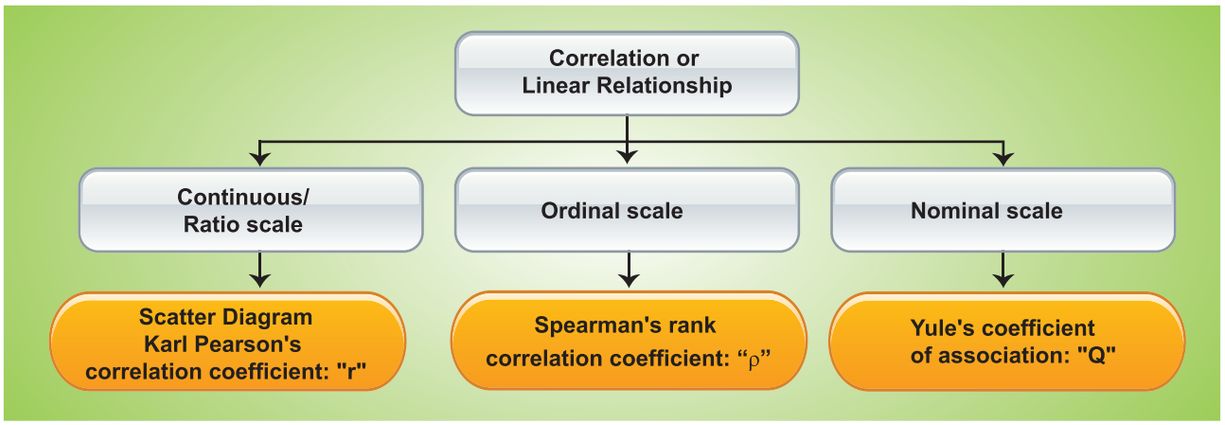
5) Perfect Negative Correlation

If x increases and y decreases ***proportionately*** or if x decreases and y increases ***proportionately***, then they are said to have perfect negative correlation.

Correlation Analysis

The purpose of correlation analysis is to find the existence of linear relationship between the variables. However, the method of calculating correlation coefficient depends on the types of measurement scale, namely, ratio scale or ordinal scale or nominal scale.

Statistical tool selection



Methods to find correlation

1. Scatter diagram
2. Karl Pearson's product moment correlation coefficient : '*r*'
3. Spearman's Rank correlation coefficient: 'ρ'
4. Yule's coefficient of Association: '*Q*'

NOTE

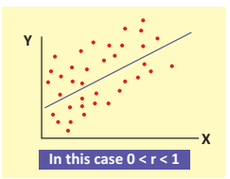
For higher order dimension of nominal or categorical variables in a contingency table, use chi-square test for independence of attributes. (Refer Chapter 2)

4.3 SCATTER DIAGRAM

A scatter diagram is the simplest way of the diagrammatic representation of bivariate data. One variable is represented along the *X*-axis and the other variable is represented along the *Y*-axis. The pair of points are plotted on the two dimensional graph. The diagram of points so obtained is known as scatter diagram. The direction of flow of points shows the type of correlation that exists between the two given variables.

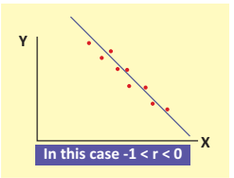
1) Positive correlation

If the plotted points in the plane form a band and they show the rising trend from the lower left hand corner to the upper right hand corner, the two variables are positively correlated.



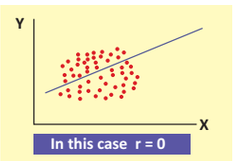
2) Negative correlation

If the plotted points in the plane form a band and they show the falling trend from the upper left hand corner to the lower right hand corner, the two variables are negatively correlated.



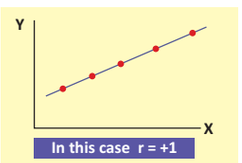
3) Uncorrelated

If the plotted points spread over in the plane then the two variables are uncorrelated.



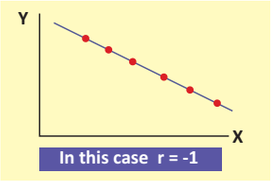
4) Perfect positive correlation

If all the plotted points lie on a straight line from lower left hand corner to the upper right hand corner then the two variables have perfect positive correlation.



5) Perfect Negative correlation

If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner, the two variables have perfect negative correlation.



4.3.1 Merits and Demerits of scatter diagram

Merits

- It is a simple and non-mathematical method of studying correlation between the variables.
- It is not influenced by the extreme items
- It is the first step in investigating the relationship between two variables.
- It gives a rough idea at a glance whether there is a positive correlation, negative correlation or uncorrelated.

Demerits

- We get an idea about the direction of correlation but we cannot establish the exact strength of correlation between the variables.
- No mathematical formula is involved.

4.4 KARL PEARSON'S CORRELATION COEFFICIENT

When there exists some relationship between two measurable variables, we compute the degree of relationship using the correlation coefficient.

Co-variance

Let (X,Y) be a bivariable normal random variable where $V(X)$ and $V(Y)$ exists. Then, covariance between X and Y is defined as

$$\text{cov}(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$$

If (x_i,y_i), i=1,2, ..., n is a set of n realisations of (X,Y), then the sample covariance between X and Y can be calculated from

$$\text{cov}\left(X,Y\right)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\,\bar{y}$$

4.4.1 Karl Pearson's coefficient of correlation

When X and Y are linearly related and (X,Y) has a bivariate normal distribution, the co-efficient of correlation between X and Y is defined as

$$r\left(X,Y\right)=\frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

This is also called as product moment correlation co-efficient which was defined by Karl Pearson.

Based on a given set of n paired observations (x_i,y_i), i=1,2, ... n the sample correlation co-efficient between X and Y can be calculated from

$$r\left(X,Y\right)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\;\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}^2}}$$

or, equivalently

$$r(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

4.4.2 Properties

1. The correlation coefficient between X and Y is same as the correlation coefficient between Y and X (*i.e*, $r_{xy} = r_{yx}$).
2. The correlation coefficient is free from the units of measurements of X and Y
3. The correlation coefficient is unaffected by change of scale and origin.

   Thus, if $u_i = \frac{x_i - A}{c}$ and $v_i = \frac{y_i - B}{d}$ with $c \neq 0$ and $d \neq 0$ $\quad i$=1,2, ..., n

$$r=\frac{n\sum_{i=1}^{n}u_iv_i-\sum_{i=1}^{n}u_i\sum_{i=1}^{n}v_i}{\sqrt{n\sum_{i=1}^{n}u_i^2-\left(\sum_{i=1}^{n}u_i\right)^2}\sqrt{n\sum_{i=1}^{n}v_i^2-\left(\sum_{i=1}^{n}v_i\right)^2}}$$

   where A and B are arbitrary values.

Remark 1: If the widths between the values of the variabls are not equal then take $c = 1$ and $d = 1$.

Interpretation

The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$

- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.
- If $|r| \geq 0.7$ then the correlation will be of higher degree. In interpretation we use the adjective 'highly'
- If X and Y are independent, then $r_{xy} = 0$. However the converse need not be true.

Example 4.1

The following data gives the heights(in inches) of father and his eldest son. Compute the correlation coefficient between the heights of fathers and sons using Karl Pearson's method.

Height of father	65	66	67	67	68	69	70	72
Height of son	67	68	65	68	72	72	69	71

Solution:

Let x denote height of father and y denote height of son. The data is on the ratio scale. We use Karl Pearson's method.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Calculation

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

$$r = \frac{8\times 37560 - 544\times 552}{\sqrt{8\times 37028 - (544)^2}\ \sqrt{8\times 38132 - (552)^2}} = 0.603$$

Heights of father and son are positively correlated. It means that on the average , if fathers are tall then sons will probably tall and if fathers are short, probably sons may be short.

Short-cut method

Let $A = 68$, $B = 69$, $c = 1$ and $d = 1$

x_i	y_i	$u_i = (xi - A)/c$ $= xi - 68$	$v_i = (y_i - B)/d$ $= y_i - 69$	u_i^2	v_i^2	$u_i v_i$
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2}}$$

$$r = \frac{8\times 24 - 0\times 0}{\sqrt{8\times 36 - (0)^2}\ \sqrt{8\times 44 - (0)^2}}$$

$$r = \frac{8\times 24}{\sqrt{8\times 36}\ \sqrt{8\times 44}}$$

$$= 0.603$$

Note: The correlation coefficient computed by using direct method and short-cut method is the same.

Example 4.2

The following are the marks scored by 7 students in two tests in a subject. Calculate coefficient of correlation from the following data and interpret.

Marks in test-1	12	9	8	10	11	13	7
Marks in test-2	14	8	6	9	11	12	3

Solution:

Let x denote marks in test-1 and y denote marks in test-2.

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	12	14	144	196	168
	9	8	81	64	72
	8	6	64	36	48
	10	9	100	81	90
	11	11	121	121	121
	1	12	169	144	156
	7	3	49	9	21
Total	70	63	728	651	676

$$r = \frac{n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$\sum_{i=1}^{n} x_i = 70 \quad \sum_{i=1}^{n} x_i^2 = 728 \quad \sum_{i=1}^{n} x_iy_i = 676$$

$$\sum_{i=1}^{n} y_i = 63 \quad \sum_{i=1}^{n} y_i^2 = 651 \quad n = 7$$

$$r = \frac{7\times 676 - 70\times 63}{\sqrt{\left[7\times 728 - 70^2\right]} \times \sqrt{\left[7\times 651 - 63^2\right]}}$$

$$= \frac{4732 - 4410}{\sqrt{[5096 - 4900]} \times \sqrt{[7\times 651 - 3969]}}$$

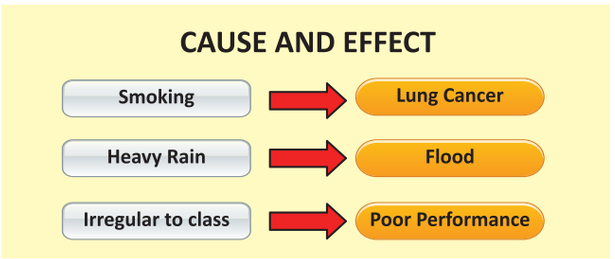
$$= \frac{322}{\sqrt{196} \times \sqrt{588}} = \frac{322}{14\times 24.25} = \frac{322}{339.5} = 0.95$$

There is a high positive correlation between test-1 and test-2. That is those who perform well in test-1 will also perform well in test-2 and those who perform poor in test-1 will perform poor in test- 2.

The students can also verify the results by using shortcut method.

4.4.3 Limitations of Correlation

Although correlation is a powerful tool, there are some limitations in using it:

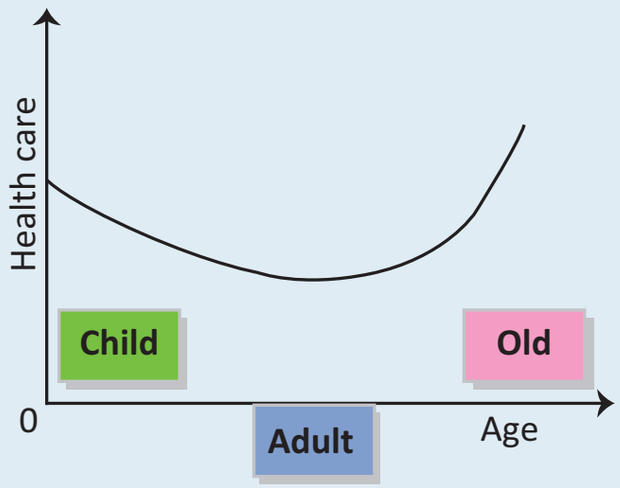


1. Outliers (extreme observations) strongly influence the correlation coefficient. If we see outliers in our data, we should be careful about the conclusions we draw from the value of r. The outliers may be dropped before the calculation for meaningful conclusion.
2. Correlation does not imply causal relationship. That a change in one variable causes a change in another.

NOTE

1. **Uncorrelated** : Uncorrelated ($r = 0$) implies no 'linear relationship'. But there may exist non-linear relationship (curvilinear relationship).

   Example: Age and health care are related. Children and elderly people need much more health care than middle aged persons as seen from the following graph.



However, if we compute the linear correlation r for such data, it may be zero implying age and health care are uncorrelated, but non-linear correlation is present.

2. **Spurious Correlation** : The word '**spurious**' from Latin means **'false'** or 'illegitimate'. *Spurious correlation means an association extracted from correlation coefficient that may not exist in reality.*

4.5 SPEARMAN'S RANK CORRELATION COEFFICIENT

If the data are in ordinal scale then Spearman's rank correlation coefficient is used. It is denoted by the Greek letter ρ (**rho**).

Spearman's correlation can be calculated for the subjectivity data also, like competition scores. The data can be ranked from low to high or high to low by assigning ranks.

Spearman's rank correlation coefficient is given by the formula

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

where $D_i = R_{1i} - R_{2i}$

R_{1i} = rank of i in the first set of data

R_{2i} = rank of i in the second set of data and

n = number of pairs of observations

Interpretation

Spearman's rank correlation coefficient is a statistical measure of the strength of a monotonic (increasing/decreasing) relationship between paired data. Its interpretation is similar to that of Pearson's. That is, the closer to the ±1 means the stronger the monotonic relationship.

Positive Range	Negative Range
0.01 to 0.19: "Very Weak Agreement"	(-0.01) to (-0.19): "Very Weak Disagreement"
0.20 to 0.39:"Weak Agreement"	(-0.20) to (-0.39): "Weak Disagreement"
0.40 to 0.59: "Moderate Agreement"	(-0.40) to (-0.59): "Moderate Disagreement"
0.60 to 0.79: "Strong Agreement"	(-0.60) to (-0.79): "Strong Disagreement"
0.80 to 1.0: "Very Strong Agreement"	(-0.80) to (-1.0): "Very Strong Disagreement"

Example 4.3

Two referees in a flower beauty competition rank the 10 types of flowers as follows:

Referee A	1	6	5	10	3	2	4	9	7	8
Referee B	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient and find out what degree of agreement is between the referees.

Solution:

Rank by 1st referee R_{1i}	Rank by 2nd referee R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			$\sum_{i=1}^{n} D_i^2 = 60$

Here n = 10 and $\sum_{i=1}^{n} D_i^2 = 60$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6\times 60}{10\left(10^2-1\right)} = 1 - \frac{360}{10(99)} = 1 - \frac{360}{990} = 0.636$$

Interpretation: Degree of agreement between the referees 'A' and 'B' is 0.636 and they have "strong agreement" in evaluating the competitors.

Example 4.4

Calculate the Spearman's rank correlation coefficient for the following data.

Candidates	1	2	3	4	5
Marks in Tamil	75	40	52	65	60
Marks in English	25	42	35	29	33

Solution:

Tamil		English		$D_i = R_{1i} - R_{2i}$	D_i^2
Marks	Rank (R_{1i})	Marks	Rank (R_{2i})		
75	1	25	5	-4	16
40	5	42	1	4	16
52	4	35	2	2	4
65	2	20	4	-2	4
60	3	33	3	0	0
					40

$$\sum_{i=1}^{n} D_i^2 = 40 \text{ and } n = 5$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 40}{5\left(5^2 - 1\right)} = 1 - \frac{240}{5(24)} = -1$$

Interpretation: This perfect negative rank correlation (-1) indicates that scorings in the subjects, totally disagree. Student who is best in Tamil is weakest in English subject and vice-versa.

Example 4.5

Quotations of index numbers of equity share prices of a certain joint stock company and the prices of preference shares are given below.

Years	2013	2014	2015	2016	2017	2008	2009
Equity shares	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Reference shares	75.1	75.9	77.1	78.2	79	74.6	76.2

Using the method of rank correlation determine the relationship between equity shares and preference shares prices.

Solution:

Equity shares	Preference share	R_{1i}	R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
97.5	75.1	4	6	-2	4
99.4	75.9	1	5	-4	16
98.6	77.1	2	3	-1	1
96.2	78.2	6	2	4	16
95.1	79.0	7	1	6	36
98.4	74.6	3	7	-4	16
97.1	76.2	5	4	1	1
					$\sum_{i=1}^{n} D_i^2 = 90$

$\sum_{i=1}^{n} D_i^2 = 90$ and $n = 7$.

Rank correlation coefficient is

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 90}{7\left(7^2 - 1\right)} = 1 - \frac{540}{7 \times 48} = 1 - \frac{540}{336} = 1 - 1.6071 = -0.6071$$

Interpretation: There is a negative correlation between equity shares and preference share prices. There is a strong disagreement between equity shares and preference share prices.

4.5.1 Repeated ranks

When two or more items have equal values (i.e., a tie) it is difficult to give ranks to them. In such cases the items are given the average of the ranks they would have received. For example, if two individuals are placed in the 8th place, they are given the rank $\frac{8+9}{2} = 8.5$ each, which is common rank to be assigned and the next will be 10; and if three ranked equal at the 8th place, they are given the rank $\frac{8+9+10}{3} = 9$ which is the common rank to be assigned to each; and the next rank will be 11.

In this case, a different formula is used when there is more than one item having the same value.

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

where m_i is the number of repetitions of i^{th} rank

Example 4.6

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

Repetitions of ranks

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

Therefore,

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{\left[81.5 + 0.5 + 2\right]}{504} = 1 - \frac{504}{504} = 0$$

Interpretation: Marks in Commerce and Mathematics are uncorrelated

4.6 YULE'S COEFFICIENT OF ASSOCIATION

Udny yule

This measure is used to know the existence of relationship between the two attributes A and B (binary complementary variables). Examples of attributes are drinking, smoking, blindness, honesty, etc.

Udny Yule (1871 – 1951), was a British statistician. He was educated at Winchester College and at University College London. After a year dong research in experimental physics, he returned to University College in 1893 to work as a demonstrator for Karl Pearson. Pearson was beginning to work in statistics and Yule followed him into this new field. Yule was a prolific writer, and was active in Royal Statistical Society and received its Guy Medal in Gold in 1911, and served as its President in 1924–26.The concept of Association is due to him.

Coefficient of Association

Yule's Coefficient of Association measures the strength and direction of association. "Association" means that the attributes have some degree of agreement.

2×2 Contingency Table

Attribute A ⇩	Attribute B		Total
	Yes B	No β	
Yes A	(AB)	$(A\beta)$	(A)
No α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

Note 1: The usage of the symbol α is not to be confused with level of significance.

Note 2: (AB): Number with attributes AB etc.

This coefficient ranges from –1 to +1. The values between –1 and 0 indicate inverse relationship (association) between the attributes. The values between 0 and +1 indicate direct relationship (association) between the attributes.

Example 4.7

Out of 1800 candidates appeared for a competitive examination 625 were successful; 300 had attended a coaching class and of these 180 came out successful. Test for the association of attributes attending the coaching class and success in the examination.

Solution:

$N = 1800$

A: Success in examination

α: No success in examination

B: Attended the coaching class

β: Not attended the coaching class

$(A) = 625,\ (B) = 300,\ (AB) = 180$

	B	β	Total
A	180	445	625
α	120	1055	1175
Total	300	1500	$N = 1800$

Yule's coefficient: $Q = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

$$= \frac{180 \times 1055 - 445 \times 120}{180 \times 1055 + 445 \times 120}$$

$$= \frac{189900 - 53400}{189900 + 53400}$$

$$= \frac{136500}{243300}$$

$$= 0.561 > 0$$

Interpretation: There is a positive association between success in examination and attending coaching classes. Coaching class is useful for success in examination.

Remark: Consistency in the data using contingency table may be found as under.

Construct a 2 × 2 contingency table for the given information. If at least one of the cell frequencies is negative then there is inconsistency in the given data.

Example 4.8

Verify whether the given data: N = 100, (A) = 75, (B) = 60 and (AB) = 15 is consistent.

Solution:

The given information is presented in the following contingency table.

	B	β	Total
A	15	60	75
α	45	(-20)	25
Total	60	40	N = 100

Notice that $(\alpha\beta) = -20$

Interpretation: Since one of the cell frequencies is negative, the given data is "Inconsistent".

POINTS TO REMEMBER

- Correlation study is about finding the linear relationship between two variables. Correlation is not causation. Sometimes the correlation may be spurious.
- Correlation coefficient lies between –1 and +1.
- Pearson's correlation coefficient provides the type of relationship and intensity of relationship, for the data in ratio scale measure.
- Spearman's correlation measures the relationship between the two ordinal variables.
- Yule's coefficient of Association measures the association between two dichotomous attributes.

EXERCISE 4

FD6ACU

I. Choose the best answer.

1. The statistical device which helps in analyzing the co-variation of two or more variables is

   (a) variance (b) probability

   (c) correlation coefficient (d) coefficient of skewness

2. "The attempts to determine the degree of relationship between variables is correlation" is the definition given by

   (a) A.M. Tuttle (b) Ya-Kun-Chou

   (c) A.L. Bowley (d) Croxton and Cowden

3. If the two variables do not have linear relationship between them then they are said to have

   (a) positive correlation (b) negative correlation

   (c) uncorrelated (d) spurious correlation

4. If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner then it is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

5. If $r = +1$, then the correlation is called

   (a) perfect positive correlation (b) perfect negative correlation

   (c) positive correlation (d) negative correlation

6. The correlation coefficient lies in the interval

   (a) $-1 \le r \le 0$ (b) $-1 < r < 1$ (c) $0 \le r \le 1$ (d) $-1 \le r \le 1$

7. Rank correlation coefficient is given by

   (a) $1+\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (b) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3-n}$ (c) $1-\frac{6\sum_{i=1}^{n} D_i^2}{n^3+n}$ (d) $1-\frac{6\sum_{i=1}^{n} D_i^3}{n\left(n^2-1\right)}$

8. If $\sum D^2 = 0$, rank correlation is

   (a) 0 (b) 1 (c) 0.5 (d) –1

9. Rank correlation was developed by

   (a) Pearson (b) Spearman (c) Yule (d) Fisher

10. Product moment coefficient of correlation is

    (a) $r=\frac{\sigma_x\ \sigma_y}{\operatorname{cov}(x,y)}$ (b) $r=\sqrt{\sigma_x\ \sigma_y}$ (c) $r=\frac{\operatorname{cov}(x,\ y)}{\sigma_x\ \sigma_y}$ (d) $r=\frac{\operatorname{cov}(x,\ y)}{\sigma_{xy}}$

11. The purpose of the study of _____ is to identify the factors of influence and try to control them for better performance.

(a) mean (b) correlation (c) standard deviation (d) skewness

12. The height and weight of a group of persons will have _____ correlation.

(a) positive (b) negative

(c) zero (d) both positive and negative

13. _____ correlation studies the association of two variables with ordinal scale.

(a) A.M. Tuttle rank (b) Croxton and Cowdon rank

(c) Karl Pearson's rank (d) Spearman's rank.

14. _____ presents a graphic description of quantitative relation between two series of facts.

(a) scatter diagram (b) bar diagram (c) pareto diagram (d) pie diagram

15. _____ measures the degree of relationship between two variables.

(a) standard deviation (b) correlation coefficient

(c) moment (d) median

16. The correlation coefficient of x and y is symmetric. Hence

(a) $r_{xy} = r_{yx}$ (b) $r_{xy} > r_{yx}$ (c) $r_{xy} < r_{yx}$ (d) $r_{xy} \neq r_{yx}$

17. If cov $(x, y) = 0$ then its interpretation is

(a) x and y are positively correlated (b) x and y are negatively correlated

(c) x and y are uncorrelated (d) x and y are independent

18. Rank correlation is useful to study data in _____ scale.

(a) ratio (b) ordinal (c) nominal (d) ratio and nominal

19. If $r = 0$ then cov(x, y) is

(a) 0 (b) +1 (c) -1 (d) α

20. If cov$(x, y) = \sigma_x, \sigma_y$ then

(a) $r = 0$ (b) $r = -1$ (c) $r = +1$ (d) $r = \alpha$

II. Give very short answer to the following questions.

21. What is correlation?

22. Write the definition of correlation by A.M. Tuttle.

23. What are the different types of correlation?

24. What are the types of simple correlation?

25. What do you mean by uncorrelated?

26. What you understand by spurious correlation?

27. What is scatter diagram?

28. Define co-variance.

29. Define rank correlation.

30. If $\sum D^2 = 0$ what is your conclusion regarding Spearman's rank correlation coefficient?

31. Give an example for (i) positive correlation

(ii) negative correlation (iii) no correlation

32. What is the value of 'r' when two variables are uncorrelated?

33. When the correlation coefficient is +1, state your interpretation.

III. Give short answer to the following questions.

34. Write any three uses of correlation.

35. Define Karl Pearson's coefficient of correlation.

36. How do you interpret the coefficient of correlation which lies between 0 and +1?

37. Write down any 3 properties of correlation?

38. If rank correlation coefficient $r = 0.8$, $\sum D^2 = 3$ then find n?

39. Write any three merits of scatter diagram.

40. Given that cov$(x, y) = 18.6$, variance of $x = 20.2$, variance of $y = 23.7$. Find r.

41. Test the consistency of the following data with the symbols having their usual meaning. $N = 1000$, $(A) = 600$, $(B) = 500$, $(AB) = 50$.

IV. Give detailed answer to the following questions.

42. Explain different types of correlation.

43. Explain scatter diagram.

44. Calculate the Karl Pearson's coefficient of correlation for the following data and interpret.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

45. Find the Karl Pearson's coefficient of correlation for the following data.

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

How are the wages and cost of living correlated?

46. Calculate the Karl Pearson's correlation coefficient between the marks (out of 10) in statistics and mathematics of 6 students.

Student	1	2	3	4	5	6
Statistics	7	4	6	9	3	8
Mathematics	8	5	4	8	3	6

47. In a marketing survey the prices of tea and prices of coffee in a town based on quality was found as shown below. Find the rank correlation between prices of tea and prices of coffee.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

48. Calculate the Spearman's rank correlation coefficient between price and supply from the following data.

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

49. A random sample of 5 college students is selected and their marks in Tamil and English are found to be:

Tamil	85	60	73	40	90
English	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

50. Calculate Spearman's coefficient of rank correlation for the following data.

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

51. Calculate the coefficient of correlation for the following data using ranks.

Mark in Tamil	29	24	25	27	30	31
Mark in English	29	19	30	33	37	36

52. From the following data calculate the rank correlation coefficient.

x	49	34	41	10	17	17	66	25	17	58
y	14	14	25	7	16	5	21	10	7	20

Yule's coefficient

53. Can vaccination be regarded as a preventive measure of Hepatitis B from the data given below. Of 1500 person in a locality, 400 were attacked by Hepatitis B. 750 has been vaccinated. Among them only 75 were attacked.

ANSWERS

I 1. (c) 2. (b) 3. (c) 4. (b) 5. (a)
6. (d) 7. (b) 8. (b) 9. (b) 10. (b)
11. (b) 12. (a) 13. (d) 14. (a) 15. (b)
16. (a) 17. (c) 18. (b) 19. (a) 20. (c)

II 30. $r = 1$

III 38. $n = 10$

40. $r = 0.85$

41. $(\alpha\beta) = -50$, The given data is inconsistent

IV 44. $r = 0.95$ it is highly positively correlated

45. $r = 0.847$ wages and cost of living are highly positively correlated.

46. $r = 0.8081$. Statistics and mathematics marks are highly positively correlated.

47. $\rho = 0.8929$ price of tea and coffee are highly positively correlated.

48. $\rho = 1$ (perfect positive correlation)

49. $\rho = 0.8$

50. $\rho = -0.905$ x and y are highly negatively

51. $\rho = -0.78$ marks in Tamil and English are negatively correlated.

52. $\rho = +0.733$

53. There is a negative association between attacked and vaccinated.

54. There is a positive association between not attacked and not vaccinated.

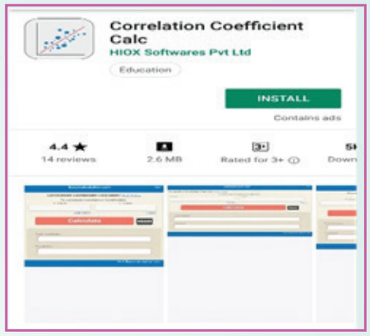
55. Hence vaccination can be regarded as a preventive measure of Hepatitis B.

ICT CORNER

CORRELATION ANALYSIS

STATS IN YOUR PALM

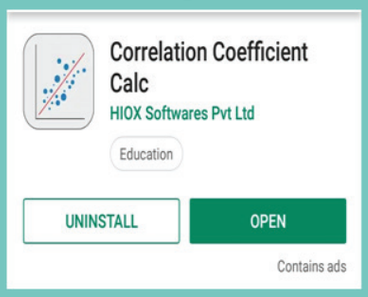
This activity is to calculate Correlation Coefficient



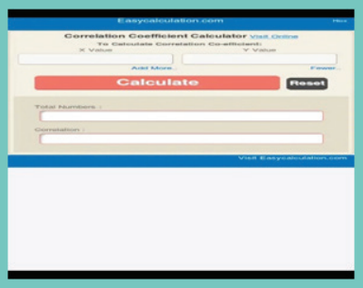
Steps:

- This is an android app activity. Open the browser and type the URL given (or) scan the QR code. (Or) search for "**Correlation Coefficient**" in google play store.
- (i) Install the app and open the app, (ii) To calculate Correlation Co-efficient in put the the values of X and Y in the given box (iii) Then click "**CALCULATE**" we will get the result.

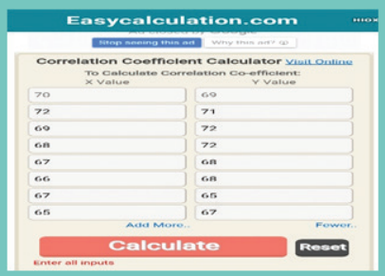
Step-1



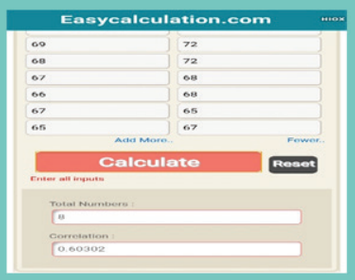
Step-2



Step-3



Result



Pictures are indicatives only*

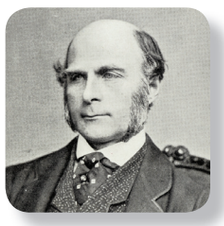
URL:

https://play.google.com/store/apps/details?id=com.hiox.CreliCoefficientCalcul

B273_12_STAT_EM

CHAPTER

5 REGRESSION ANALYSIS



Francis Galton

Francis Galton (1822-1911) was born in a wealthy family. The youngest of nine children, he appeared as an intelligent child. Galton's progress in education was not smooth. He dabbled in medicine and then studied Mathematics at Cambridge. In fact he subsequently freely acknowledged his weakness in formal Mathematics, but this weakness was compensated by an exceptional ability to understand the meaning of data. Many statistical terms, which are in current usage were coined by Galton. For example, correlation is due to him, as is regression, and he was the originator of terms and concepts such as quartile, decile and percentile, and of the use of median as the midpoint of a distribution.

The concept of regression comes from genetics and was popularized by Sir Francis Galton during the late 19th century with the publication of regression towards mediocrity in hereditary stature. Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. An examination of publications of Sir Francis Galton and Karl Pearson revealed that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression. Subsequent efforts by Galton and Pearson brought many techniques of multiple regression and the product-moment correlation coefficient.

LEARNING OBJECTIVES

The student will be able to

- ❖ know the concept of regression, its types and their uses.
- ❖ fit best line of regression by applying the method of least squares.
- ❖ calculate the regression coefficient and interpret the same.
- ❖ know the uses of regression coefficients.
- ❖ distinguish between correlation analysis and regression analysis.

GMGYF6

Introduction

The correlation coefficient is an useful *statistical tool for describing the type (positive or negative or uncorrelated) and intensity of linear relationship* (such as moderately or highly) between two variables. But it fails to give a *mathematical functional* relationship for prediction purposes. Regression analysis is a vital statistical method for obtaining functional relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one to understand how the typical value of the dependent variable (or 'response variable') changes when any one of the independent variables (regressor(s) or predictor(s)) is varied, while the other independent variables are held fixed. It helps to determine the impact of changes in the value(s) of the the independent variable(s) upon changes in the value of the dependent variable. Regression analysis is widely used for prediction.

5.1 DEFINITION

Regression analysis is a statistical method of determining the mathematical functional relationship connecting independent variable(s) and a dependent variable.

Types of 'Regression'

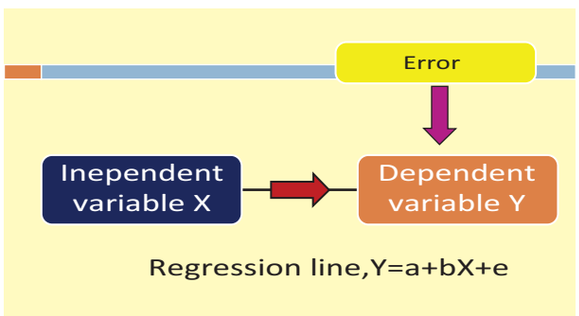
Based on the kind of relationship between the dependent variable and the set of independent variable(s), there arises two broad categories of regression *viz.*, linear regression and non-linear regression.

If the relationship is linear and there is only one independent variable, then the regression is called as simple linear regression. On the other hand, if the relationship is linear and the number of independent variables is two or more, then the regression is called as multiple linear regression. If the relationship between the dependent variable and the independent variable(s) is not linear, then the regression is called as non-linear regression.

5.1.1 Simple Linear Regression

It is one of the most widely known modeling techniques. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of relationship is linear. This relationship can be expressed using a straight line equation (linear regression) that best approximates all the individual data points.

Simple linear regression establishes a relationship between a **dependent variable (Y)** and one **independent variable (X)** using a **best fitted straight line** (also known as regression line).



NOTE

There are many reasons for the presence of the error term in the linear regression model. It is also known as measurement error. In some situations, it indicates the presence of several variables other than the present set of regressors.

The general form of the simple linear regression equation is $Y = a + bX + e$, where 'X' is independent variable, 'Y' is dependent variable, 'a' is intercept, 'b' is slope of the line and 'e' is error term. This equation can be used to estimate the value of response variable (Y) based on the given values of the predictor variable (X) within its domain.

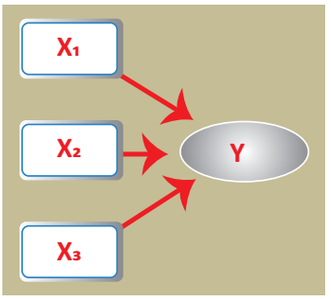
5.1.2 Multiple Linear Regression

In the case of several independent variables, regression analysis also allows us to compare the effects of independent variables measured on different scales, such as the effect of price changes and the number of promotional activities.

Multiple linear regression uses two or more independent variables to estimate the value(s) of the response variable (Y).

The general form of the multiple linear regression equation is

$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + e$



Here, Y represents the dependent (response) variable, X_i represents the i^{th} independent variable (regressor), a and b_i are the regression coefficients and e is the error term.

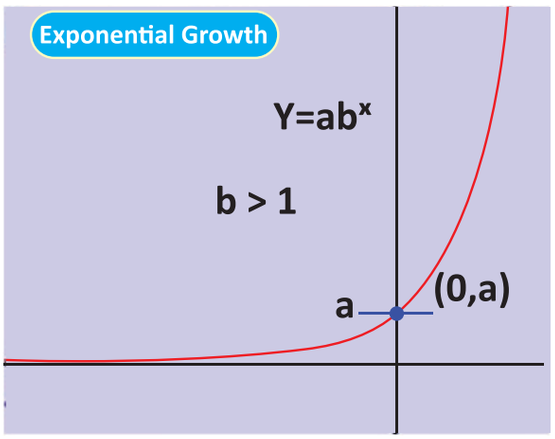
Suppose that price of a product (Y) depends mainly upon three promotional activities such as discount (X_1), instalment scheme (X_2) and free installation (X_3). If the price of the product has linear relationship with each promotional activity, then the relationship among Y and X_1, X_2 and X_3 may be expressed using the above general form as

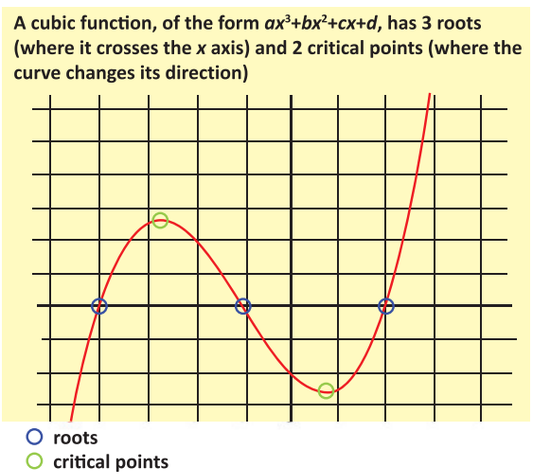
$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e\,.$$

These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building regression models for predictive purposes.

5.1.3 Non-Linear Regression

If the regression is not linear and is in some other form, then the regression is said to be non-linear regression. Some of the non-linear relationships are displayed below.





5.2 USES OF REGRESSION

Benefits of using regression analysis are as follows:

1. It indicates the **significant mathematical** relationship between independent variable (X) and dependent variable(Y). *(i.e)* Model construction
2. It indicates the **strength of impact (b)** of independent variable on a dependent variable.
3. It is used to estimate (interpolate) the value of the response variable for different values of the independent variable from its range in the given data. It means that extrapolation of the dependent variable is not generally permissible.

NOTE

Multiple linear regression and Curvilinear relationships (non-linear regression) are out of the syllabus. Basic information about them are given here, for enhancing the knowledge.

4. In the case of several independent variables, regression analysis is a way of mathematically sorting out which of those variables indeed have an impact (It answers the questions: Which independent variable matters most? Which can we ignore? How do those independent variables interact with each other?

5.3 WHY ARE THERE TWO REGRESSION LINES?

There may exist two regression lines in certain circumstances. When the variables X and Y are interchangeable with related to causal effects, one can consider X as independent variable and Y as dependent variable (or) Y as independent variable and X as dependent variable. As the result, we have (1) **the regression line of Y on X** and (2) **the regression line of X on Y**.

Both are valid regression lines. But we must judicially select the one regression equation which is suitable to the given environment.

Note: If, X only causes Y, then there is only one regression line, of Y on X.

5.3.1 Simple Linear Regression

In the general form of the simple linear regression equation of Y on X

$$Y = a + bX + e$$

the constants 'a' and 'b' are generally called as the regression coefficients.

The coefficient 'b' represents the rate of change in the value of the mean of Y due to every unit change in the value of X. When the range of X includes '0', then the intercept 'a' is $E(Y|X = 0)$. If the range of X does not include '0', then 'a' does not have practical interpretation.

If (x_i, y_i), $i = 1, 2, \ldots, n$ is a set of n-pairs of observations made on (X, Y), then fitting of the above regression equation means finding the estimates '$\hat{a}$' and '$\hat{b}$' for 'a' and 'b' respectively. These estimates are determined based on the following general assumptions:

i) the relationship between Y and X is linear (approximately).

ii) the error term 'e' is a random variable with mean zero.

iii) the error term 'e' has constant variance.

There are other assumptions on 'e', which are not required at this level of study.

Before going for further study, the following points are to be kept in mind.

- Both the independent and dependent variables must be measured at the interval scale.
- There must be **linear relationship** between independent and dependent variables.
- Linear Regression is very sensitive to **Outliers** (extreme observations). It can affect the regression line extremely and eventually the estimated values of Y too.

Meaning of line of "best fit"

Based on the assumption (ii), the response variable Y is also a random variable with mean

$$E(Y|X=x) = a + bx$$

In regression analysis, the main objective is finding the line of best fit, which provides the fitted equation of Y on X.

The line of 'best fit' is the line (straight line equation) which minimizes the error in the estimation of the dependent variable Y, for any specified value of the independent variable X from its range.

The regression equation $E(Y|X=x) = a + bx$ represents a family of straight lines for different values of the coefficients 'a' and 'b'. The problem is to determine the estimates of 'a' and 'b' by minimizing the error in the estimation of Y so that the line is a best fit. This necessitates to find the suitable values of the estimates of 'a' and 'b'.

5.4 METHOD OF LEAST SQUARES

In most of the cases, the data points do not fall on a straight line (not highly correlated), thus leading to a possibility of depicting the relationship between the two variables using several different lines. Selection of each line may lead to a situation where the line will be closer to some points and farther from other points. We cannot decide which line can provide best fit to the data.

Method of least squares can be used to determine the line of best fit in such cases. It determines the line of best fit for given observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

5.4.1 Method of Least Squares

To obtain the estimates of the coefficients 'a' and 'b', the least squares method minimizes the sum of squares of residuals. The residual for the i^{th} data point e_i is defined as the difference between the observed value of the response variable, y_i, and the estimate of the response variable, $\hat{y}_i$, and is identified as the error associated with the data. *i.e.*, $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

The method of least squares helps us to find the values of unknowns 'a' and 'b' in such a way that the following two conditions are satisfied:

- Sum of the residuals is zero. That is $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.
- Sum of the squares of the residuals $E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the least.

5.4.2 Fitting of Simple Linear Regression Equation

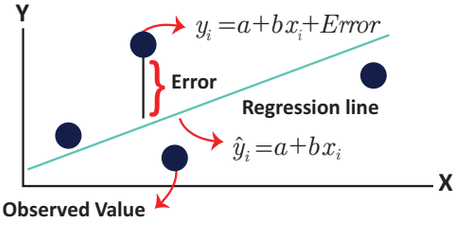
The method of least squares can be applied to determine the estimates of 'a' and 'b' in the simple linear regression equation using the given data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) by minimizing

$$E(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

i.e., $$E(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2 .$$

Here, $\hat{y}_i = a + bx_i$ is the expected (estimated) value of the response variable for given x_i.

Simple Linear Regression Model



It is obvious that if the expected value ($\hat{y}_i$) is close to the observed value (y_i), the residual will be small. Since the magnitude of the residual is determined by the values of '*a*' and *'b'*, estimates of these coefficients are obtained by minimizing the sum of the squared residuals, *E(a,b).*

Differentiation of *E(a,b)* with respect to *'a'* and *'b'* and equating them to zero constitute a set of two equations as described below:

$$\frac{\partial E(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\frac{\partial E(a,b)}{\partial b} = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

These give

$$na + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

$$a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i$$

These equations are popularly known as **normal equations.** Solving these equations for '*a*' and *'b'* yield the estimates $\hat{a}$ and $\hat{b}$.

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

and

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

It may be seen that in the estimate of '*b*', the numerator and denominator are respectively the sample covariance between *X* and *Y,* and the sample variance of *X*. Hence, the estimate of *'b'* may be expressed as

$$\hat{b} = \frac{Cov(X,Y)}{V(X)}$$

Further, it may be noted that for notational convenience the denominator of $\hat{b}$ above is mentioned as variance of *X*. But, the definition of sample variance remains valid as defined in Chapter I, that is, $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}^2\right)$.

From Chapter 4, the above estimate can be expressed using, r_{XY}, Pearson's coefficient of the simple correlation between *X* and *Y,* as

$$\hat{b} = r_{XY}\frac{SD(Y)}{SD(X)}.$$