Revised Edition

STATISTICS II

B. Sc. CSIT

Vikash Raj Satyal
Bijay Lal Pradhan
Nawaraj Poudel
Prakash Bdr Amatya

New Syllabus

Course Title: Statistics II
Course No: STA210
Nature of Course: Theory + Lab
Semester: III

Full Marks: 60 + 20 + 20
Pass Marks: 24 + 8 + 8
Credit Hrs: 3

Course Description

[illegible] concepts of sampling, testing of hypothesis, parametric and non parametric tests, [illegible] experimental designs and stochastic processes.

Course Objective

The main objective of the course is to acquire the theoretical as well as practical knowledge of estimation, testing of hypothesis, application of parametric and non-parametric statistical tests, design of experiments, multiple regression analysis, and basic concept of stochastic process with special focus to data problems related with computer science and information technology.

Course Contents

Unit 1: Sampling Distribution and Estimation **(6 Hrs.)**

Sampling distribution; sampling distribution of mean and proportion; Central Limit Theorem; Concept of inferential Statistics; Estimation; Methods of estimation; Properties of good estimator; Determination of sample size; Relationship of sample size with desired level of error Problems and illustrative examples related to computer Science and IT

Unit 2: Testing of hypothesis **(8 Hrs.)**

Types of statistical hypotheses; Power of the test, concept of p-value and use of p-value in decision making; steps used in testing of hypothesis; one sample tests for mean of normal population (for known and unknown variance), test for single proportion, test for difference between two means and two proportions, paired sample t-test; Linkage between confidence interval and testing of hypothesis Problems and illustrative examples related to computer Science and IT

Unit 3: Non parametric test **(8 Hrs.)**

Parametric vs. non-parametric test; Needs of applying non-parametric tests; One-sample test: Run test, Binomial test, Kolmogorov-Smirnov test; Two independent sample test: Median test, Kolmogorov-Smirnov test, Wilcoxon Mann Whitney test, Chi-square test; Paired-sample test: Wilcoxon signed rank test; Cochran's Q test; Friedman two way analysis of variance test; Kruskal Wallis test

Problems and illustrative examples related to computer Science and IT

Unit 4: Multiple correlation and regression **(6 Hrs.)**

Multiple and partial correlation; Introduction of multiple linear regression; Hypothesis testing of multiple regression; Test of significance of regression; Test of individual regression coefficient; Model adequacy tests

Problems and illustrative examples related to computer Science and IT

Unit 5: Design of experiment **(10 Hrs.)**

Experimental design; Basic principles of experimental designs; Completely Randomized Design (CRD); Randomized Block Design (RBD); ANOVA table, Efficiency of RBD relative to CRD, Estimations of missing value (one observation only), Advantages and disadvantages; Latin Square Design (LSD): Statistical analysis of m × m LSD for one observation per experimental unit, ANOVA table, Estimation of missing value in LSD (one observation only), Efficiency of LSD relative to RBD, Advantage and disadvantages.

Problems and illustrative examples related to computer Science and IT

Unit 6: Stochastic Process **(7 Hrs.)**

Definition and classification; Markov Process; Markov chain; Matrix approach; [illegible] distribution; Counting process; Binomial process; Poisson process; [illegible] process; Queuing system; Main component of queuing system; Little's law; [illegible] server queuing process; system with limited capacity; M/M/1 [illegible] performance.

Evaluation

S.No.	Practical Problems	No. of Practical Problems
1	Sampling distribution, random number generation, and computation of sample size	1
2	Methods of estimation (including interval estimation)	1
3	Parametric tests (covering most of the tests)	3
4	Non-parametric test(covering most of the tests)	3
5	Partial correlation	1
6	Multiple regression	1
7	Design of Experiments	3
	Stochastic process	2
	Total number of practical problems	15

Text Books:

1. Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, & Keying Ye(2012). Probability & Statistics for Engineers & Scientists. 9th Ed., Printice Hall.
2. Michael Baron (2013). Probability and Statistics for Computer Scientists. 2nd Ed., CRC Press, Taylor & Francis Group, A Chapman & Hall Book.

Reference Books:

1. Douglas C. Montgomery & George C. Runger (2003). Applied Statistics and Probability for Engineers. 3rd Ed., John Willey and Sons, Inc.
2. Sidney Siegel, & N. John Castellan, Jr. Nonparametric Statistics for the Behavioral Sciences. 2nd Ed., McGraw Hill International Editions.

Table of Contents

1 Chapter — SAMPLING DISTRIBUTION AND ESTIMATION

2 Chapter — TESTING OF HYPOTHESIS

3 Chapter — NON PARAMETRIC TEST

4 Chapter — MULTIPLE CORRELATION AND REGRESSION

5 Chapter DESIGN OF EXPERIMENT

6 Chapter STOCHASTIC PROCESS

1

Chapter

SAMPLING DISTRIBUTION AND ESTIMATION

CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Sampling distribution, Sampling distribution of mean and proportion.
- Central Limit Theorem
- Concept of inferential Statistics
- Estimation, Methods of estimation
- Properties of goods estimator
- Determination of sample size, Relationship of sample size with desired level of error Problem
- Illustrative examples related to computer science and IT
- Problems and illustrative examples related using software.

Sampling Distribution

A sample statistic is a numerical summary measure calculated from sample data. The mean, median, mode, standard deviation, sample proportion and correlation calculated for sample data are called sample statistics. On the other hand, the same numerical summary measures calculated for population data are called population parameters. A population parameter is always a constant, whereas a sample statistic is a random variable. Because every random variable most possesses a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is called sampling distribution.

If we take a sample of size n from a population of size N, then there are ${}^{N}C_n = k$ (say), possible samples. We can compute the sample statistic, say, T for each of these samples. Let $T_1, T_2, \ldots, T_k$ be the values of the k possible samples. Thus, the statistic T may be regarded as a random variable which can take any one of the values $T_1, T_2, \ldots, T_k$. The probability distribution of the statistic T is called the sampling distribution. The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection. The average value and standard deviation of sampling distribution plays vital role in statistics. The standard deviation of the sampling distribution of a statistic T is known as standard error of the statistic. Here we will discuss about the most important sampling distribution i.e. Sampling distribution of mean and proportion.

Sampling Distribution of Mean

Let us consider the sampling distribution of sample mean $\bar{X}$. Let us suppose that a random sample of size n is taken from a normal population with mean $\bar{X}$ and variance σ^2. Let $x_1, x_2, \ldots x_n$ is a random sample (without replacement) of size *n* from a finite population of size N, then

$E(\bar{X}) = \mu$ mean value of the sampling distribution of the sample mean is equal to Population mean and its variance is given by $V(\bar{X}) = \frac{\sigma^2(N - n)}{n(N - 1)}$ which measures the variability of sample mean. In the same way its standard error will be S.E. $(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{(N - n)}{N-1}}$

Where each observation in this sample, say, $x_1, x_2, \ldots x_n$, is a normally and independently distributed random variable with mean $\bar{X}$ and variance σ^2. Again

If sample is from an infinite (very large) population so that the sampling fraction *n/N* can be neglected, or if the sampling is done with replacement then variance of sample mean is given by $V(\bar{X}) = \frac{\sigma^2}{n}$ and standard deviation S.E. $(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (i.e. square root of variance is known as standard error of the sampling distribution.)

Sampling distribution of mean possesses the following properties:

(i) Sample mean $\bar{x}$ is an unbiased estimate of the population mean μ.

(ii) The variance of $\bar{x}$ depends on the sample size *(n)*.

When population standard deviation is not known then it will be estimated by Root mean square S which is given by $S = \sqrt{\frac{1}{n-1}\Sigma(x-\bar{x})^2)}$

Then standard error of sample mean is given by

$S.E.(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}}$ for simple random sampling without replacement for finite N.

$S.E.(\bar{X}) = \frac{S}{\sqrt{n}}$ for unknown population size

If we have sample standard deviation then the estimated population standard deviation will be $\sqrt{\frac{(n-1)}{n}}\, s$ Where sample SD $(s) = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2)}$

Then standard error of sample mean is given by

$S.E.(\bar{X}) = \frac{s}{\sqrt{n-1}}\sqrt{\frac{(N-n)}{N}}$ for simple random sampling without replacement for finite N

$S.E.(\bar{X}) = \frac{s}{\sqrt{n-1}}$ for unknown population size

Example 1: A sample of size 36 is drawn from a population consisting of 196 units. If the population standard deviation is 7, find the standard error of the sample mean when the sample drawn is (i) without replacement, (ii) with replacement.

Solution:

Here,

Population size (N) = 196,

Sample size (n) = 36

Population SD (σ) = 7

Sampling without replacement: The standard error of the sample mean is given by σ

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{7}{\sqrt{36}} \times \sqrt{\frac{196-36}{196-1}} = 1.0568$$

Sampling with replacement: The standard error of the sample mean is given by

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{36}} = 1.1667$$

Example 2: A random sample of 25 register has drawn from a large lot of register produced by an electronics company manufacturer. Find the standard error of sample mean if sample root mean square is found to be 10 ohms.

Solution:

Here,

Sample size (n) = 25

Root mean square (S) = 10 ohms.

Now standard error of sample mean is $\frac{S}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$

Note: If $\bar{X}_1$ and $\bar{X}_2$ denote the means calculated from independent random samples of sizes n_1 and n_2 and drawn from two normal population with standard deviations σ_1 and σ_2 respectively, then

S.E. $(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ = Standard error of difference between two means

When σ_1 and σ_2 are not known then combined variance of population is estimated by

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{[\Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2]}{n_1 + n_2 - 2} \text{ Where } s_i^2 = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x}_i)^2)} \quad i = 1,2$$

Then S.E. of difference of mean is given by S.E. $(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(S^2\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}\right)}$

Sampling Distribution of Proportion

A proportion is the number of elements with a given characteristic divided by the total number of elements in the group. Let X be the number of characteristics obtained in the population having size N. Then population proportion is given by $P = \frac{X}{N}$.

The sample proportion, $p = \frac{x}{n}$, is the point estimate of the population proportion, P. and the variance of the sample proportion is given by the formula $V(p) = \frac{PQ}{n} \times (\frac{N-n}{N-1})$

Where $Q = 1 - P$ and x = number of characteristics in the sample and n = sample size.

The standard error of population proportion is given by $S.E.(p) = \sqrt{\frac{PQ}{n}} \times \sqrt{\frac{N-n}{N-1}}$

If population size is large or the sampling is done with replacement then the variance term becomes $V(p) = \frac{PQ}{n}$ and its standard error is given as $S.E.(p) = \sqrt{\frac{PQ}{n}}$

(i.e. square root of variance is known as standard error of the sampling distribution)

Sampling distribution p of proportion possesses the following properties:

(i) Sample proportion is an unbiased estimate of the population proportion p.

(ii) The variance of p depends on the sample size (n)

If population proportion is not known then it will be estimated by using sample proportion p.

$SE(p) = \sqrt{\frac{pq}{n-1}} \cdot \sqrt{\frac{N-n}{N}}$ for simple random sampling without replacement for finite population size N.

$SE(p) = \sqrt{\frac{pq}{n}}$ for unknown population size (large population)

Note: The rule of thumb is if $\frac{n}{N} < 0.05$, then the population is said to be large and is used the formula for the variance as $\frac{PQ}{n}$.

Example 3: From the consignment of 100 apples 20 apples is drawn by simple random sampling method without replacement. If out of the 20 apples 15 apples found defective, what should be the standard error of the sample proportion?

Solution:

We have,

Population size (N) = 100, Sample size (n) = 20, No. of defectives (x) = 15, SE (p) = ?

We know,

Sample proportion of defectives $(p) = \frac{x}{n} = \frac{15}{20} = 0.75$

Sample proportion of non-defectives (q) = 1 – 0.75 = 0.25

Checking ratio

$\frac{n}{N} = \frac{20}{100} = 0.20 > 0.05$, the adjustment is needed.

Now, $SE(p) = \sqrt{\frac{pq}{n-1}} \cdot \sqrt{\frac{N-n}{N}} = \sqrt{\frac{0.75 \times 0.25}{20-1}} \sqrt{\frac{100-20}{100}} = 0.088$

Hence, the required standard error of sample proportion is 0.088.

Note: If p_1 and p_2 denote the sample proportions calculated from independent random samples of sizes n_1 and n_2 and drawn from two populations with proportions P_1 and P_2 respectively, then the standard error of difference between proportions is given by,

$$S.E.(p_1 - p_2) = \sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}$$

In particular, if it is assumed that the two population proportions P_1 and P_2 are equal, say $P_1 = P_2 = P$ then

$$S.E.(p_1 - p_2) = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ where } p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}, Q = 1 - p$$

Standard errors of some well-known statistics are

Statistic	Standard error
Mean (when σ known and population size infinite)	$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
Mean (when σ known and population size finite i.e. N)	$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$
Mean (when σ unknown and population size infinite)	$S.E.(\bar{X}) = \frac{S}{\sqrt{n}}$
Mean (when σ unknown and population size finite i.e. N)	$S.E.(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$
Difference of means (when σ's are known)	$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference of means (when σ's are unknown)	$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(S^2\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}\right)}$
Proportion (when population size is infinite)	$S.E.(p) = \sqrt{\frac{PQ}{n}}$
Proportion (when population size is finite i.e. N)	$S.E.(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$
Difference of proportions	$S.E.(p_1 - p_2) = \sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}$
Difference of proportions (two population proportion assumed to be equal)	$S.E.(p_1 - p_2) = \sqrt{\left(PQ\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}\right)}$

Central Limit Theorem

The central limit theorem states that as the sample size gets large enough, the sampling distribution of the mean is approximately normally distributed. This statement is true regardless of the shape of the distribution of the individual values in the population. If $x_1, x_2, \ldots x_n$ is a random sample from normal population with mean μ and variance σ^2, then the sample mean $\bar{x}$ is also normally distributed with mean μ and variance σ^2/n, i.e., $X \sim N(\mu, \sigma^2/n)$. This result is true even if the population from which the samples are drawn is not normal, provided the sample size is sufficiently large as stated in the following Central Limit Theorem.

"If $x_1, x_2, \ldots x_n$ is a random sample of size n from any population, then the sample mean $(\bar{x})$ is normally distributed with mean μ and variance σ^2/n provided n is sufficiently large."

"If $X_1, X_2, \ldots, X_n$ are independent random variables following any distribution, then under certain very general conditions, their sum $\Sigma X = X_1 + X_2 + \ldots + X_n$, is asymptotically normally distributed, i.e., ΣX follows normal distribution as $n \to \infty$.

By using this theorem, it has been proved that the sampling distributions of most of the statistics, like sample proportion (p), difference of sample proportions $(P_1\text{-}P_2)$, difference of sample means $(\mu_1 - \mu_2)$ difference of sample standard deviation $(s_1\text{-}s_2)$ are asymptotically normal, i.e., the standardized variates corresponding to any one of these statistics is N(0, 1) for large samples. Thus, if t is any statistic, then by central limit theorem,

$$z = \frac{t - E(t)}{S.E.(t)} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \text{ as } n \to \infty.$$

Now the issue here is what sample size to be taken as a large sample. As a general rule, statisticians have found that for many population distributions, when the sample size is at least 30, the sampling distribution of the mean is approximately normal. However, we can apply the central limit theorem for even smaller sample size if the population distribution is approximately bell shape (normal shape). In the case when the distribution is extremely skewed or has more than one mode, we have to take sample size larger than 30 to ensure the normality.

Concept of Inferential Statistics

The inductive inference may be termed as the logic of drawing statistically valid conclusions about the population characteristics on the basis of a sample drawn from it in a scientific manner. We shall develop the technique which enables us to generalize the results of the sample to the population; to find how far these generalizations are valid, and also to estimate the population parameters along with the degree of confidence. The answers to these and many other related problems are provided by a very important branch of Statistics, known as the *Statistical Inference.*

Estimation

Estimation of population parameters like mean, variance, proportion, correlation coefficient, etc., from the corresponding sample statistics is one of the very important problems of statistical inference. The theory of estimation was founded by Prof. R.A. Fisher in a series of fundamental papers about 1930 and is divided into two groups.

i. Point Estimation

ii. Interval Estimation.

In Point Estimation, a sample statistic (numerical value) is used to provide an estimate of population parameter whereas in Interval Estimation, probable range is specified within which the value of the parameter might be expected to lie.

Point Estimation

A particular value of a statistic which is used to estimate a given parameter is known as a *point estimate* or *estimator* of the parameter. A good estimator is one which is as close I true value of the parameter as possible. The following are some of the criteria which should satisfy to be a good estimator.

1. Unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

Unbiasedness

A statistic $t = t(x_1, x_2, \ldots, x_n)$,a function of the sample observations $x_1, x_2, \ldots x_n$ is said to be an unbiased estimate of the corresponding population parameter θ, if $E(t) = \theta$ i.e., if the mean value of the sampling distribution of the statistic is equal to the parameter. For example, the sample mean x is an unbiased estimate of the population mean μ; the sample proportion p is an unbiased estimate of the population proportion P, S^2 is unbiased estimate of variance.

i.e. $E(\bar{X}) = \mu$ and $E(p) = P$, $E(S^2) = \sigma^2$ where, $S^2 = \frac{1}{n-1}\Sigma(X - \bar{X})^2$

If $E(t) \neq \theta$, *then the statistic* t *is said to be a* biased estimator of θ.

Let $E(t) = \theta + Q$ then 'Q' is called the *'amount of bias'* in the estimate. If $Q > 0$, i.e., $E(t) > \theta$, *then* t *is said to be* positively biased *and if* $Q < 0$, i.e., $E\{t) < 0$, *it is said to be* negatively biased.

Consistency

A statistic t based on a sample of size n is said to be a consistent estimator of the parameter θ if it converges in probability to 0, *i.e.*, if $t_n \to 0$ as n. Symbolically, $\lim_{n\to\infty} P(t_n \to \theta) = 1$.

For any distribution, sample mean x is a consistent estimator of the population mean, sample proportion 'p' is a consistent estimator of population proportion P and sample variance s^2 is a consistent estimator of the population variance σ^2.The variance of sampling distribution of the estimator enables us to determine if the statistic is a consistent estimator of the parameter or not. The result is contained in the following theorem.

Theorem. A statistic $t = t_n = t(x_1, x_2, \ldots, x_n)$ is a consistent estimator of the parameter θ if

$$Var(t) \to 0 \text{ as } n \to \infty.$$

Efficiency

If we have more than one consistent estimator of a parameter θ, then *efficiency* is the criterion which enables us to choose between them by considering the variances of the sampling distributions of the estimators. Thus, if t_1 and t_2 are consistent estimators of a parameter θ such that

$$Var(t_1) < Var(t_2), \text{ for all } n$$

then t_1 is said to be more efficient than t_2. *In other words,* an estimator with lesser variability is said to be more efficient and consequently more reliable than the other.

If there exist more than two consistent estimators for the parameter θ, then considering the class of all such possible estimators we can choose the one whose sampling variance is minimum. Such an estimator is known as the *most efficient estimator* and provides a measure of the efficiency of the other estimators.

Definition. If t is the most efficient estimator of a parameter θ with variance v and t_1 is any other estimator with variance v_1, then the efficiency E of t_1 is defined as:

$$E = \frac{v}{v_1}$$

Since t is the most efficient estimator, its sampling variance v is minimum, i.e. $v < v_1$

The efficiency of any estimator cannot exceed unity.

Sufficiency

A statistic $t = t(x_1, x_2, \ldots x_n)$ is said to be a sufficient estimator of parameter θ if it contains all the information in the sample regarding the parameter. In other words, a sufficient statistic utilises all the information that a given sample can furnish about the parameter. If $t = t(x_1, x_2, \ldots x_n)$ is a statistic based on a random sample of size n from a population with probability function or pdf $p(x, \theta)$, then it is a sufficient estimator of θ if the conditional distribution of x_1 x_2, ..., x_n for given value of t is independent of θ, *i.e.*, if the conditional probability $P(x_1 \cap x_2 \ldots x_k t = k)$ does not depend on θ.

The sample mean $\overline{X}$ is sufficient estimator of population mean and sample proportion p is a sufficient estimator of population proportion P.

Properties of **Sufficient Estimators**

1. If a sufficient estimator exists for some parameter then, it is also the most efficient estimator.
2. It is always consistent.
3. It may or may not be unbiased.

Interval Estimation

In point estimation, a single value of a statistic (t) is used as an estimate of the population parameter (θ). But even the best possible point estimate may deviate enough from the true parameter value to make the estimate unsatisfactory. Such deviation will be controlled by the technique of *interval estimation*. This consists in the determination of two constants c_1 and c_2, say, such that

$$P(c_1 < \theta < c_2, \text{ for given value of } t) = 1 - \alpha.$$

Where α is the level of significance. The interval $[c_1, c_2]$, within which the unknown value of the parameter θ is expected to lie is known as *Confidence Interval* (Neyman) or *Fiducial Interval* (R.A. Fisher); the limits c_1 and c_2 so determined are known as *Confidence Limits or Fiducial Limits* and $1-\alpha$, is called the *confidence coefficient*, depending upon the desired precision of the estimate.

For example, $\alpha = 0.5$ (or 0.01) gives the 95% (or 99%) confidence limits.

If t is the statistic used to estimate the parameter θ, then $(1-\alpha)\%$ *confidence limits for* $\theta = t \pm SE(t) \times t_{\alpha/2}$ where α is the significant or critical value of t at level of significance α for a two-tailed test. Thus, the computation of confidence limits for a parameter, involves the following three steps:

(i) Compute the appropriate sample statistic t

(ii) Obtain the Standard Error of the sampling distribution of the statistic t *i.e., S.E. (t).*

(iii) Choose appropriate confidence coefficient $(1 - \alpha)$, depending on the precision of the estimate.

Interval Estimation of population mean for Large Samples

For large samples, the underlying distribution of the standardized variate corresponding to the sampling distribution of the statistic t will be asymptotically normally distributed i.e.

For practical purposes, samples may be regarded as large if the sample size is greater than 30 *i.e.*> 30.

From areas under normal probability curve, we have

$$P(-1.96 < Z < 1.96) = 0.95$$

Thus 95% *confidence limits for population mean are* $\bar{X} \pm \frac{1.96\sigma}{\sqrt{n}}$ where σ is assumed to be known, and the interval is the 95% confidence interval for estimating μ. Then, $(1 - \alpha)\%$ confidence limits are given by $\bar{X} \pm \frac{Z_{\alpha/2}\,\sigma}{\sqrt{n}}$ for large N.

And $(1 - \alpha)\%$ confidence limits are given by $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N - n)}{(N - 1)}}$ for large n is assumed and simple random sampling without replacement for finite N.

Two-Tailed Significant (Critical) Values of Z

Confidence coefficient $(1-\alpha)$	50%	68.27%	90%	95%	95.45%	98%	99%	99.73%
Significant value	0.6745	1	1.645	1.96	2	2.33	2.58	3

For small sample i.e. n < 30 the student t statistics is followed and given by $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ where s is given by $S = \sqrt{\frac{1}{n-1}\Sigma(x - \bar{x})^2)}$

Then, $(1-\alpha)\%$ confidence limits are given by

$\bar{x} \pm t_{\alpha/2,\, n-1} \dfrac{S}{\sqrt{n}}$ for unknown population size

$\bar{x} \pm t_{\alpha/2,}\ n-1 \dfrac{S}{\sqrt{n}} \sqrt{\dfrac{(N-n)}{N}}$ for simple random sampling without replacement for finite N.

In case of sample SD (s) = $\sqrt{\dfrac{1}{n}\Sigma(x-\bar{x})^2)}$

$(1-\alpha)\%$ confidence limits are given by

$\bar{x} \pm t_{\alpha/2,\, n-1} \dfrac{s}{\sqrt{n-1}}$ for unknown population size

$\bar{x} \pm t_{\alpha/2,\, n-1} \dfrac{s}{\sqrt{n-1}} \sqrt{\dfrac{(N-n)}{N}}$ for simple random sampling without replacement for finite N

Example 4: A random sample of 500 have the average of 68.6 with standard deviation 2.5. Find 95% and 99% confidence interval population average.

Solution:

Here, sample size n = 500

Sample mean $\bar{x}$ = 68.6

Sample standard deviation s = 2.5

95% confidence limits population mean μ is given by

$$\bar{x} \pm \frac{Z_{\alpha/2}\, s}{\sqrt{n-1}} = 68.6 \pm 1.96 \times \frac{2.5}{\sqrt{500-1}} = 68.6 \pm 0.219$$

Taking (-) sign 68.6 – 0.219 = 68.38

Taking (+) sign 69.8+0.219 = 68.819

Hence 99% confidence limits population mean μ is 68.38 to 68.819.

Example 5: A random sample of 12 records gave the average length of 163.99 minutes with standard deviation of 3.043 minutes. Find the 95% confidence limits for population mean and population total length if population consists of 100 units.

Solution:

Here, sample size (n) = 12

Sample mean ($\bar{x}$) = 163.99

Sample standard deviation (s) = 3.043

Population size (N) = 100

Confidence limit (α) = 5%

$t_{\alpha/2,\, n-1} = 2.20$

95% confidence limits population mean μ is

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n-1}} \sqrt{\frac{(N-n)}{N}}$$

$$= 163.99 \pm 2.20 \frac{3.043}{\sqrt{12-1}} \sqrt{\frac{100-12}{100}}$$

$$= 163.99 \pm 1.893$$

Taking (-) sign 163.99- 1.893 = 162.097

Taking (+) sign 163.99 + 1.893 = 165.883.

Hence 95% confidence limit for mean is 162.086min to 165.892 min

95% confidence limits for population total is given by

$$N\,[\bar{x} \pm t_{\alpha/2,n-1}\,S.E.\,(\bar{x})\,] = N\,[\,\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n-1}} \sqrt{\frac{(N-n)}{N}}\,]$$

$$= 100 \times 163.99 \pm 100 \times 2.20 \times \frac{3.043}{\sqrt{12-1}} \sqrt{\frac{100-12}{100}}$$

$$= 16399 \pm 189.3$$

Taking (-) sign, 16399 – 189.3 = 16209.7

Taking (+) sign, 16399+ 189.3 =16588.3

Hence 95% confidence limits for the population total is 16216.8 minutes and 16581.2 minutes respectively.

Confidence Limits for Proportions

Taking p as the sample proportion, according to Central Limit Theorem

$$Z = \frac{p - E(p)}{S.E.\,(p)} \sim N(0, 1)$$

Where E(p)=P and S.E. (p) = $\sqrt{\frac{PQ}{n}}$ if P and Q population proportion for success and failure and population is infinite.

For finite population and samples are drawn without replacement

$$S.E.\,(p) = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$$

Then, (1-α)% confidence limits are given by

$$p \pm Z_{\alpha/2} \sqrt{\frac{PQ}{n}}$$

If P is not known then confidence interval are

$$p \pm Z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

Also (1-α)% confidence limits are given by

$p \pm Z_{\alpha/2} \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$ for large n is assumed and simple random sampling without replacement for finite N.

If P is not known then confidence internal are $p \pm Z_{\alpha/2} \sqrt{\frac{pq}{n-1} \cdot \frac{N-n}{N}}$.

Example 6: A dice is thrown 9000 times and a throw of 3 or 4 is considered as a success. Suppose the 3240 throws of a 3 or 4 have been made out. Find the 95% confidence interval for the population proportion of outcome of 3 or 4.

Solution:

Here, sample size n = 9000

Sample proportion = $p = \frac{9240}{9000} = 0.36$

$q = 1\text{-}p = 1 - \frac{9240}{9000} = 0.64$

$\alpha = 5\%$

95% confidence limits of population proportion = $p \pm Z_{\alpha/2}\sqrt{\frac{PQ}{n}} = p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$

$= 0.36 \pm 1.96\sqrt{\frac{0.36 \times 0.64}{9000}} = 0.36 \pm 0.0099$

Taking (-) sign 0.36 – 0.0099 = 0.3501

Taking (+) sign) 0.36 + 0.0099 = 0.3699

Hence 95% confidence limit for population proportion is 0.3501 to 0.3699.

Example 7: A random sample of 500 pineapples was taken from a large consignment and 65 of them were found to be bad. Show that the Standard Error of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

(b) Also obtain 95% confidence limits for population proportion P if no estimate is available.

Solution:

(a) Here sample size (n) = 500.

Number of bad pineapples in the sample(x) = 65

Proportion of bad pineapples in the sample (p) = $\frac{x}{n} = \frac{65}{500} = 0.13$

$q = 1 - p = 1 - 0.13 = 0.87$

The standard error of proportion of bad pineapples in the sample is given by: S.E. (p) = $\sqrt{\frac{PQ}{n}}$ where P is the population proportion of bad pineapples in the whole consignment. Since P is not known, we use its unbiased estimate provided by the sample proportion p. Thus

Est. [S.E. (p)] = $\sqrt{\frac{0.13 \times 0.87}{500}} = 0.015.$

The most probable limits or the population proportion of bad pineapples are:

$p \pm 3$ S.E. (p) = $0.130 \pm 3 \times 0.015 = 0.130 \pm 0.045$ = (0.085 and 0.175)

Hence the percentage of bad pineapples in the consignment lies almost surely between the limits 8.5 and 17.5.

(b) If no estimate of P is available, then 95% confidence limits for P are given by

$$p \pm Z_{\alpha/2}\, S.E.\,(p) = p \pm 1.96\, S.E.\,(p) = p \pm 1.96\sqrt{\frac{pq}{n}}$$

$$= 0.130 \pm 1.96 \times 0.015 = 0.130 \pm 0.044 = (0.086 \text{ and } 0.174).$$

Example 8: A random sample of 700 units from a large consignment showed that 200 were damaged. Find (i) 95% and (ii) 99% confidence limits for the proportion of damaged units in the consignment.

Solution:

(i) We are given sample size $(n) = 700$.

Proportion of damaged units in the sample (p) $= \frac{x}{n} = \frac{200}{700} = 0.28$ then $q = 1 - p$

$= 1 - 0.28 = 0.72$

Hence an estimate of Standard Error of p is given by

$$S.E.(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.28 \times 0.72}{700}} = 0.017$$

95% confidence Limits for P, i.e., proportion of damaged units in the consignment are given by:

$$= p \pm Z_{a/2}\sqrt{\frac{pq}{n}} = p \pm 1.96 = 0.286 + 1.96 \times 0.017$$

$$= 0.286 \pm 0.033 = (0.253, 0.319).$$

Hence 95% confidence limit for proportion of damaged is 25.3% to 31.9%

(ii) 99% Confidence Limits for P are

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}} = p \pm 2.58\sqrt{\frac{pq}{n}}$$

$$= 0.286 \pm 2.58 \times 0.017 = 0.286 \pm 0.044 = (0.242 \text{ and } 0.330).$$

Hence 99% confidence limit for proportion of damaged is 24.2% to 33%.

Example 9: Out of 20,000 customers' ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing where in 45 mistakes were found. Assign limits within which the number of defective cases can be expected at 95% level.

Solution:

Here we are given

Sample size $(n) = 600$,

Population size $(N) = 20{,}000$.

Number of mistakes in the sample ledger accounts $(x) = 45$

Proportion of mistakes in the sample (p) $= \frac{x}{n} = \frac{45}{600} = 0.075$

$q = 1 - 0.075 = 0.925$

95% confidence limits for population proportion P are given by:

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n-1}}\sqrt{\frac{N-n}{N}}$$

$$= 0.075 \pm 1.96\sqrt{\frac{0.075 \times 0.925}{600-1}}\sqrt{\frac{20000-600}{20000}}$$

$$= 0.075 \pm 1.96 \times 0.01 \times 0.97$$

$$= 0.075 \pm 0.019 = (0.056, 0.094)$$

Hence, the number of defective cases in a lot of 20,000 are expected to lie between 20,000 × 0.056 and 20,000 × 0.094 i.e., 1,020 and 1,880.

Determination of Sample Size

Estimation of sample size by using mean

Let $\bar{x}$ be the sample mean from a random sample of size n drawn from population with mean μ and standard deviation σ.

Now, $Z = \dfrac{\bar{x} - E(\bar{x})}{SE(\bar{x})} = \dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

At α level of significance $(1-\alpha)$ confidence limit is

$$P\left(\left|\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| \le Z_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \quad P\left(|\bar{x}-\mu| \le \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right) = 1-\alpha$$

Now, $\bar{x} - \mu = d$ (margin of error) then

$$d = \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$$

$$\Rightarrow \quad \sqrt{n} = \frac{\sigma}{d} Z_{\alpha/2}$$

$$\Rightarrow \quad n = \left[\frac{\sigma}{d} Z_{\alpha/2}\right]^2 = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2}$$

⇒ In case of σ is not known take $\sigma = s$

For the finite population of size N, sample size $= \dfrac{\sigma^2 Z^2_{\alpha/2}}{d^2 + \frac{\sigma^2 Z^2_{\alpha/2}}{N}} = \dfrac{n}{1 + \frac{n}{N}}$

Example 10: In measuring reactions time, a psychologist estimates that the standard deviation is 0.05 seconds. How large a sample of measurement must be taken in order to be 99% confident that the error of his estimate will not exceed 0.01 seconds?

Solution:

Here

Sample size (n) = ?

Standard deviation (s) = 0.05

Confidence interval $(1 - \alpha) = 99\% = 0.99$

or $\alpha = 0.01$

$Z_{\alpha/2} = 2.58$

Error (d) = 0.01

Here $\sigma = s$

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2} = \frac{(0.05)^2 \times (1.96)^2}{(0.01)^2} = 166.4 \approx 167$$

Hence required sample size is 167.

Example 11: The mean systolic blood pressure of a certain group of people was found to be 125 mm of Hg with standard deviation of 15 mm of Hg. Calculate sample size to verify the result at 5% level of significance if error do not exceed 2. Also find sample size if sample is selected from population of size 500.

Solution:

Here

Standard deviation $(\sigma) = 15$

Level of significance $(\alpha) = 5\%$

Then $Z_{\alpha/2} = 1.96$

Sample size (n) = ?

Error (d) = 2

Here $\sigma = s$

Now,

$$n = \frac{\sigma^2 Z^2_{\alpha/2}}{d^2} = \frac{(15)^2 \times (1.96)^2}{(2)^2} = 216.09 \approx 216$$

When N = 500

$$\text{Sample size} = \frac{n}{1 + \frac{n}{N}} = \frac{216}{1 + \frac{216}{500}} = 150.83 \approx 151$$

Example 12: Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% level of confidence when coefficient of variation is 60%.

Solution:

Here, C.V. = 60% = 0.6

$$P(|\bar{x} - \mu| \leq 0.1\mu) = 0.95 \quad ......(i)$$

Confidence level $(1-\alpha) = 95\% = 0.95$ then $\alpha = 0.05$

Now, $P\left(|\bar{x} - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right) = 1 - \alpha$

$$\Rightarrow \quad P\left(|\bar{x} - \mu| \leq 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95 \quad ......(ii)$$

From equation (i) and (ii)

$$0.1\mu = 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \sqrt{n} = \frac{1.96}{0.1} \times \frac{\sigma}{\mu}$$

$$\Rightarrow n = (1.96/0.1 \times \sigma/\mu)^2$$

$$\Rightarrow n = 384.16 \times CV^2$$

$$\Rightarrow n = 384.16 \times (0.6)^2$$

$$\Rightarrow n = 138.29 \approx 138$$

Hence required sample size is 138.

Estimation of Sample Size by Using Proportion

Let p be sample proportion from random sample of size n drawn from population with proportion P

Now,

$$Z = \frac{p - E(p)}{SE(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

At α level of significance $(1-\alpha)$ confidence limit is

$$P\left(\left|\frac{p - P}{\sqrt{\frac{PQ}{n}}}\right| \leq Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P(|p - P| \leq \sqrt{\frac{PQ}{n}}\, Z_{\alpha/2} = 1 - \alpha$$

Now, $p - P = d$ (margin of error) then

$$d = \sqrt{\frac{PQ}{n}}\, Z_{\alpha/2}$$

$$\Rightarrow \sqrt{n} = \frac{Z_{\alpha/2}\sqrt{PQ}}{d}$$

$$\Rightarrow n = \frac{PQ\, Z^2_{\alpha/2}}{d^2}$$

In case of P is not known it is estimated by sample proportion p.

$$\Rightarrow n = \frac{pq\, Z^2_{\alpha/2}}{d^2}$$

For the finite population of size N, sample size $= \dfrac{PQ\, Z^2_{\alpha/2}}{d^2 + \dfrac{PQ\, Z^2_{\alpha/2}}{N}} = \dfrac{n}{1 + \dfrac{n}{N}}$

Example 13: A researcher wants to conduct a survey of disabled at Kathmandu valley. What should be the sample size if the prior estimate of population of disables in the population is 10% and the desired error is estimation is 2% and level of significance is 5%.

Solution:

Here

Sample size (n) = ?

Population proportion (P) = 10% = 0.1

Q = 1 - P = 0.9

Error (d) = 2% = 0.02

Level of significance (α) = 5%

$$\text{Estimated sample size } n = \frac{Z^2_{\alpha/2}\, P.Q.}{d^2} = \frac{(1.96)^2 \times 0.1 \times 0.9}{(0.02)^2} = 864.36 \approx 865$$

Hence required sample size is 865.

Example 14: For p = 0.2, d = 0.05 and z = 2 find n. Also find n if N = 1000.

Solution:

Now,

$$n = \frac{Z^2_{\alpha/2}\, pq}{d^2}$$

$$= \frac{4 \times 0.2 \times (1 - 0.2)}{0.05^2} = 256$$

When N = 1000

$$\text{Sample size} = \frac{n}{1 + \frac{n}{N}}$$

$$= \frac{256}{1 + \frac{256}{1000}}$$

$$= 203.82 \approx 204.$$

Relationship of Sample Size with Desired Level of Error

For estimation of unknown parameters population mean μ based on the sample statistic (i.e. sample mean), we wish to consider the relationship between the error, risk and sample size. Since the sampling distribution of sample means for large samples are normally distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, the standard normal variate is defined by,

$$Z_\alpha = \frac{\overline{X} - \mu}{S.E(\overline{X})} = \frac{d}{S.E(\overline{X})} = \frac{d}{\frac{\sigma}{\sqrt{n}}}$$

Where,

$\overline{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

Z_α = standard normal variate associate with risk α.

α = the risk that sample mean ($\overline{X}$) differ by population mean μ.

Example 15: If the population proportion of success is 0.65 and n = 100, what will be the value of sampling error when acceptance region is 0.95?

Solution:

Here,

Population proportion of success (P) = 0.65

$\therefore$ Q =1 - P = 0.35

Sample size (n) = 100

Sampling Error (d) = ?

Significant level (α) = 1 - 0.95 = 0.05

z_α= 1.96 [Two tailed]

We know,

$$n = \left(\frac{Z_{\alpha/2}}{d}\right)^2 PQ$$

or, $$100 = \left(\frac{1.96}{d}\right)^2 0.65\times0.35$$

Taking root both sides, we get;

or, $$10 = \frac{1.96}{d}\sqrt{0.65x0.35}$$

or, $$10d = 1.96\sqrt{0.65\times0.35}$$

or, $$d = \frac{0.935}{10} = 0.0935$$

$\therefore$ The value of sampling error = 0.0935.

EXERCISE

1. What do you mean by sampling distribution? Why this study is important in statistics?
2. Elaborate the statement "The central limit theorem forms the basis of inferential statistics".
3. Describe the different criteria of good estimator.
4. Explain how sample size affects the margin of error with examples.
5. A sample of size 25 is drawn from a finite population consisting of 150 units. If the population standard deviation is 10, find the standard error of sample mean. **Ans: 1.825**
6. A random sample of 64 projector lamp indicated a sample average life of 3500 hours. The sample standard deviation of life is 200 hours. Then calculate the standard error of average life of projector lamp. **Ans: 25**
7. A simple random sample of size 20 is drawn without replacement from a finite population of 75 units, if the number of defective units in the population is 12. Ascertain standard error of the sample proportion. **Ans: 0.07**
8. A manager has sent 300 SMS for a mass meeting and checked 25 of them either they were correctly delivered or not. While checking he found that 15 were not correctly delivered. Find out standard error of sample proportion of SMS incorrectly delivered. **Ans: 0.095**
9. A candidate prepares for a local election. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Calculate the standard error for the difference in support that this candidate getting in town A and B. **Ans: 0.0765**
10. Internet connections are often slowed by delays at nodes. Five hundred packets were sent through the same network between 5 to 6 pm and after four hours i.e between 10 to 11 pm 300 packets were sent. The earlier sample showed mean delay time of 0.8 second with standard deviation of 0.1 second whereas the second sample showed mean delay time of 0.5 seconds with a standard deviation of 0.08 second. Calculate the standard error of difference between two sample means. **Ans: 0.0086**
11. A random sample of size 64 has been drawn from a population with standard deviation 20. The mean of the sample is 80. Calculate 95% confidence limit for the population mean. How does the width of the confidence interval change if sample size is 256 instead? **Ans: 75.1, 84.9, 77.55, 82.45**
12. A random sample of size 65 was taken to estimate the mean life of 1000 laptop batteries and the mean and standard deviation were found to be 6300 hours and 9.5 hours respectively. Find a 95% confidence interval for the population mean. **Ans: 6297.75, 6302.25**

13. It is observed that 28 successes in 70 independent Bernoulli trial. Compute 90% confidence interval for population proportion. **Ans: 0.304, 0.496**

14. In a random sample of 400 chips from a large consignment, 20 items were found to defective. Find 99% confidence limits for the percentage of defective chips in the consignment. **Ans: (0.023, 0.077)**

15. In laboratory experiment, for the test of a material in good condition, a sample of 400 units was drawn. When they were tested, 80 were good. Find 95% confidence limits for the percentage of good. **Ans: (0.161, 0.239)**

16. In a sample survey of 100 professionals in a city, 23% preferred a particular brand of laptop. Find 99% confidence limits for percentage of all professionals in the city preferring the brand of laptop. **Ans: (0.122, 0.338)**

17. A factory is producing 50000 CD daily from a sample of 500 CD, 2% were found to be of substandard quality. Estimate the percentage of CD that can be reasonable expected to spoiled in the daily production at 95% confidence level. **Ans: (0.0078, 0.0322)**

18. A random sample of 100 defective computers of a university, 75 were successfully repaired and 25 were unable to repair due to motherboard problems. Find 95% confidence limits for the percentage of computer which can be repaired in that university. **Ans: 66.5%, 83.5%**

19. A sample of size 100 produces the sample mean 16. Assuming population standard deviation 3, compute 95% confidence interval for population mean.

20. Assuming population standard deviation 3, how large should a sample be to estimate population mean with margin of error not exceeding 0.5? **Ans: 138**

21. The principle of a college wants to estimate the proportion of students who were interested to develop startup. What size of a sample should he select so as to have the difference of proportion of interested students with true mean not to exceed by 10% with almost certainty? It is believed from previous records that the proportion of interested students was 0.30? **Ans: 189**

22. Mr. X wants to determine the average time to complete a project the past records show that population standard deviation is 10 days. Determine the sample size so that he may be 95% confident that sample average remains within ± 2 days of the averages. **Ans: 96**

23. The average time taken by server to execute an algorithm varies from time to time. From the past experience it is known that the time taken is normally distributed with standard deviation of 6.7 minutes. The IT manager wishes to estimate the average by drawing a random sample such that the probability is 0.95 that the mean of the sample will not deviate by more than 1 minute from the population mean. What should be sample size? **Ans: 173**

Use of Statistical Software

Find confidence interval of mean assuming normal distribution for following data.

height	78	55	68	48	65	76	57	55	65	75	51	61	68	67	76
	78	71	56	57	67	58	51	50	58	50	77	55	48	70	55	58
	70	56	52	74	61	69	76	61	68	78	56	78	57	66	66	74
	66	48	73	71	70	62	74	76	50	69	75	65	48			

Solution:

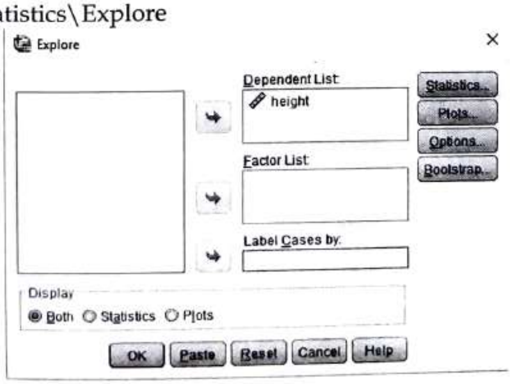
Using excel

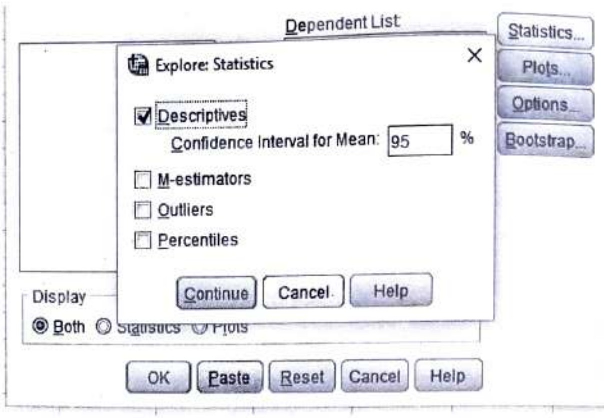
Place the data in A2:A61

	A	B	C	D	E	F	G
1	height			symbol	value	formula	
2	78		sample size	n	60	=COUNT(A2:A61)	
3	55		sample mean	ȳ	63.8833	=AVERAGE(A2:A61)	
4	68		sample st dev	s	9.55287	=STDEV.S(A2:A61)	
5	48		alpha	α	0.05		
6	65		z value	$z_{\alpha/2}$	1.95996	=NORM.S.INV(1-E5/2)	
7	76						
8	57		direct method				
9	55		margin of error	CI/2	2.41717	=CONFIDENCE.NORM(E5,E4,E2)	
10	65		lower limit	LCI	61.4662	=E3-E9	
11	75		upper limit	UCI	66.3005	=E3+E9	
12	51						
13	61		formula method				
14	68		lower limit	LCI	61.4662	=E3-E6*E4/SQRT(E2)	
15	67		upper limit	UCI	66.3005	=E3+E6*E4/SQRT(E2)	
16	76						
17	78						
18	71						
19	56						

Using SPSS

Analyze\Descriptive Statistics\Explore





Descriptive

			Statistic	Std. Error
height	Mean		63.88	1.233
	95% Confidence Interval for Mean	Lower Bound	61.42	
		Upper Bound	66.35	
	5% Trimmed Mean		63.98	
	Median		65.50	
	Variance		91.257	
	Std. Deviation		9.553	
	Minimum		48	
	Maximum		78	
	Range		30	
	Interquartile Range		17	
	Skewness		-0.136	0.309
	Kurtosis		-1.258	0.608

Note: SPSS uses t statistics to determine confidence interval.

Using STATA

ci means height

```
. ci means height
```

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
height	60	63.88333	1.23327	61.41557	66.3511

❑❑❑

2
Chapter

TESTING OF HYPOTHESIS

CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Types of statistical hypotheses
- Power of the test, concept of p-value and use of p-value in decision making
- Steps used in testing of hypothesis,
- One sample tests for mean of normal population (for known and unknown variance)
- Test for single proportion, test for difference between two means and two proportions, paired sample t-test
- Linkage between confidence interval and testing of hypothesis problems
- Problems and illustrative examples related using software.

Introduction

One of the important application of statistical inference is the testing of hypothesis. Modern theory of probability plays an important role in decision making and the branch of statistics which helps us in arriving at the criterion for such decisions is called testing of hypothesis. It was initiated by J. Neyman and E.S. Pearson. It employees statistical techniques to arrive at decisions in certain situation where there is an element of uncertainty on the basis of sample whose size is fixed in advance. We draw conclusion about the population parameter on the basis of sample.

According to Webster hypothesis is defined as "A tentative theory or supposition provisionally adopted to explain certain facts and to guide in the investigation of others"

Hypothesis

A hypothesis is a tentative theory or supposition provisionally adopted to explain certain facts and to guide in the investigation of others.

A statistical hypothesis which is tentative statement or supposition about the estimated value of one or more parameter of the population is called parametric hypothesis .A statistical hypothesis about attributes is called non parametric hypothesis.

If a hypothesis completely determines the population, it is called a simple hypothesis, otherwise composite hypothesis.

In testing of hypothesis a statistic is computed from a sample drawn from the parent population and on the basis of the statistic it is observed whether the sample so drawn has come from the population with certain specified characteristic.

Terminology are the different terms used in testing of hypothesis. By the knowledge of terminology it becomes easier to know the terms during the numerical calculation.

Types of Statistical Hypothesis

Null hypothesis

The supposition about the population parameter is called null hypothesis. It is set up for testing a statistical hypothesis only to decide whether to accept or reject the null hypothesis. According to R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that is true.

It is the hypothesis of no difference between sample statistic and parameter. It is hypothesis of no difference between parameters.

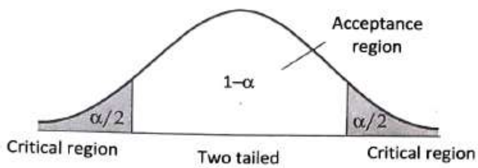
Null hypothesis is denoted by H_0 .It is set up as $H_0 : \mu = \mu_0$

Suppose we want to test the average score of students in B.Sc. entrance exam is 55 then to start testing the hypothesis we assume the average score is 55.There is no difference between sample average and population average. Then the null hypothesis is $H_0 : \mu = 55$

Alternative Hypothesis

A hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. Any hypothesis which is not null is also called alternative hypothesis. It is hypothesis of difference between sample statistic and parameter. It is hypothesis of difference between parameters.

Alternative hypothesis is denoted by H_1.



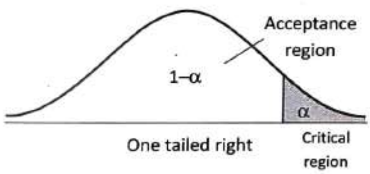
Alternative hypothesis are

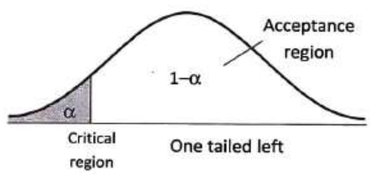
(i) two tailed

(ii) one tailed right

(iii) one tailed left

Alternative hypothesis is set up as H_1: $\mu \neq \mu_0$ for two tailed or H_1: $\mu > \mu_0$ for one tailed right or H_1: $\mu < \mu_0$ for one tail left.





The setting of alternative hypothesis is very important since it enable us to decide whether we have to use two tailed or one tailed test.

If the null hypothesis is rejected in the above example then average score of students in B.Sc. entrance exam is different from 55 or greater than 55 or less than 55. Among these only one situation arise. Then the alternative hypothesis is H_1:$\mu \neq 55$ or H_1: $\mu > 55$ or H_1: $\mu < 55$.

Errors in Hypothesis Testing

The main objective of sampling theory is to draw valid inference about the population parameters on the basis of the sample. We decide to accept or reject the hypothesis after examining result from sample. In testing of hypothesis two types of errors are introduced.

Type I Error

It is the error of rejecting null hypothesis H_0 when it is true. The probability of type I error is denoted by α, called the level of significance.

α = Probability (Type I error) = Probability (Rejecting H_0/H_0 is true).

Type II Error

It is error of accepting null hypothesis H_0 when it is false. It means it is error of accepting null hypothesis H_0 when alternative hypothesis H_1 is true. The probability of type II error is denoted by β.

β = Probability (Type II error) = Probability (Accepting H_0/ H_1 is true).

		Truth	
		H_0 is true	H_0 is false
Decision	Accept H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

In testing of hypothesis, accepting H_0 when H_0 is true is the correct decision and also rejecting H_0 when H_0 is false is correct decision .The remaining conditions rejecting H_0 when H_0 is true and accepting H_0 when H_0 is false are errors.

The probability of type I error (α) and the probability of type II error (β) are called producer's risk and consumer's risk respectively.

Power of Test (1-β)

β is the probability of accepting H_0 when H_0 is false. It is the probability of wrong decision. The probability of rejecting H_0 when H_0 is false is probability of correct decision is called power of test and is denoted by $1-\beta$. It is also called probability of accepting H_1 when H_1 is true. It is the probability of not making type II error.

Test statistic

The test statistic is the statistic based upon appropriate probability distribution. It is used to test whether the null hypothesis set up should be accepted or rejected .It helps us to decide whether to accept or reject the null hypothesis. Different probability distribution values are used in appropriate cases while testing hypothesis. The commonly used test statistic are

Z test: We use Z distribution under the normal curve for large sample(sample size n>30)

The Z test statistic is $Z = \frac{t - E(t)}{S.E.\ (t)} \sim N(0,1)$ as $n \to \infty$

t test :We use t distribution for small sample (sample size < 30)

The t test statistic is $t = \frac{\bar{x} - \mu}{S.E.\ (\bar{x})} \sim$ t distribution with n-1 degree of freedom.

Level of Significance

In testing of hypothesis two types of errors are introduced namely type I error and type II error. The two errors are inversely related. Both errors can not be minimized at the same time. Usually we fix type I error and minimize type II error. The maximum size of type I error prepared in testing of hypothesis is called level of significance. It is denoted by α. Commonly used level of significance are 5%, 1%.

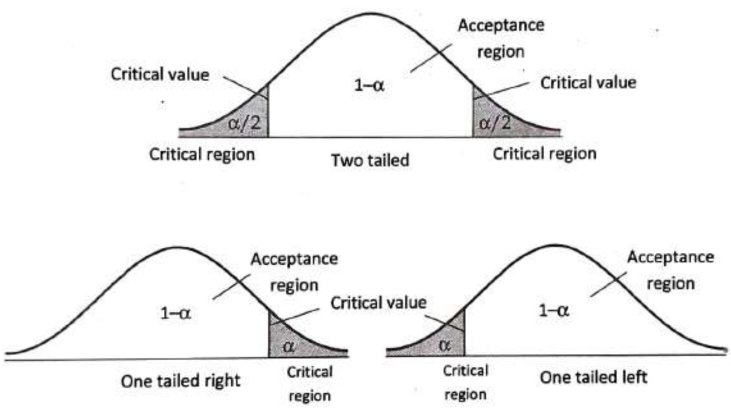
The level of significance should be chosen on the basis of power of test. If the power of test is too low then high level of significance should be chosen. In such case level of significance are 10%, 20%.

Critical Region

It is also called the rejection region. The set of all possible values of statistic is divided into two regions one leading to the rejection of H_0 and other to the acceptance of H_0. The division is made on the basis of level of significance and H_1. The region which leads to the rejection of H_0(null hypothesis) is called rejection region and is denoted by ω .While those region which leads to the acceptance of H_0 is called acceptance region and is denoted by $\bar{\omega}$.

If the test statistic falls into the rejection region, the null hypothesis is rejected. If the test statistic falls into the acceptance region, the null hypothesis is accepted.

The critical region is situated on both the tail or any one tail depending upon the alternative hypothesis.



Critical Value

It is also called significant value or tabulated value. The value of statistic which separates critical region and acceptance region is called critical value. It depend upon the level of significance and alternative hypothesis.

The critical value of Z for a singled tailed test at a level of significance α is the same as the critical value of Z for a two tailed test at a level of significance 2α.

For right tailed test $P(Z > Z_\alpha) = \alpha$

For left tailed test $P(Z < -Z_\alpha) = \alpha$

For two tailed test $P(|Z| > Z_\alpha) = \alpha$

or $\quad P(Z > Z_\alpha) + P(Z < -Z_\alpha) = \alpha$

or $P(Z > Z_\alpha) + P(Z > Z_\alpha) = \alpha$

or $2P(Z > Z_\alpha) = \alpha$

or $P(Z > Z_\alpha) = \frac{a}{2}$

Area of each tail is $\frac{\alpha}{2}$ in two tailed test.

Degree of Freedom

The number of independent variates which make up statistic is called degree of freedom. It is simply denoted by d.f.

One Tailed Test

A test of statistical hypothesis in which the alternative hypothesis H_1 looks for a definite increase (right tail) or definite decrease (left tail) in parameter is called one tailed test.

$H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ (right tail)

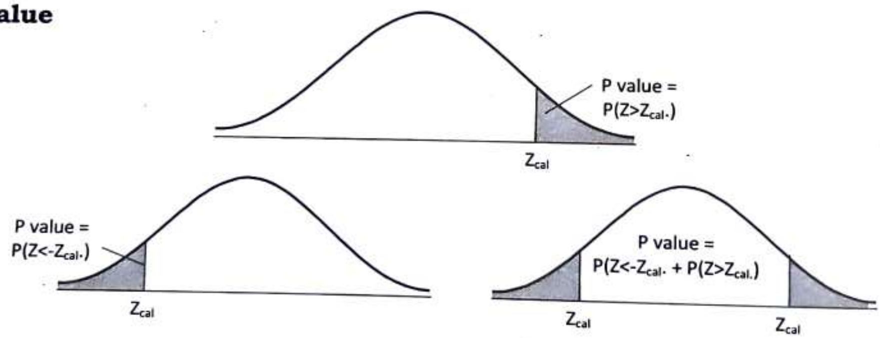
$H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$ (left tail)

Two Tailed Test

A test of statistical hypothesis in which the alternative hypothesis looks for a definite change in the parameter is called two tailed test.

$H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

P Value



It is the probability value of tail area under the curve of the test statistic used in hypothesis. Instead of critical value we can make decision on testing of hypothesis using p value.

The probability of obtaining a test statistic at least as extreme as the one that was actually observed assuming H_0 is true is called p value.

If p value is less than α then reject H_0. It is simply a measure of how likely the data were to have occurred by chance assuming H_0 is true

For Z test, p value = Probability ($Z \geq |Z_{calculated}|$)

For right tailed test, p value is the area to the right of the computed value of the test statistic under H_0

For left tailed test, p value is the area to the left of the computed value of test statistic under H_0.

For two tailed test, p value is twice the area to the right of computed value of the test statistic or twice the area to the left of computed value of the statistic under H_0.

If p value < α then reject H_0 at α level of significance, accept otherwise.

Steps Use in Testing of Hypothesis

The following are various steps in testing of hypothesis in a systematic manner;

1: Set up null hypothesis H_0 .

2: Set up alternative hypothesis H_1. In H_1 decide to use whether one tailed or two tailed.

3: Choose the appropriate level of significance α depending upon the reliability of the estimates and permissible risk.

4: Identify the sample statistic to be computed and its sampling distribution.

5: Compute the test statistic under H_0.

6: Obtain the critical value of the test statistic from the appropriate table.

7: Compare the calculated value of test statistic with the critical value and then accordingly the decision to accept or reject H_0 is made.

For Z test, accept H_0 if $|Z| < Z_{tabulated}$, otherwise reject H_0.

Relationship between Hypothesis Testing and Confidence Interval

There is ordinarily close relationship between a test of hypothesis concerning a parameter or parameters and the corresponding confidence interval. To illustrate this relationship we will examine the $(1-\alpha)$ level confidence interval for μ, the mean of normal distribution with known variance σ^2 and sample size n.

The interval limits are $\bar{x} \pm Z_{tabulated}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$

This is acceptance region for testing $H_0 : \mu = \mu_0$ in two tailed test at α level of significance for known σ^2 and sample size n . This indicates that the value μ lying within the $(1-\alpha)$ confidence interval corresponds to a value of test statistic $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ that would lead to acceptance of the hypothesis. If we start with a value of μ lying outside the confidence interval, we find that corresponding value of the test statistic will fall into the rejection region.

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Here, $\bar{x} - Z\alpha_{/2}\frac{\sigma}{\sqrt{n}} \le \mu_0$

Or, $\bar{x} \le \mu_0 + Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$ .....(i)

Also, $\mu_0 \leq \bar{x} + Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$

Or, $\mu_0 - Z\alpha_{/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x}$ .....(ii)

From i and ii

$$\mu_0 - Z\alpha_{/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$$

The confidence limit for the statistic $\bar{x}$ of two tailed test is $\mu_0 \pm Z\alpha_{/2}\frac{\sigma}{\sqrt{n}}$. If $\bar{x}$ is found to be within the interval then H_0 is accepted, otherwise H_0 is rejected.

Large Sample Tests (Z test)

It is important parametric test based upon the normality assumption. When the samples are selected from population of known parameter with sample size more than 30 Z test is used. We consider that if sample size is more than 30 then sample selected from non normal population is also approximately normal distributed.

Z test is defined as the ratio of difference between t and E(t) to the S.E.(t)

$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0, 1)$, where t = statistic , E(t) = Expected value of statistic and S.E.(t) = Standard error of the statistic.

Z test is used to test

- significance of single mean.
- significance of difference between two means.
- significance of single proportion.
- significance of difference between two proportions

One Sample Test for Mean of Normal Population for Known Variance

Let us consider sample of size $n(n > 30)$ has been drawn from the normal population with known variance $N(\mu,\sigma^2)$ then the sample mean $\bar{X} \sim N(\mu,\sigma^2/n)$.

Different steps in the test are;

Problem to test

$H_0: \mu = \mu_0$ (sample is drawn from population with mean μ_0)

$H_1: \mu \neq \mu_0$ (Two tailed test) or $H_1: \mu > \mu_0$ (One tailed right) or $H_1: \mu < \mu_0$ (One tailed left)

Test statistic

For the sample selected from the population of unknown size

$$Z = \frac{\bar{X} - E(\bar{X})}{S.E.\ (\bar{X})} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ for known variance}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ for unknown variance (for large sample size } \hat{\sigma} = s)$$

For the sample selected from the population of known size

$$Z = \frac{\bar{X} - E(\bar{X})}{S.E.\ (\bar{X})} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}} \quad \text{for known variance}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}} \quad \text{for unknown variance}$$

Where $\bar{X}$ = sample mean, μ = population mean, σ = population s.d., s = sample s.d., N = population size, n = sample size

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject $|Z| > Z_{tabulated}$, accept otherwise.

Using p value approach

p value = Probability ($Z > |Zcal|$) (It can be obtained from standard normal table)

Decision

Reject H_0 at α level of significance if $p < \alpha$ for one tail

$2p < \alpha$ for two tail

Example 1: A sample of 400 students is found to have mean height of 170 cm. Can it be reasonably regarded as a sample from a large population with mean height 169.5 cm and standard deviation 3.5 cm?

Solution:

Here

Sample size (n) = 400, Sample mean height ($\bar{x}$) = 170,

Population mean (μ) = 169.5, Population SD (σ) = 3.5

Problem to test

H_0: Mean height of students is 169.5 cm ($\mu = 169.5$)

H_1: Mean height of students is not 169.5 ($\mu \neq 169.5$) (Two tailed)

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{170 - 169.5}{\frac{3.5}{\sqrt{400}}} = \frac{0.5 \times 20}{3.5} = 2.857$$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value for two tailed test is $Z_{tab} = Z_{\alpha/2} = 1.96$.

Decision

Here $Z = 2.857 > Z_{tab} = 1.96$, reject H_0 at 5% level of significance.

Conclusion

The sample of 400 students can not be regarded as sample from large population with mean height 169.5 cm and standard deviation 3.5 cm.

Using p value approach

$p = \text{Prob}(Z > 2.857) = 0.0021$

$2p = 2 \times 0.0021 = 0.0042$

At $\alpha = 5\% = 0.05$

$2p = 0.0042 < \alpha = 0.05$

Reject H_0 5% level of significance.

Example 2: A manufacturer claim that their widget is more reliable than their main competitors. In order to verify this a sample of 40 widget from manufacturer's range was taken. The mean pass rate found to be 992 per 1000 with s.d. of 15 per 1000. Previous studies have shown that the mean pass rate of all widget on the market is 979 per 1000. Test the manufacturer's claim at 1% level of significance.

Solution:

Here,

Sample size (n) = 40, Sample mean ($\bar{x}$) = 992, Sample Sd (s) = 15

Population mean (μ) = 979, Level of significance (α) = 1%

Problem to test

H_0: Average pass is 979 ($\mu = 979$)

H_1: Average pass is more than 979 ($\mu > 979$) (one tailed right)

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{992 - 989}{\frac{15}{\sqrt{40}}} = 5.48$$

Critical value

At $\alpha = 1\% = 0.01$ critical value for one tailed test is $Z_{tab} = Z_{\alpha} = 2.32$

Decision

Here $|Z| = 5.48 > Z_{tab} = 2.32$, reject H_0 at 1% level of significance.

Conclusion

The claim of manufacture is correct.

Example 3: A random sample of 100 pen drive selected from a batch of 2000 pen drive shows that the average thickness of the pen drive is 0.354 with a standard deviation 0.048 Are the samples from the lot having average thickness 0.35?

Solution:

Here, Sample size (n) = 100, Population size (N) = 2000

Sample mean $(\bar{x}) = 0.354$, Sample SD (s) = 0.048, Population mean $(\mu) = 0.35$

Problem to test

H_0: Average thickness of pen drive is 0.35 $(\mu = 0.35)$

H_1: Average thickness of pen drive is not 0.35 $(\mu \neq 0.35)$ (Two tailed)

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

$$= \frac{0.354 - 0.35}{\frac{0.048}{\sqrt{100}}\sqrt{\frac{2000-100}{2000-1}}}$$

$$= \frac{0.004}{0.0047} = 0.854$$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value for two tailed test is $Z_{tab} = Z_{\alpha/2} = 1.96$.

Decision

Here $Z = 0.854 < Z_{tab} = 1.96$, accept H_0 at 5% level of significance.

Conclusion

Sample are from lot having average thickness of pen drive 0.35.

Example 4: If the mean breaking strength of a copper wire is 575 lbs with s.d. of 8.3 lbs. How many large sample must be used in order that there be one chance in 100 that the mean breaking strength of the sample is less than 572 lbs?

Solution:

Here,

Population mean $(\mu) = 575$, Population SD $(\sigma) = 8.3$, Sample size (n) = ?

P (mean breaking strength < 572) $= \frac{1}{100}$, Sample mean $(\bar{X}) = 572$, $\alpha = 1\%$

Problem to test

H_0: Mean breaking length of copper wire is 575 lbs ($\mu = 575$)

H_1: Mean breaking length of wire is less than 575 lbs ($\mu < 575$) (one tail left)

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{572 - 575}{\frac{8.3}{\sqrt{n}}} = -0.3614\sqrt{n}$$

$|Z| = 0.3614\sqrt{n}$

Critical value

At $\alpha = 1\% = 0.01$ critical value for one tailed test is Z tab = $Z_{\alpha/2} = 2.32$.

Here, $0.3614\sqrt{n} = 2.32$

or $\sqrt{n} = \frac{2.32}{0.3614} = 6.419$

or $n = 41.2 \approx 41$.

Example 5: An ambulance service claims that it takes on the average 8.9 minutes to reach to its destination in emergency calls. To check the claim, the agency which licenses ambulance services has them timed on 50 emergency calls getting a mean of 9.3 minutes with s.d. of 1.6 minutes. What can they conclude at the level of significance $\alpha = 0.05$? Use the confidence limit to make conclusion.

Solution:

Here, Population mean (μ) = 8.9, Sample size (n) = 50, Sample mean ($\bar{x}$) = 9.3,

Sample Sd (s) = 1.6, Level of significance (α) = 0.05

Problem to test

H_0: Average time taken by ambulance to reach destination is 8.9 min ($\mu = 8.9$)

H_1: Average time taken by ambulance to reach destination is not 8.9 min ($\mu \neq 8.9$) (Two tailed)

Critical value

At $\alpha = 0.05$ critical value for two tailed test is $Z_{tab} = Z_{\alpha/2} = 1.96$.

The limits of the acceptance region for $\bar{x} = \mu \pm Z_{\alpha/2}\frac{s}{\sqrt{n}} = 8.9 \pm 1.96 \times \frac{1.6}{\sqrt{50}} = 8.9 \pm \frac{3.136}{7.07}$

$= 8.9 \pm 0.443$

Taking + sign, $\bar{x} = 8.9 + 0.443 = 9.343$

Taking – sign, $\bar{x} = 8.9 - 0.443 = 8.457$

Decision

$\bar{x} = 9.3$ lies between 8.457 and 9.343, accept H_0 at 5% level of significance.

Conclusion: The claim of ambulance service that it takes on average 8.9 minutes to reach the destination is true.

Example 6: The quality control manager at light bulb factory needs to determine whether the mean life of a large shipment of light bulbs is equal to 375 hours. The population standard deviation is 100 hours. A random sample of 100 light bulbs indicate a sample mean 350 hours At the 0.01 level of significance is there evidence that the mean life is different from 375 hours? Use p value method to draw conclusion.

Solution:

Here, Population mean (μ) = 375, population SD (σ) = 100, Sample size (n) = 100,

Sample mean ($\bar{x}$) = 350, level of significance (α) = 0.01

Problem to test

H_0 :Mean life of light bulb is 375 hours ($\mu = 375$)

H_1: Mean life of light bulb is not 375 hours ($\mu \neq 375$) (Two tailed)

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{350 - 375}{\frac{100}{\sqrt{100}}} = \frac{-25 \times 10}{100} = -2.5$$

Now, $\text{Prob}(Z \geq |Z_{cal}|) = \text{Prob}(Z \geq 2.5) = 0.5 - \text{Prob}(0 \leq Z \leq 2.5) = 0.5 - 0.4938 = 0.0062$

For two tailed test p value = $2\,\text{Prob}(Z \geq |Z_{cal}|) = 2 \times 0.0062 = 0.0124$

Here $\alpha = 0.01$

Decision

P value = $0.0124 > \alpha = 0.01$, accept H_0 at 0.01 level of significance.

Conclusion

Mean life time of light bulbs is not different from 375 hours.

Test of Significance Difference between Two Means

Let us consider two independent samples of size n_1 and n_2 be drawn from population having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means.

For large n_1 and n_2,

$\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$

$\bar{X}_2 \sim N(\mu_1, \sigma_2^2/n_2)$ then

$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$

Different steps in the test are

Problem to test

$H_0 : \mu_1 = \mu_2$ Two means are not significantly different.

$H_1: \mu_1 \neq \mu_2$ (two tailed) or $H_1 : \mu_1 < \mu_2$ (one tailed left) or $H_1 : \mu_1 > \mu_2$ (one tailed right)

Test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{S.E.\ (\bar{X}_1 - \bar{X}_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ when population means and variances are known

$$= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ when population variances are known (Under H_0, $\mu_1 = \mu_2$)

$$= \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ when population variances are unknown

for large sample size $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$

where $\bar{x}_1$ = sample mean of size n_1, $\bar{x}_2$ = sample mean of size n_2

σ_1^2 = population variance of first population, σ_2^2 = population variance of second population

s_1^2 = sample variance of first sample, s_2^2 = sample variance of second sample.

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|Z > Z_{tabulated}$, accept otherwise.

Example 7: In a random sample of 500 the mean is found to be 20. In another independent sample of 400 the mean is 15. Could the samples have been drawn from the same population with S.D. 4?

Solution:

Here, First sample size $(n_1) = 500$, First sample mean $(\bar{x}_1) = 20$,

Second sample size $(n_2) = 400$, Second sample mean $(\bar{x}_2) = 15$, Population SD $(\sigma_1 = \sigma_2) = 4$

Let, first population mean $= \mu_1$ and second population mean $= \mu_2$

Problem to test

H_0: There is no significant difference between two population $(\mu_1 = \mu_2)$

H_1: There is significant difference between two population $(\mu_1 \neq \mu_2)$ (Two tailed)

Test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(20 - 15)}{\sqrt{\frac{4^2}{500} + \frac{4^2}{400}}} = \frac{5}{\sqrt{0.032 + 0.04}} = \frac{5}{0.27} = 18.51$$

Critical value

Let $\alpha = 5\%$ be the level of significance the critical value for two tailed test is

$Z_{tabulated} = Z_{\alpha/2} = 1.96$

Decision

$Z = 18.51 > Z_{tabulated} = 1.96$, reject H_0 at 5% level of significance.

Conclusion

Samples are not drawn from same population with s.d.4.

Example 8: A random sample of 1000 workers from Pokhara show that their mean wages of \$47 per week with a standard deviation of \$ 28. A random sample of 1500 workers from Kathmandu show that their mean wage of \$ 49 per week with a standard deviation of \$ 40. Is there any significant difference between their mean level of wages?

Solution:

Here

Sample workers from Pokhara $(n_1) = 1000$

Sample mean wage from Pokhara $(\bar{X}_1) = 47$, Sample Sd of wage from Pokhara $(s_1) = 28$

Sample workers from Kathmandu $(n_2) = 1500$,

Sample mean wage from Kathmandu $(\bar{X}_2) = 49$, Sample Sd of wage from Kathmandu $(s_2) = 40$

Let μ_1 = Population mean wage of workers from Pokhara and μ_2 = Population mean wage of workers from Kathmandu.

Problem to test

H_0 : There is no significant difference in mean wages of workers between Pokhara and Kathmandu $(\mu_1 = \mu_2)$

H_1: There is significant difference in mean wage of workers between Pokhara and Kathmandu $(\mu_1 \neq \mu_2)$ (Two tailed)

Test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(47 - 49)}{\sqrt{\frac{28^2}{1000} + \frac{40^2}{1500}}} = \frac{-2}{\sqrt{0.748 + 1.0666}} = \frac{-2}{\sqrt{1.8506}} = -1.47$$

$Z| = 1.47$

Critical value

Let $\alpha = 0.05$ be the level of significance then the critical value for two tailed test is

$Z_{tabulated} = Z\,\alpha_{/2} = 1.96$

Decision

$|Z| = 1.47 < Z_{tabulated} = 1.96$, accept H_0 at 0.05 level of significance

Conclusion

There is no significant difference between mean wage of workers in Pokhara and Kathmandu.

Example 9: A sample survey is conducted to compare the weights of two makes of laser printer. A random sample of weight of 50 canon laser printer showed a mean weight 3.3 kg and s.d.1 kg. Another random sample of weight of 40 brother lazer printer showed a mean weight 2.81 kg and s.d. 0.5 kg. Do the data support the research hypothesis that the weight of canon laser printer on average heavier than the weight of brother laser printer? Use p value method at 1% level of significance.

Solution:

Here,

Sample size of canon printer (n_1) = 50, Sample mean weight of canon printer ($\overline{X}_1$) = 3.3,

Sample SD of weight of canon printer (s_1) = 1, Sample size of brother printer (n_2) = 40,

Sample mean weight of brother printer ($\overline{x}_2$) = 2.81, Sample SD of weight of brother printer (s_2) = 0.5, Level of significance (α) = 1%.

Let μ_1 = population mean weight of canon, μ_2 = population mean weight of brother

Problem to test

H_0: There is no difference in weight of canon and brother ($\mu_1 = \mu_2$)

H_1: Weight of canon is more than weight of brother ($\mu_1 > \mu_2$)

Test statistic

$$Z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3.3 - 2.81}{\sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}}} = \frac{0.49}{\sqrt{0.02 + 0.00625}} = \frac{0.49}{0.162} = 3.024$$

Now, $\text{Prob}(Z \geq Z_{\text{calculated}}) = \text{Prob}(Z \geq 3.024) = 0.5 - \text{Prob}(0 \leq Z \leq 3.024) = 0.5 - 0.49874 = 0.00126$

Here $\alpha = 1\% = 0.01$

It is one tail test, hence p value = $\text{Prob}(Z \geq Z_{\text{calculated}}) = 0.00126$

Decision

P value = $0.00126 < \alpha = 0.01$, reject H_0 at 1% level of significance.

Conclusion

The weight of canon laser printer on average heavier than the weight of brother laser printer.

Test for Single Proportion

Let P be the population proportion i.e. proportion of units possessing a certain characteristic in the population. Let a random sample of size n has been drawn from the population. Let X be the number of units possessing the characteristic in the sample then sample proportion is $p = X/n$ .For large n binomial distribution can be approximated by normal distribution.

$$X \sim N(nP, nPQ)$$

$$p = \frac{x}{n} \sim N\left(P, \frac{PQ}{n}\right), \; Q = 1 - P$$

Different steps in the test are;

Problem to test

$H_0: P = P_0$

$H_1: p \neq P_0$ (Two tailed test) or $H_1: P > P_0$ (One tailed right) or $H_1: P < P_0$ (One tailed left)

Test statistic

For the sample selected from the population of unknown size

$$Z = \frac{p - E(p)}{s.e.\ (p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

For the sample selected from the population of known size

$$Z = \frac{p - E(p)}{s.e.\ (p)} = \frac{p - P}{\sqrt{\frac{(N - n)}{(N - 1)} \frac{PQ}{n}}}$$

Where, p = sample proportion, P = population proportion, Q = 1 – P, N = population size,

n = sample size

Level of Significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical Value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|z| > Z_{\text{tabulated}}$, accept otherwise.

Example 10: A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin is unbiased?

Solution:

Here, Sample size (n) = 400, No of head turns up (x) = 216

Sample proportion of head(p) $= \frac{x}{n} = \frac{216}{400} = 0.54$

Population proportion of head (P) = 0.5

Problem to test

H_0: Coin is unbiased (P = 0.5)

H_1: Coin is biased ($P \neq 0.5$) (two tailed)

Test statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{400}}} = \frac{0.04 \times 20}{0.5} = 1.6$$

Critical value

Let $\alpha = 5\%$ be the level of significance the critical value for two tailed test is $Z_{\text{tabulated}} = Z_{\alpha/2} = 1.96$

Decision

$Z = 1.6 < Z_{tabulated} = 1.96$, accept H_0 at 5% level of significance.

Conclusion

The coin is unbiased.

Example 11: In a random sample of 400 persons from a large population 120 are females. Can it be said that males and females are in the ratio 5:3 in the population? Use 10% level of significance.

Solution:

Here,

Sample size (n) = 400, no of females(x) = 120

Sample proportion of female (p) $= \frac{x}{n} = \frac{120}{400} = 0.3$

Population proportion of female (P) $= \frac{3}{3+5} = \frac{3}{8} = 0.375$, $Q = 1 - P = 0.625$

Problem to test

H_0 : Male and female are in ratio 5:3 (P = 0.375)

H_1 : Male and female are not in ratio 5:3 ($P \neq 0.375$) (Two tailed)

Test statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.3 - 0.375}{\sqrt{\frac{0.375 \times 0.625}{400}}} = \frac{-0.075 \times 20}{\sqrt{0.23475}} = -3.099$$

Critical value

At $\alpha = 10\% = 0.1$ level of significance, then critical value for two tailed test is $Z_{tabulated} = Z_{\alpha/2}$ 1.645

Decision

$|Z| = 3.099 > Z_{tabulated} = 1.645$, reject H_0 at 10 % level of significance.

Conclusion

Males and females are not in ratio 5:3 in the population.

Example 12: A developer has claimed that at least 98% of the software which he supplied to a tribhuvan university conformed to specifications an examination of a sample of 500 software revealed that 30 were defective. Test the claim at a significance level of 0.01

Solution:

Here,

Population proportion of software conforming specification (P) = 0.98

Sample size (n) = 500

No of software conforming specification in sample (x) = 500 – 30 = 470

Level of significance (α) = 0.01

Sample proportion of software conforming specification (p) = $\frac{x}{n} = \frac{470}{500} = 0.94$

Problem to test

H_0 : Atleast 98% software confirmed specification ($P \geq 0.98$)

H_1 : Less than 98% software confirmed specification ($P < 0.98$)

(one tailed left)

Test statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.94 - 0.98}{\sqrt{\frac{0.98 \times 0.02}{500}}} = \frac{-0.04}{\sqrt{0.0000392}} = -6.45$$

Critical value

At $\alpha = 0.01$ level of significance, critical value for one tailed test is

$Z_{tabhlated} = Z_{\alpha} = 2.326.$

Decision

$|Z| = 6.45 > Z_{tabulated} = 2.326$, Reject H_0 at 0.01 level of significance.

Conclusion

The claim of the manufacturer that at least 98% software supplied to a Tribhuvan University conformed to specification is not correct.

Example 13: In a random sample of 600 cars making a right turn at a certain intersection, 157 pulled out into a wrong lane. Test the hypothesis that actually 30% of all drivers make this mistake at the given intersection at $\alpha = 0.05$ using p value.

Solution:

Here, Sample size of car (n) = 600, Car in wrong lane (x) = 157,

Sample proportion of car in wrong lane (p) = $\frac{x}{n} = \frac{157}{600} = 0.2616$

Population proportion (P) = 0.3, Q = 1 – P = 0.7, $\alpha = 0.05$

Problem to test

H_0 : 30% of drivers make mistake (P = 0.3)

H_1 : 30% of drivers do not make mistake ($P \neq 0.3$) (Two tailed)

Test statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.2616 - 0.3}{\sqrt{\frac{0.7 \times 0.3}{600}}} = \frac{-0.0384}{\sqrt{0.00035}} = -2.13$$

Now, $|Z| = 2.13$

$\text{Prob}(Z \geq |Z_{calculated}|) = \text{Prob}(Z \geq 2.13) = 0.5 - \text{Prob}(0 \leq Z \leq 2.13) = 0.5 - 0.4834 = 0.0166$

For two tailed test, p value $= 2\ \text{Prob}(Z \geq |Z_{\text{calculated}}|)$

$$= 2 \times 0.0166 = 0.0322$$

Here $\alpha = 0.05$

Decision

P value $= 0.0322 < \alpha = 0.05$, reject H_0 at 0.05 level of significance.

Conclusion

30% of all the drivers do not make mistake at the given intersection.

Test of Difference Between Two Proportions

Let P_1 and P_2 be the two population proportions possessing a certain characteristic. Let two independent samples of sizes n_1 and n_2 be drawn from the two population. Also p_1 and p_2 be the proportion of units possessing certain characteristic in the two samples.

For large sample size

$$p_1 \sim N(P_1, \frac{P_1Q_1}{n_1})$$

$$p_2 \sim N(P_2, \frac{P_2Q_2}{n_2}) \text{ then}$$

$$P_1 - P_2 \sim N\left(P_1 - P_2, \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)$$

Different steps in the test are;

Problem to test

H_0: $P_1 = P_2$

H_1: $P_1 \neq P_2$ (Two tailed test) or H_1: $P_1 > P_2$ (One tailed right)

or H_1 : $P_1 < P_2$ (One tailed left)

Test statistic

$$Z = \frac{p_1 - p_2 - E(p_1 - p_2)}{S.E.\ (p_1 - p_2)} = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

$$= \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \text{ if population proportions are given}$$

$$= \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ if population proportions are not given}$$

Where, P_1 = population proportion of first population, P_2 = Population proportion of second population, p_1 = sample proportion of first sample of size n_1, p_2 = sample proportion of second sample of size n_2, $P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$, $Q = 1 - P$

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|Z| > Z_{\text{tabulated}}$, accept otherwise.

Example 14: There are two software developer in incubation officer. First developer faces 21 box while applying 500 logics. Similarly another developer faces 3 box in 100 logics. Are the two developer differ significant in their performance?

Solution:

Here,

No. of bugs by first developer $(x_1) = 21$

No. of logic of first developer $(n_1) = 500$

No. of bugs by second developer $(x_2) = 3$

No. of logic of second developer $(n_2) = 100$

Sample proportion of bugs by first developer $(p_1) = \frac{x_1}{n_1} = \frac{21}{500} = 0.042$

No. of logic of second developer bugs by second developer $(p_2) = \frac{x_2}{n_2} = \frac{3}{100} = 0.03$.

Let P_1 = Population proportion of bugs by first developer (P_2) = Population proportion of bugs by second developer

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

$$= \frac{500 \times 0.042 + 100 \times 0.03}{5000 + 100}$$

$$= \frac{24}{600}$$

$$= 0.04,$$

$Q = 1 - P = 1 - 0.04 = 0.96$

Problem to test

H_0: There is no significant difference in performance of two developer $(P_1 = P_2)$

H_1: There is significant difference in performance of two developer $(P_1 \neq P_2)$ (Two tailed)

Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.042 - 0.03}{\sqrt{0.04 \times 0.96\left(\frac{1}{500}+\frac{1}{100}\right)}} = \frac{0.012}{\sqrt{0.0004608}} = 0.571$$

Now, Prob($Z \geq Z_{calculated}$) = Prob ($Z \geq 0.571$) = 0.5 - Prob ($0 \leq Z \leq 0.571$) = 0.5 - 0.2157 = 0.284

For two tailed test, p value = 2Prob($Z \geq Z_{calculated}$) = 2 × 0.284 = 0.568

Here $\alpha = 1\% = 0.01$

Decision

P value = 0.568 > α = 0.01, accept H_0 at 1 % level of significance.

Conclusion

There is no significant difference in performance of two developer.

Example 15: A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts 3 imperfect articles in a batch of 100. Has the machine improved?

Solution:

Here,

Imperfect articles by machine before overhauled (x_1) = 16, Sample size of articles before machine is overhauled (n_1) = 500, Imperfect articles by machine after overhauled (x_2) = 3, Sample size of articles after machine is overhauled (n_2) = 100, Sample proportion of imperfect articles before overhauled (p_1) = $\frac{x_1}{n_1} = \frac{16}{500}$ = 0.032, Sample proportion of imperfect articles after overhauled (p_2) = $\frac{x_2}{n_2} = \frac{3}{100}$ = 0.03.

Let P_1 and P_2 be the population proportion of imperfect articles before and after overhauled respectively,

$$P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{16 + 3}{500 + 100} = \frac{19}{600} = 0.03166,\ Q = 1\text{-}P = 1\text{-}\ 0.03166 = 0.96833$$

Problem to test

H_0: $P_1 = P_2$

H_1: $P_1 < P_2$ (one tailed left)

Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.032 - 0.03}{\sqrt{0.03166 \times 0.96833\left(\frac{1}{500}+\frac{1}{100}\right)}}$$

$$= \frac{0.002}{\sqrt{0.00036789}}$$

$$= \frac{0.002}{0.019} = 0.105$$

Critical value

Let α = 5% = 0.05 be the level of significance, then critical value for one tailed test is $Z_{\text{tabulated}} = Z_\alpha = 1.645$.

Decision

$Z = 0.105 < Z_{\text{tabulated}} = 1.645$, accept H_0 at 5 % level of significance.

Conclusion

Machine has not improved after overhauled.

Example 16: 1000 apples kept under one type of storage were found to show rotting of the extent of 4%. 1500 apples kept under another kind of storage showed 3% rotting. Can it be reasonably concluded that the second type of storage is superior to the first?

Solution:

Here, sample size of apples in first storage (n_1) = 1000, Sample proportion of rotting apples in first storage (p_1) = 4% = 0.04, Sample size of apples in second storage (n_2) = 1500, Sample proportion of rotting apples in second storage (p_2) = 3% = 0.03

Let P_1 = Population proportion of rotting apples kept in one storage, P_2 = Population proportion of rotting apples kept in second storage,

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{1000 \times 0.04 + 1500 \times 0.03}{100 + 1500} = \frac{85}{2500} = 0.034,$$

$$Q = 1 - P = 1 - 0.034 = 0.966$$

Problem to test

$H_0 : P_1 = P_2$

$H_1 : P_1 > P_2$ (One tailed right)

Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.04 - 0.03}{\sqrt{0.034 \times 0.966\left(\frac{1}{1000} + \frac{1}{1500}\right)}}$$

$$= \frac{0.01}{\sqrt{0.00005474}} = \frac{0.01}{0.0074} = 1.351$$

Critical value

Let α = 5% = 0.05 be the level of significance then critical value for one tailed test is $Z_{\text{tabulated}} = Z_a = 1.645$.

Decision

$Z = 1.351 < Z_{\text{tabulated}} = 1.645$, accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between first storage and second storage.

Example 17: In city A, there are 856 births in a year of which 51% were males. In cities A and B combined, the proportion of male births in a total of 1360 was 47%. Is there any significant difference in the proportion of male births in the two cities?

Solution:

Here,

Sample size in city $A(n_1) = 856$, Sample proportion of male birth in city A $(p_1) = 51\% = 0.51$

Total proportion of male birth $(P) = 47\% = 0.47$

$Q = 1 - P = 0.53,\ n_1 + n_2 = 1360, n_2 = 1360 - 856 = 504$

$$P = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

or $$0.47 = \frac{856 \times 0.51 + 504 \times p^2}{1360}$$

or $$p_2 = 0.40$$

Sample proportion of male birth in city B $(p_2) = 0.4$

Problem to test

H_0 : There is no significant difference in proportion of male birth in city A and city B $(P_1 = P_2)$

H_1 : There is significant difference in proportion of male birth in city A and city B $(P_1 \neq P_2)$

Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.51 - 0.40}{\sqrt{0.47 \times 0.53\left(\frac{1}{856} + \frac{1}{504}\right)}}$$

$$= \frac{0.11}{0.027}$$

$$= 4.07$$

Critical value

Let $\alpha = 5\%$ be the level of significance then the critical value for two tailed test is $Z_{\text{tabulated}} = Z_{\alpha/2} = 1.96$.

Decision

$Z = 4.07 > Z_{\text{tabulated}} = 1.96$, reject H_0 at 5% level of significance.

Conclusion

There is significant difference in the proportions of male births in two cities.

Example 18: A computer manufacturing firm claims that its brand A computer outsells its brand B by 8%. If it is found that 42 out of a sample of 200 users prefer brand A and 18 out of another sample of 100 user prefer brand B, test whether the 8% is a valid claim. Use 5% level of significance.

Solution:

Here

First sample user (n_1) = 200, Brand A prefer from first sample (x_1) = 42

Second sample user (n_2) = 100, Brand B prefer from second sample (x_2) = 18

Sample proportion of brand A prefer user (p_1) $= \frac{x_1}{n_1} = \frac{42}{200} = 0.21$

Sample proportion of brand B prefer user (p_2) $= \frac{x_2}{n_2} = \frac{18}{100} = 0.18$

$$P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{42+18}{200+100} = \frac{60}{300} = 0.2,\ Q = 1 - P = 1 - 0.2 = 0.8$$

Let P_1 = Population proportion of sells of brand A computer, P_2 = Population proportion of sells of brand B computer.

Problem to test

$H_0 : P_1 - P_2 = 0.08$

$H_1 : P_1 - P_2 \neq 0.08$ (Two tailed)

Test statistic

$$Z = \frac{p_1 - p_2 - (P_1 - P_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.21 - 0.18 - 0.08}{\sqrt{0.2 \times 0.8\left(\frac{1}{200} + \frac{1}{100}\right)}} = \frac{-0.05}{\sqrt{0.0024}} = -1.02$$

Critical value

At α = 5% level of significance, critical value for two tailed test is Z tabulated $= Z_{\alpha/2} = 1.96$.

Decision

$|Z| = 1.02 < Z_{\text{tabulated}} = 1.96$, accept H_0 at 5% level of significance.

Conclusion

We can conclude that a difference of 8% in the sale of two brands of computer is a valid claim.

Small Sample Tests

When sample selected from population is less than or equal to 30 is called small sample size. In such cases sampling distribution of statistic is not approximately normally distributed .For small samples the statistic value estimated vary from sample to sample and also far from population parameter .So that the inference drawn from small sample is less precise in comparison to the inference from large sample. Hence modification in the hypothesis testing is made and are called exact sample test or small sample test .Different small sample test are based upon exact sampling distribution i.e. t distribution , F distribution ,Fisher's Z distribution, χ^2 distribution.

t test

When the sample size is less than or equal to 30 then the sampling distribution of the sample mean follows student's t distribution. The t distribution is also similar to normal distribution having shape as in normal distribution but little bit flatter. As the sample size is more than 30 then shape of t distribution is more likely to normal curve.

t test is

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \text{t distribution with n-1 degree of freedom.}$$

where $\bar{X} = \frac{\Sigma X}{n}$, $S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ is the unbiased estimate of the population standard deviation.

t test is based upon the assumption that

- sample size $n \leq 30$
- sample is selected from normal population.
- population standard deviation is not known.
- samples are independent.

It is used to test

- significance of single mean.
- significance of difference between means.
- significance of correlation coefficient.
- significance of regression coefficient.

One Sample Test For Mean Of Normal Population With Unknown Variance

Consider a random sample of size $n(n \leq 30)$ selected from normal population having mean μ and unknown variance. Let $x_1, x_2, x_3, \ldots x_n$ be sample of size n.

It is based upon the assumption that samples are selected from normal population with unknown variance and the samples observations are independent.

Different steps in the test are;

Problem to test

H_0: $\mu = \mu_0$ (sample is drawn from population with mean μ_0)

H_1: $\mu \neq \mu_0$ (Two tailed test) or H_1: $\mu > \mu_0$ (One tailed right) or $H_1 : \mu < \mu_0$ (One tailed left)

Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim \text{t distribution with n-1 degree of freedom.}$$

where, $\bar{x} = \frac{\Sigma x}{n}$, S^2 (sample mean square) $= \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{1}{n-1} = \frac{1}{x-1}[\Sigma x^2 - n\bar{x}]$

$$= \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\}, \text{ where } d = x-A$$

If sample variance $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$ then

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$$

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of t is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|t| > t_{tabulated}$, accept otherwise.

Confidence limit of population mean for small sample size

At α% level of significance for n-1 degree of freedom the critical value of t is $t_{\alpha/2\,(n-1)}$, then (100 – α%) confidence or fudicial limits for population mean μ is given by

$$\bar{x} \pm t_{\alpha/2\,(n-1)} \frac{S}{\sqrt{n}}, \text{ where } S^2 = \frac{\Sigma(X-\bar{x})^2}{n-1} = \frac{1}{n-1}[\Sigma x^2 - n\bar{x}^2]$$

If the sample standard deviation is given then confidence limits for population mean μ is given by $\bar{x} \pm t_{\alpha/2\,(n-1)} \frac{s}{\sqrt{n-1}}$, where $s^2 = \frac{\Sigma(x-\bar{x})^2}{n}$

Example 19: A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm with a standard deviation of 0.002 cm. Test the significance of the deviation of mean.

Solution:

Here

Population mean (μ) = 0.025, Sample size (n) = 10, Sample mean ($\bar{X}$) = 0.024

Sample SD (s) = 0.002

Problem to test

H_0 : Average thickness is 0.025cm ($\mu = 0.025$)

H_1 : Average thickness is not 0.025 ($\mu \neq 0.025$) (Two tailed)

Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{0.024 - 0.025}{\frac{0.002}{\sqrt{10-1}}} = \frac{-0.001 \times 3}{0.002} = -1.5$$

Critical value

Let α =5% be the level of significance then critical value for two tailed test is

$t_{tabulate} = t_{\alpha/2(n-1)} = 2.262.$

Decision

$|t| = 1.5 < t_{tabulated} = 2.262$, accept H_0 at 5% level of significance.

Conclusion

There is no significant deviation between the sample mean and population mean.

Example 20: A university librarian suspects that the average number of books checked out to each student per visit has changed recently. In the past average of 3.4 books were checked out. However, a recently sample of 23 students averaged 4.3 books per visit with a s.d. of 1.5 books. At 0.01 level of significance has the average check out changed? Use confidence limit to draw conclusion.

Solution:

Here, Average books checked out (μ) = 3.4, Sample size (n) = 23 ,

Sample average books checked out ($\bar{x}$) = 4.3, Sample Sd (s) = 1.5

Problem to test

H_0 : Average books checked out is 3.4 ($\mu = 3.4$)

H_1 : Average books checked out is not 3.4 ($\mu \neq 3.4$) (Two tailed)

Critical value

Here

$\alpha = 0.01$ then critical value $t_{\alpha/2\,(n-1)} = 2.819$

Now

$$\text{Confidence limit for } \bar{x} = \mu \pm t_{\alpha/2\,(n-1)} \frac{s}{\sqrt{n-1}}$$

$$= 3.4 \pm 2.819 \times \frac{1.5}{\sqrt{23-1}}$$

$$= 3.4 \pm \frac{4.2285}{4.69} = 3.4 \pm 0.901$$

Taking + sign, $\bar{x} = 3.4 + 0.901 = 4.301$

Taking – sign, $\bar{x} = 3.4 - 0.901 = 2.499$

$\bar{X} = 4.3$ lies between 2.499 and 4.301, accept H_0 at 5% level of significance.

Conclusion

The average number of books checked out to each student per visit has not changed.

Example 21: A random sample download speed of 10 network points of Subisu ISP give the following data in mbps?

70 ,120 ,110 ,101 ,88 , 83 ,95 ,107 ,100 ,98

Do these data support the assumption of population mean of 100 mbps?

Solution:

Here,

Population mean (μ) = 100

Sample size (n) = 10

Download speed (X)	d= X – A (A=90)	d^2
70	-20	400
120	30	900
110	20	400
101	11	121
88	-2	4
83	-7	49
95	5	25
107	17	289
100	10	100
98	8	64
	$\Sigma d = 72$	$\Sigma d^2 = 2352$

Now, $\bar{x} = A + \frac{\Sigma d}{n} = 90 + \frac{72}{10} = 97.2$

$$S^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{10-1}\{2352 - 10 \times 7.2^2\} = \frac{1833.6}{9} = 203.74$$

$S \doteq 14.27$

Problem to test

H_0: Download speed level in population is 100 mg/dl ($\mu = 100$)

H_1: Download speed in population is not 100 mbps ($\mu \neq 100$) (Two tailed)

Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{97.2 - 100}{\frac{14.27}{\sqrt{10}}} = -0.621$$

Critical value

Let $\alpha = 0.05$ be the level of significance then critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2(n-1)} = 2.262.$

Decision

$|t| = 0.621 < t_{tabulated} = 2.262$, accept H_0 at 0.05 level of significance.

Conclusion

The data support the assumption of mean download speed of 100 mbps in the population.

Example 22: The heights of 10 males of a given locality are found to be 70 , 67 ,62 , 68 , 61 , 68, 70 , 64, 64 ,66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level.

Solution:

Here,

Sample size (n) = 10, Population mean (μ) = 64, Level of significance (α) = 5% = 0.05

Height (x)	d = x – A (A = 66)	d^2
70	4	16
67	1	1
62	-4	16
68	2	4
61	-5	25
68	2	4
70	4	16
64	-2	4
64	-2	4
66	0	0
	$\Sigma d = 0$	$\Sigma d^2 = 90$

$$\bar{x} = A + \frac{\Sigma d}{n} = 66 + 0 = 66$$

$$S^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{9}\{90 - 10 \times 0^2\} = 10$$

$S = 3.16$

Problem to test

H_0 : Average height of male is 64 inches ($\mu = 6.4$)

H_1 : Average height of male is more than 64 inches ($\mu > 64$) (One tailed right)

Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{66 - 64}{\frac{3.16}{\sqrt{10}}} = 2$$

Critical value

Tabulated value of t for $\alpha = 5\%$ level of significance for one tailed test is $t_{tabulated} = t_{\alpha(n-1)} = 1.833$.

Decision

$t = 2 > t_{tabulated} = 1.833$, reject H_0 at 5% level of significance.

Conclusion

We conclude that average height is greater than 64 inches.

Example 23: The heights of 10 children selected at random from a given locality had a mean 63.2 cms and variance 6.25 cms. Test at 5% level of significance the hypothesis that the children of the given locality are on the average less than 65 cms in all.

Solution:

Here,

Sample size (n) = 10, Sample mean $(\bar{x}) = 63.2$, Sample variance $(s^2) = 6.25$

$s = \sqrt{6.25} = 2.5$, Level of significance $(\alpha) = 5\%$, Population mean $(\mu) = 65$

Problem to test

H_0: Mean height of children is 65 cms $(\mu = 65)$

H_1 : Mean height of children is less than 65 cms $(\mu < 65)$ (one tailed left)

Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{63.2 - 65}{\frac{2.5}{\sqrt{10-1}}} = \frac{-1.8 \times 3}{2.5} = -2.158$$

Critical value

The tabulated value of t at $\alpha = 0.05$ level of significance for one tailed test is $t_{tabulated} = t_{\alpha(n-1)} = 1.83$

Decision

$|t| = 2.158 > t_{tabulated} = 1.83$, reject H_0 at 5% level of significance.

Conclusion

The mean height of children is less than 65 cms.

Example 24: A random sample of size 25 showed a mean of 65 inches with a standard deviation of 25 inches. Determine 98% confidence intervals for the mean of the population.

Solution:

Here

Sample size (n) = 25, sample mean $(\bar{x}) = 65$, Sample SD (s) = 25

Confidence limit $(1 - \alpha) = 98\% = 0.98$, Level of significance $(\alpha) = 0.02$

$$\text{Confidence limit for } \mu \text{ is } \bar{x} \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n-1}} = 65 \pm 2.492 \times \frac{25}{\sqrt{25-1}}$$

$$= 65 \pm 12.717.$$

Taking – sign

$65 - 12.717 = 52.283$

Taking + sign

$65 + 12.717 = 77.717$

Hence confidence limit is 52.283 inches to 77.717 inches.

Example 25: A city health department wishes to determine the mean bacteria count per volume of water at a lake beach. Researchers have collected 10 water sample of unit volume and have found the bacteria counts to be 175, 190, 215, 198, 184, 207, 210, 193, 196, 180. Obtain the 95% confidence limit for the mean bacteria per count per unit volume of water at the lake beach.

Solution:

Here

Sample size (n) = 10, Confidence limit $(1 - \alpha) = 95\%$, Level of significance $(\alpha) = 0.05$

Number of bacteria(x)	d = x- 196	d^2
175	-21	441
190	-6	36
215	19	361
198	2	4
184	-12	144
207	11	121
210	14	196
193	-3	9
196	0	0
180	-16	256
	$\Sigma d = -12$	$\Sigma d^2 = 1568$

$$\bar{x} = A + \frac{\Sigma d}{n} = 196 + \frac{-12}{10} = 194.8$$

$$S^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{9}\{1568 - 10 \times (-1.2)^2\} = 172.62$$

$$S = \sqrt{172.62} = 13.139$$

Now 95% confidence limit for mean is $\bar{x} \pm t_{\alpha/2(n-1)} \dfrac{S}{\sqrt{n}}$

$$= 194.8 \pm 2.262 \times \frac{13.139}{\sqrt{10}} = 194.8 \pm 9.398$$

Taking - sign

$194.8 - 9.398 = 185.402$

Taking + sign

$194.8 + 9.398 = 204.198$

Hence the 95% confidence limit for the mean bacteria per count per unit volume of water at the lake beach lies between 185.402 to 204.198.

Test for Difference between Two Means (Small Sample)

Let us consider two independent samples of sizes n_1 ($n_1 \leq 30$) and n_2 ($n_2 \leq 30$) be drawn from two normal populations of means μ_1 and μ_2 with unknown variances respectively. Let x_1, x_2, x_3 .... x_{n_1} be sample of size n_1 and y_1, y_2 y_3, ... y_{n_2} be sample of size n_2 respectively.

Different steps in the test are;

Problem to test

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$ (Two tailed test) or H1: $\mu_1 > \mu_2$ (One tailed right) or $H_1 : \mu_1 < \mu_2$ (One tailed left)

Test statistic

$$t = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \text{t distribution with } n_1 + n_2 - 2 \text{ degree of freedom.}$$

$$= \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where, $\bar{x} = \frac{\Sigma x}{n_1}$, $\bar{y} = \frac{\Sigma y}{n_2}$, $S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$

$$= \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2},\ S_1^2 = \frac{\Sigma(x - \bar{x})^2}{n_1 - 1},\ S_2^2 = \frac{\Sigma(y - \bar{y})^2}{n_2 - 1}$$

When sample variances are given then, $S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(X - \bar{X})^2 + \Sigma(Y - \bar{Y})^2]$

$$= \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2},\ s_1^2 = \frac{\Sigma(x - \bar{x})^2}{n_1},\ s_2^2 = \frac{\Sigma(y - \bar{y})^2}{n_2}$$

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of t is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|t| > t_{tabulated}$, accept otherwise.

Example 26: The mean life of a sample of 10 lamps of projector was found to be 1456 hours with a standard deviation of 423 hours. A second sample of 17 lamps of projector chosen from a different batch showed a mean life of 1280 hours with a standard deviation of 398 hours. Is there significant difference between the means of the two batches?

Solution:

Here, First sample size (n_1) = 10, First sample mean ($\bar{x}$) = 1456, First sample SD (s_1) = 423,

Second sample size (n_2) = 17, Second sample mean ($\bar{y}$) = 1280, Second sample SD (s_2) = 398

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10 \times 423^2 + 17 \times 398^2}{10 + 17 - 2}$$

$$= \frac{1789290 + 2692869}{25} = \frac{4482159}{25} = 179286.36$$

Let μ_1 and μ_2 be population mean life of first type of projector lamp and second type of projector lamp respectively.

Problem to test

H_0 : There is no significant difference between mean life of two types of projector lamps ($\mu_1 = \mu_2$)

H_1: There is significant difference between mean life of two types of projector lamps ($\mu_1 \neq \mu_2$) (two tailed)

Test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1456 - 1280}{\sqrt{179286.36\left[\frac{1}{10} + \frac{1}{17}\right]}}$$

$$= \frac{176}{168.52} = 1.04$$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2\,(n_1 + n_2 - 2)} = 2.06$

Decision

$t = 1.04 < t_{tabulated} = 2.06$, accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between mean of two type of projector lamps.

Example 27: A study was conducted to IT students to complete a certain maze. The following data reveals the time taken in second to complete the maze by eight students each of BCA and BSc CSIT. Can't be conclude that B.Sc. CSIT students more efficients that of BCA students. Use 1% level of significant.

D_1	8	12	13	9	3	8	10	9
D_2	10	8	12	15	6	11	12	12

Solution:

Let μ_1 and μ_2 be the average time to complete maze by BCA and BSc. CSIT students.

Here, $n_1 = n_2 = 8$, $\alpha = 1\% = 0.01$

D_1 (x)	D_2 (y)	x^2	y^2
8	10	64	100
12	8	144	64
13	12	169	144
9	15	81	225
3	6	9	36
8	11	64	121
10	12	100	144
9	12	81	144
$\Sigma x = 72$	$\Sigma y=86$	$\Sigma x^2 =712$	$\Sigma y^2 = 978$

$$\bar{x} = \frac{\Sigma X}{n_1} = \frac{72}{8} = 9 \qquad \bar{y} = \frac{\Sigma Y}{n_2} = \frac{86}{8} = 10.75$$

$$S_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1 - 1} = \frac{\Sigma x^2 - n\bar{x}^2}{n_1 - 1} = \frac{1}{7}[712 - 8 \times 9^2] = \frac{64}{7}$$

$$S_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2 - 1} = \frac{\Sigma y^2 - n\bar{y}^2}{n_2 - 1} = \frac{1}{7}[978 - 8 \times 10.75^2] = \frac{53.5}{7}$$

$$S^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = \frac{7 \times \frac{64}{7} + 7 \times \frac{53.5}{7}}{8 + 8 - 2} = \frac{117.5}{14} = 8.392$$

Problem to test

H_0: BCA and BSc CSIT students are equally efficient ($\mu_1 = \mu_2$).

H_1: BSc CSIT students one more efficient than BCA students.

Test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{9 - 10.75}{\sqrt{8.392\left[\frac{1}{8} + \frac{1}{8}\right]}} = \frac{-1.75}{\sqrt{2.098}} = -1.208$$

Critical value

At $\alpha = 0.01$ level of significance for one tailed test the critical value is $t_{tabulated} = t_{\alpha(n_1 + n_2 - 2)} = 2.624$

Decision

$|t| = 1.208 < t_{tabulated} = 2.624$, accept H_0 at 1% level of significance.

Conclusion

BCA and BSc CSIT students are equally efficient.

Example 28: Measuring specimen of nylon yarn taken from two machines, it was found that 8 specimens from the first machine had a mean denier of 9.67 with a standard deviation 1.81 while 10 specimens from the second machine had a mean denier of 7.43 with a standard deviation of 1.48. Assuming that the population sampled are normal and have same variance, test the hypothesis $\mu_1 - \mu_2 = 1.5$ against the alternative hypothesis $\mu_1 - \mu_2 > 1.5$ at 0.05 level of significance.

Solution:

Here, Sample from first machine $(n_1) = 8$, Sample mean from first machine $(\overline{x}) = 9.67$, Sample Sd from first machine $(s_1) = 1.81$, Sample size from second machine $(n_2) = 10$, Sample mean from second machine $(\overline{y}) = 7.43$, Sample Sd from second machine $(s_2) = 1.48$,

Level of significance $(\alpha) = 0.05$

Let μ_1 and μ_2 be the population mean of first machine and second machine respectively.

$$S^2 = \frac{n_1s_1^2 + n_2s_2^2}{n_1 + n_2 - 2} = \frac{8 \times 1.81^2 + 10 \times 1.48^2}{8 + 10 - 2} = \frac{26.2088 + 21.904}{16} = \frac{48.1128}{16} = 3.00705$$

Problem to test

$H_0 : \mu_1 - \mu_2 = 1.5$

$H_1 : \mu_1 - \mu_2 > 1.5$

Test statistic

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(9.67 - 7.43) - 1.5}{\sqrt{3.00705\left[\frac{1}{8} + \frac{1}{10}\right]}} = \frac{0.74}{\sqrt{0.6765}} = 0.902$$

Critical value

At $\alpha = 0.05$ level of significance, the critical value for one tailed test is $t_{tabulated} = t_{\alpha(n_1+n_2-2)} = 1.746$

Decision

$t = 0.902 < t_{tabulated} = 1.746$, accept H_0 at 0.05 level of significance.

Conclusion

$\mu_1 - \mu_2 = 1.5$

Example 29: Samples of two types of electric bulbs were tested for length of life and the following data were obtained:

	Type I	Type II
Number in the sample	8	7
Mean of the sample (in hours)	1134	1024
Standard deviation of sample(in hrs)	35	40

Is mean life of first type of bulbs more than second types of bulbs? Test at 5% level of significance.

Solution:

Here

$n_1 = 8, n_2 = 7, \overline{x} = 1134, \overline{Y} = 1024, s_1 = 35, s_2 = 40, \alpha = 5\% = 0.05$

$$S^2 = \frac{n_1s_1^2 + n_2s_2^2}{n_1 + n_2 - 2} = \frac{8 \times 35^2 + 7 \times 40^2}{8 + 7 - 2} = \frac{9800 + 11200}{13} = 1615.38$$

Let μ_1 and μ_2 be the mean life of first type of electric bulbs and second type of electric bulbs respectively.

Problem to test

H_0 : Mean life of type I bulb and type II bulb is same ($\mu_1 = \mu_2$)

H_1 : Mean life of type I bulb is more than type II bulb ($\mu_1 > \mu_2$)

Test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1134 - 1024}{\sqrt{1615.38\left[\frac{1}{8} + \frac{1}{7}\right]}} = \frac{110}{\sqrt{432.69}} = \frac{110}{20.8} = 5.28$$

Critical value

At $\alpha = 0.05$ level of significance critical value for one tailed test is

$$t_{tabulated} = t_{\alpha(n_1 + n_2 - 2)} = 1.771$$

Decision

$t = 5.28 > t_{tabulated} = 1.771$, reject H_0 at 5% level of significance.

Conclusion

Mean life of first type of bulbs is more than that of second type of bulbs.

Test of Difference between Two Means of Dependent Samples (Paired t test)

Let us consider two dependent (related) samples of sizes $n(n \leq 30)$ be drawn from two normal populations of means μ_1 and μ_2 with unknown variances respectively. Let $x_1, x_2, x_3 \ldots\ldots x_n$ and $y_1, y_2\, y_3, \ldots\ldots y_n$ be samples of size n. Let $d_i = x_i - y_i$ be difference between the observations in the i^{th} sample.

Different steps in the test are;

Problem to test

$H_0 : \mu_d = 0(\mu_1 - \mu_2 = 0)$

H_1: $\mu_d \neq 0$ (Two tailed test) or $H_1 : \mu_d > 0$(One tailed right) or $H_1 : \mu_d < 0$ (One tailed left)

Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} \sim \text{t distribution with n-1 degree of freedom.}$$

Where, $d = X - Y$, $\bar{d} = \frac{\Sigma d}{n}$, $S_d^2 = \frac{\Sigma(d - \bar{d})^2}{n - 1} = \frac{1}{n - 1}\{\Sigma d^2 - n\bar{d}^2\}$

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of t is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|t| > t_{tabulated}$, accept otherwise.

Example 30: A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure:

5 , 2 , 8 , -1 , 3 , 0 , -2 , 1 , 5 , 0 , 4 and 6

Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.

Solution:

Here

$d = X - Y, n = 12$

d	5	2	8	-1	3	0	-2	1	5	0	4	6	$\Sigma d = 31$
d^2	25	4	64	1	9	0	4	1	25	0	16	36	$\Sigma d^2 = 185$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{31}{12} = 2.58$$

$$S_d^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{11}\{185 - 12 \times 2.58^2\} = 9.5382, S_d = \sqrt{9.5382} = 3.08$$

Problem to test

$H_0 : \mu_d = 0 \ (\mu_1 = \mu_2)$

$H_1 : \mu_d < 0 \ (\mu_1 < \mu_2)$ (one tailed)

Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{2.58}{\frac{3.08}{\sqrt{12}}} = \frac{2.58 \times \sqrt{12}}{3.08} = \frac{2.58 \times 3.46}{3.08} = 2.89$$

Critical value

Let $\alpha = 0.05$ be the level of significance, then critical value for one tailed test is

$t_{tabulated} = t_{\alpha(n-1)} = 1.8$

Decision

$t = 2.89 > t_{tabulated} = 1.8$, reject H_0 at 5% level of significance.

Conclusion

The stimulus in general be accompanied by an increase in blood pressure.

Example 31: Ten students were given a test in SPSS. Then they were given a month's training and another test was held. The marks obtained by the 10 students in the two tests are given below:

S N of students	1	2	3	4	5	6	7	8	9	10
Test I	12	15	10	13	18	10	8	17	9	7
Test II	12	17	12	12	14	12	16	16	18	12

Test whether the students have benefited by the training or not.

Solution:

Here

Sample size (n) = 10

Students	Marks in test I(X)	Marks in test II(Y)	d= X-Y	d^2
1	12	12	0	0
2	15	17	-2	4
3	10	12	-2	4
4	13	12	1	1
5	18	14	4	16
6	10	12	-2	4
7	8	16	-8	64
8	17	16	1	1
9	9	18	-9	81
10	7	12	-5	25
			$\Sigma d = -22$	$\Sigma d^2 = 200$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-22}{10} = -2.2$$

$$S_d^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{9}\{200 - 10 \times (-2.2)^2\} = 16.84,\ S_d = \sqrt{16.84} = 4.1$$

Problem to test

H_0 :Students have not benefited by training ($\mu_d = 0$)

H_1 : Students have benefited by training ($\mu_d < 0$) (one tailed)

Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{-2.2}{\frac{4.1}{\sqrt{12}}} = -1.69$$

Critical value

Let $\alpha = 5\%$ be the level of significance, then critical value for one tail test is

$t_{tabulated} = t_{\alpha(n-1)} = 1.83$

Decision

$|t| = 1.69 < t_{tabulated} = 1.83$, accept H_0 at 5% level of significance.

Conclusion

The students have not benefited by the training.

Example 32: A training was done to increase the efficiency of worker for dissembling of computer. The decrease in assembly time of 10 workers after training were as follows:

Solution:

Here

Sample size (n) = 10, Level of significance (α) = 5% = 0.05.

Patient	1	2	3	4	5	6	7	8	9	10	
Decrease in assembly time	6	3	-2	4	-3	4	6	0	0	2	$\Sigma d = 20$
d^2	36	9	4	16	9	16	36	0	0	4	$\Sigma d^2 = 130$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{20}{10} = 2$$

$$S_d^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{9}\{130 - 10 \times 2^2\} = 10, S_d = \sqrt{10}$$

Problem to test

H_0 : The training has no effect on decreasing the assembly time. ($\mu_d = 0$)

H_1 : The training significantly decreases the assembly time. ($\mu_d \neq 0$) (Two tailed)

Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{2}{\frac{\sqrt{10}}{\sqrt{10}}} = 2$$

Critical value

At $\alpha = 0.05$ level of significance, the critical value for two tailed test is $t_{tabulated} = t_{\alpha/2(n-1)} = 2.26$.

Decision

$t = 2 < t_{tabulated} = 2.26$, accept H_0 at 5% level of significance.

Conclusion

The drug has no effect on change of blood pressure.

Example 33: An I.Q. test was administered to 5 persons before and after they were given the nourishing food Horlicks. The results are given below.

Candidates	I	II	III	IV	V
I.Q. before Horlicks	110	120	123	132	125
I.Q. after Horlicks	120	118	125	136	121

Test whether there is any change in I.Q. after the Horlicks at 1% level of significance.

Solution:

Here

Sample size (n) = 5, Level of significance (α) = 1%.

Candidates	I.Q. before(X)	I.Q. after (Y)	d=X-Y	d^2
I	110	120	-10	100
II	120	118	2	4
III	123	125	-2	4
IV	132	136	-4	16
V	125	121	4	16
			$\Sigma d = -10$	$\Sigma d^2 = 140$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-10}{5} = -2$$

$$S_d^2 = \frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\} = \frac{1}{4}\{140 - 5 \times (-2)^2\} = \frac{120}{4} = 30,\ S_d = \sqrt{30}$$

Problem to test

H_0 : There is no significant difference in IQ before and after Horlicks ($\mu_d = 0$)

H_1 : There is significant difference in IQ before and after Horlicks ($\mu_d \neq 0$) (two tailed)

Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{-2}{\frac{\sqrt{30}}{\sqrt{5}}} = \frac{-2}{\sqrt{6}} = \frac{-2}{2.45} = -0.816$$

Critical value

At α = 1% level of significance for two tailed test critical value is $t_{tabulated} = t_{\alpha/2(n-1)} = 4.6$

Decision

$|t| = 0.816 < t_{tabulated} = 4.6$, accept H_0 at 1% level of significance.

Conclusion

There is no significant difference in I.Q. after Horlicks was given.

EXERCISE

1. What do you mean by hypothesis? Describe null hypothesis and alternative hypothesis.
2. Differentiate between acceptance region and rejection region.
3. Define with examples i) Level of significance ii) Degree of freedom iii) Critical value iv) Type II error v) Type I error
4. Explain the procedure for testing significance of mean for large sample.
5. Explain the procedure for testing significance of difference between two means for large samples.
6. Write down different steps in testing of hypothesis.
7. Describe one tailed and two tailed test
8. What is Z test? Write down it's uses.
9. Explain the procedure for testing significance of mean for small sample.
10. Explain the procedure for testing significance of difference between two means for small samples.
11. Explain the procedure for testing significant of difference between two means for small dependent samples.
12. Explain the procedure for testing significance of proportion for large sample.
13. Explain the procedure for testing significance of difference between two proportions for large samples.
14. A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and standard deviation 8 at 5% level of significance. **Ans:** $Z = -2.5$ sig.
15. The mean life time of 400 laptop cells produced by a company is found to be 1570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of the laptop cells produced by the company is 16000 hours against the alternative hypothesis that it is greater than 1600 hours at 1% level of significance. **Ans:** $Z=-4$, sig, $p < 0.01$
16. The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in manufacturing process it is claimed that the breaking strength of the cables have increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0.01 level of significance? **Ans:** $Z = 3.535$, $p < 0.01$, sig.
17. A sample of 400 male students is found to have a mean height of 171.38 cm. Can it be reasonably regarded as a sample from a large population with mean 171.17 cm and standard deviation 3.30 cm. Use confidence limit to draw conclusion. **Ans:** 170.84, 171.49

18. For a sample of 100 women taken from a large population enrolled in a weight reducing program the sample mean diastolic blood pressure is 101 and the s.d. is 42. Can you conclude that on the average the women enrolled in the program have diastolic blood pressure exceed the value of 75 recommended by medical societies? **Ans:** Z= 6.19, sig.

19. Suppose that chest circumference of presumably normal newborn baby girls is normally distributed with mean 13 inch and s.d. 0.7 inch. A group of 49 newborn baby girls from a population group living in a remote region and thought perhaps to constitute a genetic isolate are studied and found to have an average chest circumference of 12.6 inch. In this evidence that the group of 49 came from a population with mean different from 13 inch.

Ans: Z= – 4, sig

20. A new variety of potato grown in 250 plots gave rise to a mean yield of 82. Quintals per hectare with a s.d. of 14.6 quintals per hectare. Is it reasonable to assert that the new variety is superior in yield to the standard variety with an estimated yield of 80.2 quintals per hectare? **Ans:** Z=1.94, sig.

21. In a certain factory there are two independent processes for manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gms with a standard deviation of 12 gms, while the other process are 124 and 14 in a sample of 400 items. Is the difference between the mean weights significant at 10% level of significance. Use p value method. **Ans:** Z=-3.874, p = 0.0001 sig.

22. The nicotine content of two brands of cigarettes has to be compared. A sample of 50 cigarettes of brand A has a mean nicotine content 20.5 with a s.d. of 2.5 . A sample of 50 cigarettes of brand B has a mean of 17.5 with a s.d. of 2.1. Is there any reason to believe that two brands are different so far as the nicotine content is considered? Use α =5%
Ans: Z= 6.49 sig.

23. A random sample of 300 families showed that the average birth rate is 30. Another sample of 400 families showed that the average birth rate is 28. Could the samples reasonably be regarded as the samples drawn from the same population with standard deviation 4? **Ans:** Z= 6.54, sig.

24. A random sample of size 35 taken from a normal population with standard deviation 5.2 has a mean 81. A second sample of size 36 taken from other normal population with a standard deviation 3.4 has a mean 76. Test whether the two sample means do not differ significantly at 5% level of significance. **Ans:** Z=4.78 sig.

25. Two research laboratories have independently produced drugs that provide relief to arthritis suffers. The first drug was tested on a group of 90 arthritis victims and produced an average of 8.5 hours relief with a standard deviation of 1.8 hours. The second drug was tested on 80 arthritis victims producing an average of 7.9 hours of relief with a standard deviation of 2.1 hours. At 5% level of significance does the second drug provide a significantly shorter period of relief? **Ans:** Z=1.98, sig

26. Two random samples of Nepalese people taken from rural and urban region gave the following data of income:

Sample	Size	Average monthly income	s.d.
I from rural region	150	800	50
II from urban region	100	1250	30

Test whether the average monthly income of rural people is significantly less than that of the urban people. **Ans:** $Z = -88.82$ sig.

27. An investigation of two kinds of photocopying equipment showed that 71 failures of first kind of equipment took on average 83.2 minutes to repair with a standard deviation of 19.3 minutes, while 75 failures of the second kind of equipment took on average 90.8 minutes to repair with a standard deviation of 21.4 minutes. Test the hypothesis that on the average it takes an equal amount of time to repair either kind of equipment.

Ans: $Z = -2.25$ sig.

28. A dice was thrown 9000 times and of these 3220 yielded a 3 or 4. Is this consistent with the hypothesis that the dice is unbiased? **Ans:** $Z = 4.97$ sig.

29. A sample of size 600 persons selected at random from a large city shows that the percentage of male in the sample is 53%. It is believed that male to the total population ratio in the city is ½. Test whether this belief is confirmed by the observation.

Ans: $Z = 1.47$ insig.

30. The coordinator in a college claimed that at least 98% of the students submit their assignment on time. Taking the sample of 250 students, 15 were not submitting assignment in whole semester. Test his claim at 10% level of significance.

Ans: $Z=-4.5$ sig.

31. To check on an ambulance service's claim that at least 40% of its calls are life threatening emergencies, a random sample was taken from its files and it was found that only 49 out of 150 calls were life threatening emergencies. Test the claim at 1% level of significance using p value method. **Ans:** $p = 0.033$ insig.

32. It is claimed that both tea and coffee are equally popular in Kathmandu district. If in random sample of 600 persons, 310 were regular consumer of tea. Is the claim justified at 1% level of significance? **Ans:** $z=0.78$ insig.

33. A certain process produces 10% defective articles. A supplier of new raw material claims that the use of his material would reduce the proportion of defectives. A random sample of 400 units using this new material was taken out of which 34 were defective units. Can the supplier's claim be accepted? Test at 1% level of significance. **Ans:** $z = -1$ insig.

34. A machine produces 20 defective articles in a batch of 400. After overhauling it produces 10 defectives in a batch of 300. Has the machine improved. **Ans:** $z=1.1$ insig.

35. At a certain date in a large city 400 out of a random sample of 500 men were found to be smokers. After the tax on tobacco had been heavily increased, another random sample of 600 men in the same city included 400 smokers. Was the observed decrease in the proportion of smokers significant? Test at 5% level of significance. **Ans:** $z = 4.98$ sig.

36. In a random samples of 600 and 1000 men from two cities 400 and 600 men are found to be literate. Do the data indicate that the population are significantly different in the percentage of literacy? Ans: $z=2.4$ sig.

37. Random samples of 250 bolts manufactured by machine A and 200 bolts manufactured by machine B showed 24 and 10 defective bolts respectively .Do the machines showing same quality of performance? Use 5% level of significance. **Ans:** $z=1.84$ insig.

38. In a sample of 600 students of a certain college 400 are found to use dot pens. In another college from a sample of 900students 450 were found to use dot pens. Test whether the two colleges are significantly different with respect to the habit of using the dot pens at 1% level of significance? Use p value method. **Ans:** $p=0$ sig.

39. Rainfall records of a particular place for last 12 years for the month of July showed that average rainfall was 50 mm and standard deviation of 30 mm . Do you agree that the average rainfall at the place was less than 512 mm? Use 10% level of significance.

Ans: $t = 1.36$ sig.

40. Ten patients are selected at random from a population of patients and their blood pressure recorded are as follows;

125 ,147 ,118 , 145 ,140 , 128 , 155 , 150 , 160 ,149.

Do the data support the hypothesis that the population average blood pressure of patients is 135?Use 5% level of significance. Ans: ($t = 1.53$) insig.

41. A fertilizer mixing machine is set to give 12 kg of nitrate for every quintal bag of fertilizer. Ten 100 kg bags are examined . The percentage of nitrate are as follows.

11, 14, 13, 12, 13, 12, 13, 14, 11, 12

Is there reason to believe that the machine is defective? Use confidence limit to draw conclusion. **Ans:** (11.72, 13.27)

42. A random sample of size 16 has the sample mean 53. The sum of the square of deviation taken from the mean value is 150. Can this sample be regarded as taken from the population having 56 as its mean at 99% confidence limit? **Ans:** $t = -3.79$ sig.

43. In the past a machine has produced washers having a mean thickness of 0.05 cm. To determine whether the machine is in proper order , a sample of 10 washers is taken of which mean thickness is 0.053 cm and s.d. is 0.003. Test the hypothesis that the machine is working in proper order. **Ans:** Ans: $t = 3$, sig.

44. The time (in minutes) spent by 10 randomly selected customers using internet in a cyber cafe are as follows;

35, 20, 30, 45, 60, 40, 65, 40, 25, 50

Can you say average time spent by customers is more than 30 minutes at 5% level of significance? **Ans:** $t = 2.4$, sig.

45. A random sample of 10 bulbs has the following life in months; 24, 26, 32, 28, 20, 20, 23, 34, 30 and 43.Obtain the 95% confidence limit for the population mean life of bulbs.

Ans: 22.9, 33.07

46. A random sample of size 10 showed a mean life of 28 years with the standard deviation of 7.1 years. Determine the fiducial limits for population mean. **Ans:** 22.64, 33.35

47 The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance.

Ans: t = – 7.6, sig.

48. Two kinds of manure were applied to sixteen one hectare plot, other condition remaining the same. The yields in quintals are given below:

Manure I	18	20	36	50	49	36	34	49	41
Manure II	29	28	26	35	30	44	46		

Is there any significant difference between the mean yields? Use 5% level of significance.

Ans: t = 0.57, insig.

49. To test the effect of a fertilizer on rice production, 24 plots of land having equal areas were chosen. Half of these plots were treated with fertilizer and the other half were untreated. Other condition were the same. The mean yield of rice on untreated plots was 4.8 quintals with a standard deviation of 0.4 quintal, while the mean yield on the treated plots was 5.1 quintals with a standard deviation of 0.36 quintal. Can we conclude that there is significant improvement in rice production because of fertilizer at 5% level of significance. **Ans:** t = – 1.86 sig.

50. Two new drugs A and B are given to two independent groups of 10 and 12 patients with heart disease respectively. The reduction of blood pressure due to the two new drugs A and B are given below;

Drug A	7	16	14	9	10	11	6	8	10	9		
Drug B	10	12	16	14	11	12	13	8	12	15	9	12

Would you conclude that the drug A is less effective than the drug B in reducing the blood pressure of patients with heart disease at 5% level of significance? **Ans:** t = – 1.75, sig

51. The means of two random samples of size 9 and 7 are 196.42 and 198.82 respectively. The sum of square of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population?

Ans: t = – 2.63, sig.

52. A group of five patients treated with medicine A weigh 42,39,48,60 and 41 kgs. A second group of 7 patients from the same hospital treated with medicine B weigh 38,42,56,64,68,69 and 62 kgs. Do you agree with the claim the medicine B increase the weight significantly?

Ans: t = – 1.7, insig.

53. A drug was given to 10 patients. The increment in their blood pressure were recorded to be 8, 10,-2 ,0, 5, -1 ,9 ,12 ,6 and 5. Is it reasonable to believe that the drug has no increase on change of blood pressure?
Ans: $t = 3.41$ sig

54. Following data gave the yields of rice in 10 experimental plots in two successive years;

Serial no of plots	1	2	3	4	5	6	7	8	9	10
Yield in 1st year	23	20	19	21	18	20	18	22	16	18
Yield in 2nd year	24	19	22	18	20	22	20	20	18	17

Test whether the mean of difference between the yields of two successive years is zero or not at 5% level of significance.
Ans: $t = -0.76$ insig.

55. Ten students were given intensive training on python for a month. The scores obtained in tests I and V given below.

S No of students	1	2	3	4	5	6	7	8	9	10
Marks in 1st test	50	52	53	60	65	67	48	69	72	80
Marks in 5th test	65	55	65	65	60	67	49	82	74	86

Does the score from test I to test V show an improvement? Test at 5% level of significance.
Ans: $t = -2.57$ sig.

56. Memory capacity of 10 students was tested before and after training, state whether the training was effective or not from the following scores;

Roll No.	1	2	3	4	5	6	7	8	9	10
Before training	12	14	11	8	7	10	3	0	5	6
After training	15	16	10	7	5	12	10	2	3	8

Ans: -1.36 insig.

USE OF SOFTWARE

Test of Proportion Two Samples

Test if sample 1 has lower proportion males than in sample 2 at 5 % level of significance and find 95% confidence limits of difference of proportion.

Using excel

	A	B	C	D	E
1		Sample 1		Sample 2	
2	sample size	n1	50	n2	30
3	no. of males	x1	22	x2	15
4	proportion	p1	0.44	p2	0.5
5		q1	0.56	q2	0.5

6			
7	H0:	p1=p2	
8	H1:	p1<p2	
9			
10		value	formula
11	S.E.	0.115157863	=SQRT((C4*C5/C2)+(E4*E5/E2))
12	z cal	-0.521023911	=(C4-E4)/B11
13	alpha	0.05	
14	z tab	-1.644853627	=NORMSINV(B13)
15	p value	0.301175057	=NORMSDIST(B12)
16			

	A	B	C
17	decision:		
18	singnificant approach method		
19	H0 is accepted		=IF(ABS(B12)>ABS(B14)," H0 is rejected,
20			i.e., H1 is accepted", "H0 is accepted")
21			
22	p value approach method		
23	It is insignificant		=IF(B15<B13,"It is significant","It is insignificant")
24	z value	1.959963985	=NORMSINV(1-B13/2)
25			
26	limit of difference of proportion		
27	lower limit	-0.285705263	=(C4-E4)-B24*B11
28	upper limit	0.165705263	=(C4-E4)+B24*B11

Test of Single mean (Z-test)

Test the equality of mean with given value of population mean and determination of confidence interval.

Using EXCEL

	A	B	C
1		symbol	value
2	population mean	μ	50
3	population standard deviation	σ	12
4	sample size	n	150
5	sample mean	m	47.5
6			

7			
8	H0:	μ=50	
9	H1:	μ≠50	formula
10	S.E.	0.980	=C3/SQRT(C4)
11	z cal	-2.552	=(C5-C2)/B10
12	alpha	0.05	
13	z tab	1.960	=NORMSINV(1-B12/2)
14	p value	0.011	=2*(NORMSDIST(B11))

	A	B	C	D	E
16	decision:				
17	significant approach				
18	H0 is rejected, i.e., H1 is accepted		=IF(ABS(B11)>ABS(B13)," H0 is rejected, i.e., H1 is accepted", "H0 is accepted")		
19					
20	p value approach				
21	It is significant		=IF(B14<B12,"It is significant","It is insignificant")		
22					
23	z value	1.960	=NORMSINV(1-B12/2)		
24					
25	limit of difference of proportion				
26	lower limit	45.580	=C5-B23*B10		
27	upper limit	49.420	=C5+B23*B10		
28					

Test of difference of two sample mean

Test whether two sample means are significantly different if they are selected prom population with standard deviations 840 and 920 respectively with following score values.

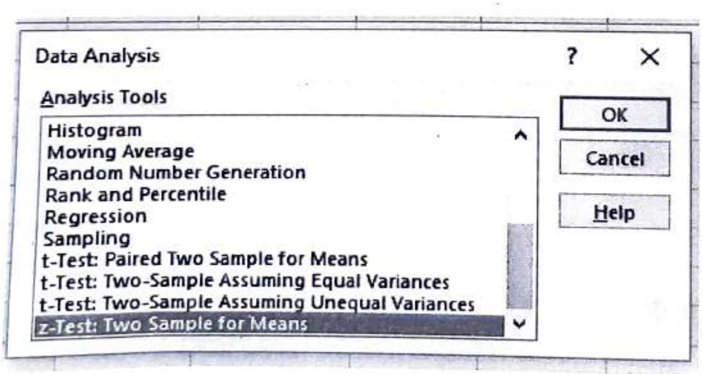
x	y	x	y
52	22	6	14
27	34	20	1
2	8	16	65
49	83	91	95
76	68	88	18
72	64	13	27
77	115	5	62
71	57	59	75
95	32	69	8
55	99	19	84
68	98	73	704
29	39	39	63
97	38	86	9
17	96	7	6
98	92	85	11

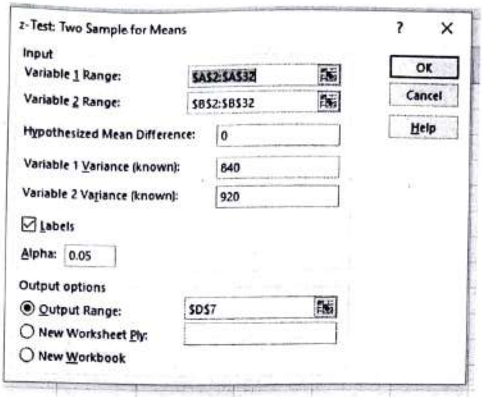
Using Excel

	A	B	C	D	E
34		x		y	
35	sample size	n1	30	n2	30
36	sample mean	m1	52.0333	m2	72.9
37	population variance	σ_1^2	840	σ_2^2	920
38					

Using data analysis tool

Data\Data analysis\Z test:For two sample means





Output

38			
39	Ztest using data analysis tool		
40	z-Test: Two Sample for Means		
41			
42		*x*	*y*
43	Mean	52.0333	72.9
44	Known Variance	840	920
45	Observations	30	30
46	Hypothesized Mean Difference	0	
47	z	-2.7243	
48	P(Z<=z) one-tail	0.00322	
49	z Critical one-tail	1.64485	
50	P(Z<=z) two-tail	0.00644	
51	z Critical two-tail	1.95996	

Using EXCEL Formula:

	A	B	C	D	E	F	G
53	Z test using regular test procedure						
54	H0:	m1=m2					
55	H1:	m1≠m2	formula				
56	S.E.	7.65942	=SQRT((C37/C35)+(E37/E35))				
57	z cal	-2.7243	=(C36-E36)/B56				
58	alpha	0.05					
59	z tab	1.9600	=NORMSINV(1-B58/2)				
60	p value	0.0064	=2*(NORMSDIST(B57))				
61							
62	decision:						
63	significant approach						
64	H0 is rejected, i.e., H1 is accepted		=IF(ABS(B57)>ABS(B59)," H0 is rejected, i.e., H1 is accepted", "H0 is accepted")				
65							
66	p value approach						
67	It is significant		=IF(B60<B58,"It is significant","It is insignificant")				
68							

Hypothesis test for Single sample mean

The time (in minutes) spent by 10 randomly selected customers using internet in a cyber cafe is as follows;35 ,20 ,30 ,45 ,60 ,40 ,65 ,40 ,25 ,50.Can you say that the average time spent by customers is more than 30 minutes at 5% level of significance?

Using EXCEL

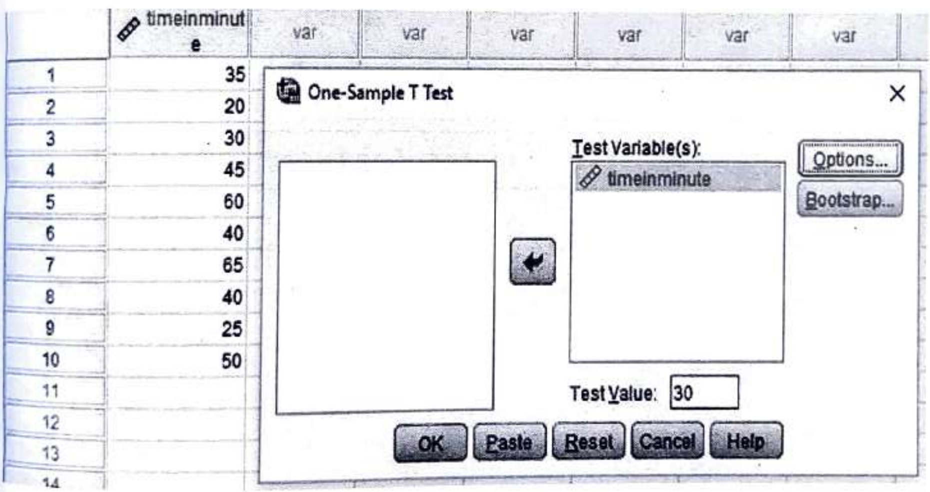
	A	B
6	time in minute	
7	35	
8	20	
9	30	
10	45	
11	60	
12	40	
13	65	
14	40	
15	25	
16	50	

	A	B	C	D	E	F
18	test hypothesis					
19	null hypothesis	$H_0: \mu = \mu_0$				
20	altenative hypothesis	$H_1: \mu > \mu_0$				
21						
22		symbol	value	formula		
23	population mean	μ	30			
24	sample size	n	10	=COUNT(A7:A16)		
25	sample mean	$\bar{y}$	41	=AVERAGE(A7:A16)		
26	sample st dev	s	14.49138	=STDEV.S(A7:A16)		
27	level of significance	α	0.05			
28	degrees of freedom	df	9	=C24-1		
29	tabulated t (two tailed)	$t_{\alpha/2,\, n-1}$	2.262157	=T.INV.2T(C27,C28)		
30	tabulated t (one tailed)	$t_{\alpha,\, n-1}$	1.833113	=T.INV(1-C27,C28)		
31						
32	standard error	S.E.	4.582576	=C26/SQRT(C24)		
33	calculated t	t cal	2.400397	=(C25-C23)/(C26/SQRT(C24))		
34						
35	p value	p	0.039872	=T.DIST.2T(C33,C28)		
36		p rt tailed	0.019936	=T.DIST.RT(C33,C28)		

	A	B	C	D	E	F	G	H
38	Decision							
39	significant approach							
40	No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.							
41	=IF(C33<C30, "Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between sample and population mean.", "No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.")							
42	p- value approach							
43	It is significant,No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.							
44	=IF(C36>C27, "It is insignificant, Hence it can be concluded that there is no significant difference between sample and population mean.", "It is significant,No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.")							
45								

Using SPSS

Analyze\Compare means\one sample t test



➔ **T-Test**

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
time in minute	10	41.00	14.491	4.583

[DataSet3]

One-Sample Test

	Test Value = 30					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
time in minute	2.400	9	.040	11.000	.63	21.37

Using STATA

ttest timeinminute == 30

```
. ttest timeinminute == 30

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
timein~e |      10          41    4.582576    14.49138    30.63349    51.36651
------------------------------------------------------------------------------
    mean = mean(timeinminute)                                     t =   2.4004
Ho: mean = 30                                    degrees of freedom =        9

    Ha: mean < 30               Ha: mean != 30                 Ha: mean > 30
 Pr(T < t) = 0.9801         Pr(|T| > |t|) = 0.0399          Pr(T > t) = 0.0199
```

Two sample mean

Two kinds of manure were applied to sixteen one-hectare plot, other condition remaining the same. The yields in quintals are given below:

Manure I	18	20	36	50	49	36	34	49	41
Manure II	29	28	26	35	30	44	46		

Is there any significant difference between the mean yields? Use 5% level of significance.

Using excel

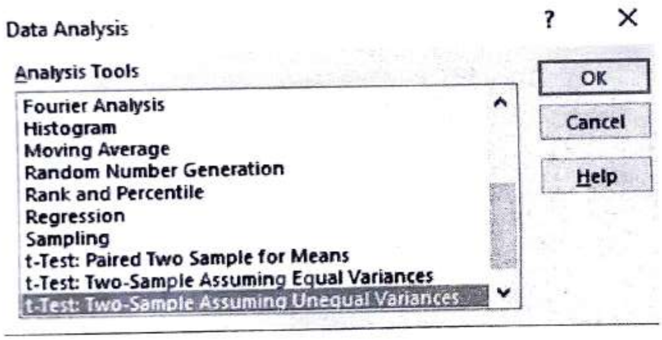
	A	B
7	Manure I	Manure II
8	18	29
9	20	28
10	36	26
11	50	35
12	49	30
13	36	44
14	34	46
15	49	
16	41	

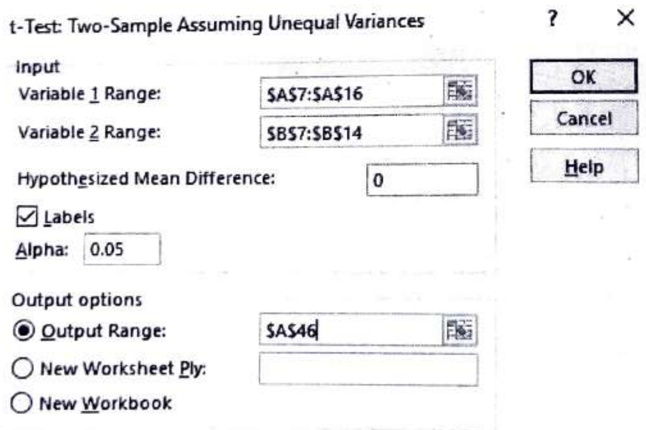
	A	B	C	D	E
19	test hypothesis				
20	null hypothesis	H_0: $\mu_1 = \mu_2$			
21	altenative hypothesis	H_1: $\mu_1 > \mu_2$			
22					
23		symbol	value	formula	
24	sample size	n1	9	=COUNT(A8:A16)	
25		n2	7	=COUNT(B8:B16)	
26	sample mean	ȳ1	37	=AVERAGE(A8:A16)	
27		ȳ2	34	=AVERAGE(B8:B14)	
28	sample st dev	s1	11.9058809	=STDEV.S(A8:A16)	
29		s2	8.020806277	=STDEV.S(B8:B14)	
30	level of significance	α	0.05		
31	degrees of freedom	df	14	=C24+C25-2	
32	tabulated t (two tailed)	$t_{\alpha/2,\, n-1}$	2.144786688	=T.INV.2T(C27,C28)	
33	tabulated t (one tailed)	$t_{\alpha,\, n-1}$	1.761310136	=T.INV(1-C27,C28)	
34	standard error	S.E.	4.994044072	=SQRT(((C28^2)/C24)+((C29^2)/C25))	
35	calculated t	t cal	0.600715564	=(C26-C29)/C35	
36	p value	p	0.557630091	=T.DIST.2T(C33,C28)	
37		p rt tailed	0.278815045	=T.DIST.RT(C33,C28)	

	A
39	Decision
40	significant approach
41	Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between two sample means.
42	=IF(C36<C33, "Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between two sample means.", "No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between two sample means.")
43	p- value approach
44	It is insignificant, Hence it can be concluded that there is no significant difference between two sample means.
45	=IF(C39>C30, "It is insignificant, Hence it can be concluded that there is no significant difference between two sample means.", "It is significant,No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between two sample means.")

Using Data analysis tool

Data\Data Analysis





Click O.K.

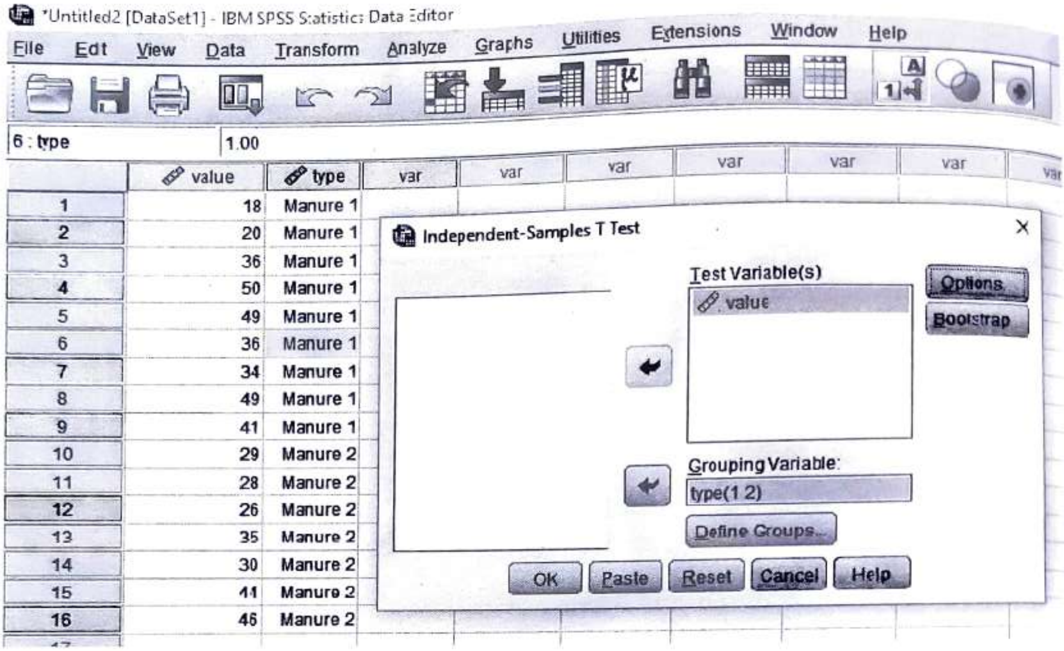
	A	B	C
46	t-Test: Two-Sample Assuming Unequal Variances		
47			
48		*Manure I*	*Manure II*
49	Mean	37	34
50	Variance	141.75	64.3333333
51	Observations	9	7
52	Hypothesized Mean Difference	0	
53	df	14	
54	t Stat	0.60072	
55	P(T<=t) one-tail	0.27882	
56	t Critical one-tail	1.76131	
57	P(T<=t) two-tail	0.55763	
58	t Critical two-tail	2.14479	

	A	B	C	D
60	Decision			
61	significant approach			
62	Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between two sample means.			
63	=IF(B54<B56, "Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between two sample means.", "No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between two sample means.")			
64	p- value approach			
65	It is insignificant, Hence it can be concluded that there is no significant difference between two sample means.			
66	=IF(B55>0.05, "It is insignificant, Hence it can be concluded that there is no significant difference between two sample means.", "It is significant,No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between two sample means.")			

Using SPSS

Create variable for Manure type and value

Analyze\Compare means\Independent sample t test



Click OK

T-Test

[DataSet1]

Group Statistics

	Manure type	N	Mean	Std. Deviation	Std. Error Mean
Manure I	Manure 1	9	37.00	11.906	3.969
	Manure 2	7	34.00	8.021	3.032

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Manure I	Equal variances assumed	.756	.399	.571	14	.577	3.000	5.251	-8.262	14.262
	Equal variances not assumed			.601	13.797	.558	3.000	4.994	-7.726	13.726

Using STATA

. ttest value, by(type) unequal

```
. ttest value, by(type) unequal

Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Manure 1 |       9          37    3.968627    11.90588    27.84833    46.15167
Manure 2 |       7          34     3.03158    8.020806    26.58199    41.41801
---------+--------------------------------------------------------------------
combined |      16     35.6875    2.545779    10.18312     30.2613     41.1137
---------+--------------------------------------------------------------------
    diff |                   3    4.994044               -7.725987    13.72599
------------------------------------------------------------------------------
    diff = mean(Manure 1) - mean(Manure 2)                       t =   0.6007
Ho: diff = 0                     Satterthwaite's degrees of freedom =  13.7967

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7211         Pr(|T| > |t|) = 0.5578          Pr(T > t) = 0.2789
```

If two separate variables are created for manure 1 and Manure 2

. ttest ManureI == ManureII, unpaired unequal

```
. ttest ManureI == ManureII, unpaired unequal

Two-sample t test with unequal variances

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 ManureI |       9          37    3.968627    11.90588    27.84833    46.15167
ManureII |       7          34     3.03158    8.020806    26.58199    41.41801
---------+--------------------------------------------------------------------
combined |      16     35.6875    2.545779    10.18312     30.2613     41.1137
---------+--------------------------------------------------------------------
    diff |                   3    4.994044               -7.725987    13.72599
------------------------------------------------------------------------------
    diff = mean(ManureI) - mean(ManureII)                        t =   0.6007
Ho: diff = 0                     Satterthwaite's degrees of freedom =  13.7967

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7211         Pr(|T| > |t|) = 0.5578          Pr(T > t) = 0.2789
```

Paired t test

Memory capacity of 10 students was tested before and after training, state whether the training was effective or not from the following scores;

Roll No.	1	2	3	4	5	6	7	8	9	10
Before training	12	14	11	8	7	10	3	0	5	6
After training	15	16	10	7	5	12	10	2	3	8

Using EXCEL

	A	B	C	D
7	Roll No.	Before training	After training	d
8	1	12	15	-3
9	2	14	16	-2
10	3	11	10	1
11	4	8	7	1
12	5	7	5	2
13	6	10	12	-2
14	7	3	10	-7
15	8	0	2	-2
16	9	5	3	2
17	10	6	8	-2

	A	B	C
20	test hypothesis		
21	null hypothesis	$H_0: \mu_1 = \mu_2$	
22	altenative hypothesis	$H_1: \mu_1 < \mu_2$	
23			
24	Paired t test		
25		symbol	value
26	population mean	μ	0
27	sample size	n	10
28	sample mean	$\bar{y}$	-1.2
29	sample st dev	s	2.78089
30	level of significance	α	0.05
31	degrees of freedom	df	9
32	tabulated t (two tailed)	$t_{\alpha/2, n-1}$	2.26216
33	tabulated t (one tailed)	$t_{\alpha, n-1}$	1.83311
34	standard error	S.E.	0.87939
35	calculated t	t cal	-1.36458
36	p value	p	0.20553
37		p rt tailed	0.10276

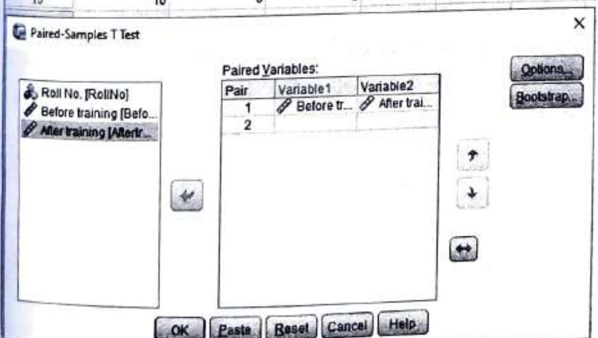
	A	B	C	D	E	F	G	H
39	Decision							
40	significant approach							
41	No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.							
42	**=IF(C36>C33, "Null hypothesis H0 is accepted, Hence it can be concluded that there is no significant difference between sample and population mean.", "No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.")**							
43	p- value approach							
44	It is insignificant, Hence it can be concluded that there is no significant difference between sample and population mean.							
45	**=IF(C39>C30, "It is insignificant, Hence it can be concluded that there is no significant difference between sample and population mean.", "It is significant,No reason to accept Null hypothesis H0 , Hence it can be concluded that there is significant difference between sample and population mean.")**							

Using SPSS

Paired t test SPSS

Analyze\Compare means\paired t test

	RollNo	Beforetraining	Aftertraining	var	var	var
1	1	12	15			
2	2	14	16			
3	3	11	10			
4	4	8	7			
5	5	7	5			
6	6	10	12			
7	7	3	10			
8	8	0	2			
9	9	5	3			
10	10	6	8			



Click OK.

Result will be as follows.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Before training	7.60	10	4.300	1.360
	After training	8.80	10	4.733	1.497

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Before training & After training	10	.815	.004

Paired Samples Test

		Paired Differences					t	df	Sig. (2-t
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Before training - After training	-1.200	2.781	.879	-3.189	.789	-1.365	9	

Using STATA

. ttest Beforetraining == Aftertraining

```
. ttest Beforetraining == Aftertraining

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Before~g |      10         7.6    1.359739    4.299871    4.524058    10.67594
Aftert~g |      10         8.8    1.496663    4.732864    5.414313    12.18569
---------+--------------------------------------------------------------------
    diff |      10        -1.2    .8793937    2.780887   -3.189327    .7893268
------------------------------------------------------------------------------
     mean(diff) = mean(Beforetraining - Aftertraining)            t =  -1.3646
 Ho: mean(diff) = 0                              degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.1028         Pr(|T| > |t|) = 0.2055          Pr(T > t) = 0.8972
```

3
Chapter

NON PARAMETRIC TEST

CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Parametric Vs. non-parametric test, Needs of applying non-parametric tests,
- One-sample test, Run test, Binomial test, Kolmogorov-Smirnov test
- Two independent sample test, Median test, Kolmogrov-Smirnov test, Wilconxon Mann Whitney test, Chi-square test
- Paired-sample test: Wilcoxon signed rank test, Cochran's Q test, Fredman two way analysof test variance test, Kruskal Wallis test
- Problems and illustrative examples related using software.

Most of the hypothesis testing procedures so far such as Z test, t test, F test are based upon the assumption that the random samples are selected from a normal population. If this is true, these methods can extract all the information that is available in sample and they usually give the best possible precision. Parametric tests depends on parameters, viz., mean or proportion or standard deviation of the population from which sample is taken.

In practice there are many circumstances in which sample are selected from non normal population. In such case, we can have no assumptions about parameters or normality about the population. In such special cases, Parametric are inevitable and are used for testing of hypothesis. Manynon parametric procedures are based upon ranked data or even categorical data. For such data no parametric tests are available.

Non parametric tests are often used in place of their parametric counterparts when certain assumptions about the underlying population are in weak state. However, if the sample size is large enough most non parametric tests can be viewed as (the usual normal theory based procedures applied to ranks- make this statement more clear).

The following table shows a comparative view of some normal distribution based tests (that is parametric tests) and its non parametric counterparts

Normal Distribution based Test	Corresponding non Parametric Test	Purpose of Test
t test for independent samples	Mann Whitney U test	Compare two independent samples
Paired t test	Wilcoxon matched pair signed rank test	Compare dependent samples
One way ANOVA	Kruskalwallis H test	Compare three or more groups
Two way ANOVA	Friedman test	Compare groups classified by two factors

Advantages of Non-parametric Test

- Can be applied to qualitative data (rank, ordinal, categorical data) as well as quantitative data.
- For relatively small sample it is the only possible test
- It is simple to understand, quicker an easier to apply.
- It is less time consuming.
- It needs no assumption about the population from which sample is selected.
- It has greater range of applicability because of milder assumption.
- It does not require complicated sampling theory.

Disadvantages of Non-parametric Test

- It can not be used to estimate the parameters of population.
- These tests are less reliable and less powerful than parametric tests.
- These tests are less efficient than parametric tests.
- In these test many times lot of information are discarded or unused.
- Lot of tables are needed for tests.

Uses of non parametric test

(i) When data size is small
(ii) When assumptions of parametric procedure are not satisfied.
(iii) When quick or preliminary data analysis is required.
(iv) When data are in weak scaled.
(v) When data show highly skewed nature (large positive/negative skewness)
(vi) When basic question of interest is distribution free in nature.

Measurement Scale

The measurement is the process of assigning number codes to observations with or without true numerical meaning. Measurement scales are used to categorize and/or quantify variables. Any statistical data must have measurement. There are four measurement scale in which any data is measured.

(i) Nominal scale
(ii) Ordinal scale
(iii) Interval scale
(iv) Ratio scale

Nominal Scale

It is a process of assigning numbers or symbols to events such that the numbers assigned have no numerical meaning. It is the weakest scale. It is numeric in name only because it does not satisfy ordinary arithmetic properties such as addition, subtraction, multiplication, division etc. eg. Numbers on football player's jerseys, numbers assigned to sex, numbers assigned to marital status.

Ordinal Scale

It is system of assigning numbers to events such that numbers assigned have grading to events. It is rank order of the events. In comparing one is higher or more than the other ordinal scale is used. e.g. grading system in government job , preference to different cold drinks such as Coca-Cola, Sprite, Fanta, Pepsi.

Interval Scale

It is extended form of ordinal scale in which distance (interval) between two objects are exactly known. In this scale the numerals with quantitative meaning are associated to the objects. In this scale ratio of any two interval is independent of unit of measurements and zero point(origin).eg. temperature recorded.

Ratio Scale

It is extended form of interval scale including a true zero point as it's origin. In this case ratio of any two points is independent of unit of measurement but not of zero point. The numbers associated to ratio scale are true numbers with true zero. eg. heights recorded, weights recorded.

Difference between Parametric and Non Parametric test

Parametric test	Non parametric test
1. It specifies certain condition about parameter of the population from which sample is selected.	1. It does not specify certain condition about parameter of the population from which sample is selected.
2. It is used in testing of hypothesis and estimation of parameters.	2. It is used in testing of hypothesis but not in estimation of parameters.
3. Mostly it is used in data measured in interval and ratio scale.	3. It is used in data measured in nominal and ordinal scale.
4. It is most powerful.	4. It is less powerful.
5. It requires complicated sampling technique.	5. It does not require complicated sampling technique.

Assumptions of Non-parametric Test

Non parametric tests do not depend upon assumption about parameter of the population from which sample is selected. However, following basic and weak assumptions are made in the non parametric tests.

(i) The sample observations are independent.

(ii) The variable under study is continuous.

(iii) Samplep.d.f. is continuous.

(iv) Lower order moments exists(mean and variance).

One Sample Test

Run Test

It is non parametric test used to determine the randomness of the selected samples.

Run is a set of identical or related symbols contained between two different symbols or none at all. In sequence HHHTTHHHTHHHTTH HHHTT number of run is 8.

Let us consider a random sample of size n is selected from non normal population. Let $x_1, x_2, x_3, x_4 \ldots\ldots\ldots\ldots x_n$ be samples.

Different steps in the test are

Problem to test

H_0 : Sample observations are in random order

H_1: Sample observations are not in random order.

First find the median of sample and assign a symbol (say A) for x_i if $x_i > Md$, assign next symbol (say B) for x_i if $x_i < Md$ and omit if $x_i = Md$ (called tie)

Count the number of runs (r) of the symbols, number of observations of symbol A denoted by n_1 and number of observations of symbol B denoted by n_2.

For small sample size ($n_1, n_2 \leq 20$)

Test statistic

Number of runs (r)

Level of significance

Let α be the level of significance. Usually we take $\alpha = 0.05$ unless we are given.

Critical value

Critical or tabulated value $\underline{r}$ and $\bar{r}$ is obtained from table according to the level of significance α, degree of freedom n_1 and n_2 and alternative hypothesis.

Decision

Accept H_0 at α level of significance if $r \in (\underline{r}, \bar{r})$, reject otherwise.

For large sample size (n_1 or $n_2 > 20$)

In case of large sample size r is approximately normally distributed with mean $\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$

And variance $\sigma_r^2 = \frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}$

Test statistic

$$Z = \frac{r - \mu_r}{\sigma_r} \sim N(0, 1)$$

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of Z is obtained from table according to the level of significance, and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $|Z| > Z_{\text{tabulated}}$, accept otherwise.

Note: For α other than 5% even if sample size is small use Z test in which $Z = \frac{|r - \mu_r| - 0.5}{\sigma_r}$.

Example 1: In 15 toss of a coin the following sequence of heads (H) and tails (T) is obtained H H T TT H T H HH TT H H T .

Test at 0.05 level of significance whether the sequence is random.

Solution:

Here, Number of heads (n_1) = 8, Number of tails (n_2) = 7, Number of runs (r) = 8, Level of significance (α) = 0.05

Problem to test

H_0 : The sequence is in random order

H_1: The sequence is not in random order

Test statistic

r = 8

Critical value

At $\alpha = 0.05$ level of significance, the critical value for n_1=8 and n_2 =7 degree of freedom are $\underline{r} = 4$ and $\bar{r} = 13$ for two tailed test.

Decision

$r = 8 \in (\underline{r} = 4, \bar{r} = 13)$, accept H_0 at 0.05 level of significance.

Conclusion

The sequence of H and T are in random order.

Example 2: The following is the arrangement of defective (d) and non defective (n) pieces of keyboard produced in the given order by a certain machine:

n n n n n d d d d n n n n n n n n n n d d n n d d d d n n n d d n n d n n n

Test the randomness at the 0.01 level of significance.

Solution:

Here, number of n $(n_1) = 25$

Number of d $(n_2) = 13$ number of runs $(r) = 11$

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2 \times 25 \times 13}{25 + 13} + 1 = 8.105$$

$$\sigma_r^2 = \frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}$$

$$= \frac{2 \times 25 \times 13(2 \times 25 \times 13 - 25 - 13)}{(25 + 13)^2\,(25 + 13 - 1)}$$

$$= \frac{397800}{53428} = 7.44$$

$$\sigma_r = \sqrt{7.44} = 2.72.$$

Problem to test

H_0 : Arrangement is in random order

H_1 : Arrangement is not in random order.

Test statistic

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{11 - 8.105}{2.72} = 1.064$$

Critical value

At $\alpha = 0.01$ level of significance critical value is Ztabulated = $Z_{\alpha/2} = 2.576$.

Decision

$Z = 1.064 < Z_{tabulated} = 1.96$, accept H_0 at 0.01 level of significance

Conclusion

The defective and non defective pieces produced by machine are in random order.

Example 3: The height (in inches) of 16 BSC CSIT students are as follows:

68.2,71.6,69.3,71.6,70.4, 65.0, 63.6,64.7,65.3,64.2,67.6,68.6,66.8,68.9,66.8, 70.1

Test whether the order of the heights is random or not? Use 5% level of significance.

Solution:

Here, n = 16, to find run first of all find median. To find the median, arranging data in ascending order

63.6, 64.2, 64.7, 65.0, 65.3, 66.8, 66.8, 67.7, 68.2, 68.6, 68.9, 69.3, 70.1, 70.4, 71.6, 71.6

$$Md = \frac{(n+1)}{2}^{th} \text{item} = \frac{16+1}{2} = 8.5^{th} \text{ item}$$

Hence, median = (8^{th} item + 9^{th} item)/2 = (67.7+68.2)/2 = 67.95

Let the number greater than 67.95 is denoted by A and number less than 67.95 is denoted by B, then the arrangement of given sample are

A AAAA B BBBBB A B A B A

Here, number of A (n_1) = 8, number of B (n_2) = 8, number of run (r) = 7

Problem to test

H_0 : Order of heights are in random order

H_1: Order of heights are not in random order.

Test statistic

r = 7

Critical value

Let α = 0.05 be the level of significance, then critical value for n_1 = 8 and n_2 = 8 are $\underline{r}$ = 4 and $\bar{r}$ = 14

Decision

$r = 7 \in (\underline{r} = 4, \bar{r} = 14)$, accept H_0 at 0.05 level of significance.

Conclusion

The order of heights are in random order.

Example 4: The following are the no. of emails arrived within one hour in a communication room: 46, 58, 60, 56, 70, 66, 48, 54, 62, 41, 39, 52, 45, 62, 53, 69, 65, 65, 67, 76, 52, 52, 59, 59, 67, 51, 46, 61, 40, 43, 42, 77, 67, 63, 59, 63, 63, 72, 57, 59, 42, 56, 47, 62, 67, 70, 63, 66, 69 and 73. Given that median is 59.5

Test the randomness at the 0.05 level of significance.

Solution:

Here, Sample size (n) = 50 , Median (Md) = 59.5, Level of significance (α) = 0.05

Assign A for number greater than 59.5 and assign B for number less than 59.5 then arrangement of sample becomes

B B A B A A B B A B BBB A B AAAAA B BBB A B B A B BB A AA B A AA B BBBB A AAAAAA.

Number of B (n_1) = 25 , Number of A (n_2) = 25, Number of run (r) = 20

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2 \times 25 \times 25}{25 + 25} + 1 = 26$$

$$\sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} = \frac{2 \times 25 \times 25(2 \times 25 \times 25 - 25 - 25)}{(25 + 25)^2(25 + 25 - 1)} = 12.2,$$

$$\sigma_r = \sqrt{12.2} = 3.5$$

Problem to test

H_0 : Emails are in random order

H_1: Emails are not in random order.

Test statistic

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{20 - 26}{3.5} = -1.71$$

Critical value

At $\alpha =$ 0.05 level of significance, critical value is $Z_{tabulated} = Z_{\alpha/2} = 1.96$.

Decision

$| Z | = 1.71 < Z_{tabulated} = 1.96$, accept H_0 at 0.05 level of significance.

Conclusion

The sample is in random order.

Binomial Test

It is non parametric test used for data present in either nominal or ordinal scale. It is used to test whether the binomial population has two distinct groups of two equal numbers of outcomes or not.

Let us consider a sample of size n is selected from binomial population (dichotomized population)

Let $x_1, x_2, x_3, \ldots\ldots x_n$ be independent sample of size n selected from binomial population having probability mass function $P(x) = c(n,x)\, p^x(1 - p)^{n-x}$ such that probability associated with each trial is equal. x is number of observations having certain characteristic and n – x is number of observations having next characteristic then $p = \frac{x}{n}$

Different steps in the test are

Problem to test

H_0: $P = P_0$

H_1: $P \neq P_0$ (two tailed) or H_1: $P > P_0$ one tail) or H_1: $P < P_0$ (one tail)

Let n_1 and n_2 be the number of observations belonging to two groups from n samples.

$X_0 = \min(n_1, n_2)$

Small sample size ($n \leq 25$)

Test statistic

$X_0 = \min\{n_1, n_2\}$

Level of significance

Let α be the level of significance. Generally fix α= 0.05 unless we are given.

Critical value

Using the binomial distribution $p = \text{Prob}(X \le x_0) = \sum_{x=0}^{x_0} C(n, x)\, p^x (1-p)^{n-x} = \sum_{x=0}^{x_0} C(n, x) \left(\frac{1}{2}\right)^n$

Decision

Accept H_0 at α level of significance if $p > \alpha$ for one tailed test and $2p > \alpha$ for two tailed test, reject otherwise

Large sample size (n > 25)

For large sample size, x_0 is normally distributed with mean np and variance npq

Test statistic

$$Z = \frac{X_0 - \mu}{\sigma} = \frac{X_0 - np}{\sqrt{npq}}$$

Since X_0 is discrete so that continuity correction is made as

$$Z = \frac{(x_0 \pm 0.5) - np}{\sqrt{npq}}$$, use + 0.5 if $x_0 < np$ and use – 0.5 if $x_0 > np$

Level of significance

Let α be the level of significance. Generally fix α = 0.05 unless we are given.

Critical value

Critical value $Z_{tabulated}$ is obtained from table according to level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $Z > Z_{tabulated}$, Accept otherwise.

Example 5: The following are defective (D) and non defective (N) electronic items produced in the given order by a certain machine:

N N D D N D N N N D D N D D N N D N D N N N

Test whether defective and non defective are equally produced or not. Use Binomial test at the 0.01 level of significance.

Solution:

Here number of N $(n_1) = 13$

Number of $D(n_2) = 9$

Problem to test

$H_0 : P = \frac{1}{2}$ $\qquad$ $H_1 : P \neq \frac{1}{2}$

Test statistics

$X_0 = \min\{n_1, n_2\} = \min(13, 9) = 9$

Level of significance

$\alpha = 0.01$

Critical value

$$P = \sum_{x=0}^{x_0} C(n, x)\left(\frac{1}{2}\right)^n = \sum_{x=0}^{9} C(n, x)\left(\frac{1}{2}\right)^n = 0.262$$

$2p = 2 \times 0.262 = 0.524$

Decision

$2P = 0.524 > \alpha = 0.01$

Accept H_0 at 0.01 level of significance.

Conclusion

Defective and non defective are produced equally.

Example 6: Out of 50 students willing to express opinion of laptop 30 expressed preferences to brand Dell and 20 for brand Lenovo. Use binomial test to test hypothesis that both brand of laptop are equally popular

Solution:

Here number of female students prefer Dell laptop $(n_1) = 30$

Number of female students prefer Lenovo laptop $(n_2) = 20$

Problem to test

$H_0 : P = \frac{1}{2}$

$H_1 : P \neq \frac{1}{2}$

Test statistics

$X_0 = \min\{n_1, n_2\} = \min(20, 30) = 20$

$np = 50 \times \frac{1}{2} = 25$

$$\begin{aligned} Z &= \frac{(x_0 \pm 0.5) - np}{\sqrt{npq}} \\ &= \frac{(x_0 + 0.5) - np}{\sqrt{npq}} \\ &= \frac{(20 + 0.5) - 25}{\sqrt{12.5}} \\ &= -1.27 \end{aligned}$$

Level of significance

Let $\alpha = 0.05$

Critical value

$Z_{\alpha/2} = 1.96$

Decision

$|Z| = 1.27 < Z_{\alpha/2} = 1.96$

Accept H_0 at 0.05 level of significance.

Conclusion

Both brand of laptop are equally popular

Kolmogorov Smirnov Test

Kolmogorov Smirnov one sample test is a test of goodness of fit.It is alternate to chi square test for goodness of fit when sample size is small.

Let $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ be random sample of size n from population having distribution function F(x).

Let us consider n observations of random variable x is classified into k classes with their respective frequencies

Different steps in the test are;

Problem to test

H_0: Samples come from population with distribution $F_0(x)$

H_1: Samples do not come from population with distribution $F_0(x)$

Obtain the observed relative frequency F_o and expected relative frequency F_e respectively. Then find the absolute deviation of F_e and F_0 i.e., $|F_e - F_0|$

Test statistic

$D_0 = \text{Max } |F_e - F_0|$

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

Critical value $D_{n,\,a}$ is obtained according to the level of significance α and sample size n

Decision

Reject H_0 at α level of significance if $D_0 \geq D_{n,\,a}$ accept otherwise.

Example 7: The number of laptop in 10 different department are given below. Test whether the laptops are uniformly distributed over the entire office use Kolmogorov smirnov test.

Department No.	1	2	3	4	5	6	7	8	9	10
No of laptop	8	10	9	12	15	7	5	12	13	9

Solution:

Her expected frequency for each class = Sum of frequencies/No of class

Department no.	No of laptop	Observed cf (cf_0)	Observed relative freq (F_0)	Expected freq (f_e)	Expected cf (cf_e)	Expected relative freq(F_e)	$\lvert F_e - F_0 \rvert$
11	8	8	8/100	10	10	10/100	2/100
2	10	18	18/100	10	20	20/100	2/100
3	9	27	27/100	10	30	30/100	3/100
4	12	39	39/100	10	40	40/100	1/100
5	15	54	54/100	10	50	50/100	4/100
6	7	61	61/100	10	60	60/100	1/100
7	5	66	66/100	10	70	70/100	4/100
8	12	78	78/100	10	80	80/100	2/100
9	13	91	91/100	10	90	90/100	1/100
10	9	100	100/100	10	100	100/100	0/100

Problem to test

H_0: The disease infected plants are uniformly distributed over the entire area.

H_1: The disease infected plants are not uniformly distributed over the entire area.

Test statistic

$D_0 = \text{Max } \lvert F_e - F_0 \rvert = 4/100 = 0.04$

Critical value

Let 5% be the level of significance then critical value is $D_{100,0.05} = \dfrac{1.36}{\sqrt{100}} = 0.136$.

Decision

$D_0 = 0.04 < D_{10,0.05} = 0.136$, accept H_0 at 5% level of significance.

Conclusion

The disease infected plants are uniformly distributed over the entire area.

Example 8: A random sample of 20 volume based internet connected have following speed of internet connection in mps:

2.7 , 2.9 , 3.0 , 3.1 , 2.8 , 3.0 , 2.9 , 3.0 , 2.6 , 3.1 , 3.2 , 3.1 , 3.0 , 2.9 , 3.3 , 3.0 , 2.8 , 2.9 , 3.0 , 2.9

Apply the Kolmogorov Smirnov test for testing that the internet speed are equally distributed.

Solution:

Expected frequency for each class = Σf_0/no of class = 20/8 = 2.5

Speed	Tally bar	No of internet connection	cf_0	F_0	Expected Freq (f_e)	cf_e	F_e	$\lvert F_e - F_0 \rvert$
2.6	|	1	1	1/20	2.5	2.5	2.5/20	1.5/20
2.7	|	1	2	2/20	2.5	5	5/20	3/20
2.8	||	2	4	4/20	2.5	7.5	7.5/20	3.5/20
2.9	卌	5	9	9/20	2.5	10	10/20	1/20
3.0	卌 |	6	15	15/20	2.5	12.5	12.5/20	2.5/20
3.1	|||	3	18	18/20	2.5	15	15/20	3/20
3.2	|	1	19	19/20	2.5	17.5	17.5/20	1.5/20
3.3	|	1	20	20/20	2.5	20	20/20	0/20

Problem to test

H_0: Internet speeds are equally distributed.

H_1: Internet speed is not equally distributed.

Test statistic

$D_0 = \text{Max } |F_e - F_0| = 3.5/20 = 0.175$

Critical value

Let 5% be the level of significance then critical value is $D_{20,0.05} = 0.294$

Decision

$D_0 = 0.175 < D_{20,\,0.05} = 0.294$, accept H_0 at 5% level of significance.

Conclusion

Children of different heights are equally enrolled.

EXERCISE

1. What do you mean by non parametric test? Write down advantages of non parametric test over parametric test.
2. Differentiate between parametric test and non parametric test.
3. Discuss assumptions of non parametric test.
4. Write down steps of non parametric test.
5. Describe the function and procedure of run test.
6. What is binomial text? Why is it used?
7. Describe the method of binomial test
8. Discuss process of kolmogorov Smirnov test.
9. In 31 toss of a coin the following sequence of heads (H) and tails (T) is obtained.

   H T T H T H HH T H H TT H T H T H H T H T T H T H H T H T H

   Test at 0.05 level of significance level whether the sequence is random. **Ans:** r = 23, sig.

10. A random samples of 15 adults living in a small town is selected to estimate the proportion of voting favoring a certain candidate for Mayer. Each individual was also asked if he or she was a college graduate. By letting Y and N designate the response of yes and no to the education question, the following sequence was obtained N NNN Y Y N Y Y N Y N NNN. Use the run test to determine if the sequence supports the condition that the sample was selected at random. **Ans:** r = 7, insig.

11. In production line touch screen are inspected periodically for defectives .The following is sequence of defective item D and non defective item N produced by production line.
D D N NN D N N D D N NNNN D DD N N D N NNN D N D.
Use run test with a significance level 0.05 to determine whether the defectives are occurring at random order. **Ans:** r = 13, insig.

12. The following are the number of IT students absent from a college on 24 consecutive college days: 29, 25, 31, 28, 30, 28, 33, 31, 35, 29, 31, 33, 35, 28, 36, 30, 33, 26, 30, 28, 32, 31, 38 and 27.Test for randomness at the 0.01 level of significance. **Ans:** z = 0, insig.

13. The height (in inches) of 15 football players are as follows:
65.3, 67.7, 68.4,71, 70.2, 67, 69.8, 64, 67.6, 71.5, 64.4, 63.6, 66.8, 70.4, 69.
Test whether the order of the heights is random or not .Use 1% level of significance.
Ans: z = 0, insig.

14. The following are defective (D) and non defective (N) electric cable produced in the given order by a certain machine in a manufacturing industry:
D N D D N D NDN N D N N D D N N D N D D N N N N
Test whether defective and non defective cables are equally produced. Use Binomial test at the 5% level of significance. **Ans:** p = 0.345, insig.

15. 100 workers of sales house willing to express opinion of cell phones, 40 expressed preferences to brand Nokia and 60 for brand Oppo. Are both brand of cell phone equally popular at 5% level of significance. Use Binomial test **Ans:** p = – 1.9, insig.

16. In a certain computer hardware manufacturing industry six different types of machines are working to cut pieces of wires. The number of wires of unequal length recorded in a day is as follows:

Machine	1	2	3	4	5	6
No of wire	2	0	4	8	5	11

Do these data provide sufficient evidence that the six machines equally cut the wires of unequal length? Apply Kolmogorov Smirnov test at 5% level of significance. **Ans:** D_0 = 0.3, sig.

17. The number of virus infected computers of five different capacity of hard disk is given below:

Capacity of hard disk (GB)	500	320	1000	2000	4000
No of virus infected	11	15	20	3	1

Test whether the computers of five hard disk are uniformly infected using Kolmogorov Smirnov test.
Ans: D_0 =0.333, insig.

18. A game consists of four pairs of color cards. Twenty chimpanzees of same age were taught the matching game of color cards for a specified period of time. At the end of the training 4 pairs of color cards are given to each of the chimpanzee for matching. The result were as follows;

Matched set	0	1	2	3	4
Frequency	1	0	5	7	7

Does chimpanzees recognize colors? Use Kolmogorov Smirnov test at 5% level of significance.
Ans: D = 0.35, insig.

Two Independent Sample Test

Median Test

It is non parametric test used to test the significance difference between two independent distribution.

Let $x_1, x_2, x_3 \ldots\ldots x_{n_1}$ and $y_1, y_2, y_3 \ldots\ldots y_{n_2}$ be two independent samples of sizes n_1 and n_2 respectively be selected from continuous populations with unknown medians Md_1 and Md_2 respectively.

Different steps in the test are

Problem to test

$H_0: Md_1 = Md_2$

$H_1: Md_1 \pm M_2$ (Two tailed test) or $H_1: Md_1 > Md_2$ (One tailed right) or $H_1: Md_1 < Md_2$ (one tailed left)

Small sample size ($n_1 \leq 10, n_2 \leq 10$)

Combine n_1 and n_2 such that $n = n_1 + n_2$ and obtain median of n observation Find number of observations in $x_i \leq Md$ and denote by a.

Test statistic

$$P(A=a) = \frac{c(n_1, a)\, c(n_2, k-a)}{c(n_1 + n_2, k)} : a = 0, 1, 2, \ldots\ldots \min(n_1, k),\ k = \frac{n_1 + n_2}{2} = \frac{n}{2}$$

Level of significance

Let α be the level of significance. Generally we fix $\alpha = 0.05$ unless we are given.

Critical value

Critical value is given by $p = P(A \geq a)$

Decision

Accept H_0 at α level of significance if $p > \alpha$ for one tailed and $2p > \alpha$ for two tailed test, reject otherwise.

Large sample size ($n_1 > 10, n_2 > 10$)

Combine n_1 and n_2 such that $n = n_1+n_2$ and obtain median of n observation. Find number of observations in $x_i \leq Md$ and denote by

a. Find number of observations in $y_i \leq Md$ and denote by b.

Also find the number of observations in x_i>Md an y_i>Md and denote by c and d respectively.

	No. of obs.≤ Md	No. of obs.>Md	Total
Sample x	a	c	a+c
Sample y	b	d	b+d
Total	a+b	c+d	N=a+b+c+d

Test statistic

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \sim \chi^2_{(1)}$$

If any cell frequency is less than 5 then

$$\text{Corrected } \chi^2 = \frac{N(|ad - bc| - \frac{N}{2})^2}{(a + c)(b + d)(a + b)(c + d)} \sim \chi^2_{(1)}$$

Level of significance

Let α be the level of significance. Generally we fix $\alpha = 0.05$ unless we are given.

Critical value

At α level of significance for 1 degree of freedom critical value is $\chi^2_{\alpha(1)}$

Decision

Reject H_0 at α level of significance if $\chi^2 > \chi^2_{\alpha(1)}$, accept otherwise.

Example 1: Two independent samples are given below;

Sample I	10	11	8	8	14		
Sample II	9	12	13	9	15	9	17

Test whether the two samples have come from the same population with respect to their medians. Use median test at 0.05 level of significance.

Solution:

Here, to find the median of combined group, arranging data in ascending order,

8, 8, 9, 9, 9, 10, 11, 12, 13, 14, 15, 17.

$$n = 12, \text{ Md} = \frac{(n + 1)}{2}^{\text{th}} \text{ item} = \frac{(12 + 1)}{2} = 6.5^{\text{th}} \text{ item.}$$

$$\text{Hence median} = \frac{10 + 11}{2} = 10.5$$

Now no of observations in first sample less than or equal to median (a) = 3

First sample size (n_1) = 5

$$\text{Second sample size } (n_2) = 7, k = \frac{n_1 + n_2}{2} = \frac{(5 + 7)}{2} = 6$$

Let Let Md_1 and Md_2 be median of I population and II population respectively.

Problem to test

H0: There is no significant difference between median of I population and median of II population ($Md_1 = Md_2$)

H_1: There is significant difference between median of I population and II population ($Md_1 \neq Md_2$)

Test statistic

$$P(A=a) = \frac{c(n_1, a)\, c(n_2, k-a)}{c(n_1 + n_2, k)} = \frac{c(5, a)\, c(7, 6 - a)}{c(12, 6)}, a = 0, 1, 2, 3, 4, 5$$

Critical value

$$P = P(A \geq a) = P(A \geq 3) = \sum \frac{c(5, a)\, c(7, 6 - a)}{c(12, 6)}$$

$$= \frac{c(5, 3)\, c(7, 6 - 3)}{c(12, 6)} + \frac{c(5, 4)\, c(7, 6 - 4)}{c(12, 6)} + \frac{c(5, 5)\, c(7, 6 - 5)}{c(12, 6)}$$

$$= 462/9 \quad 24 = 0.5$$

Decision

$2P = 1 > \alpha = 0.05$, accept H_0 at 0.05 level of significance.

Conclusion

Two samples have come from same population with respect to median.

Example 2: An IQ test was given to a random sample of 15 male and 20 female students of a university. Their scores were recorded as follows;

Male: 56, 66, 62, 81, 75, 73, 83, 68, 48, 70, 60, 77, 86, 44, 72

Female: 63, 77, 65, 71, 74, 60, 76, 61, 67, 72, 64, 65, 55, 89, 45, 53, 68, 73, 50, 81

Use median test to determine whether IQ of male and female students is same in the university. (Given that the median of combined sample =68)

Solution:

Here Number os male (n_1) = 15, Number of female (n_2) = 20, $N = n_1 + n_2 = 15 + 20 = 35$,

a(no. of obs. of male ≤ Md) = 7 , b(no. of obs. of female ≤ Md) = 12

c(no. of obs. of male > Md) = 8, d(no. of obs. of female > Md) = 8

The 2 × 2 contingency table is

	Sample I	Sample II	Total
No. of obs. ≤ Md	7 (a)	12 (b)	19
No. of obs > Md	8 (c)	8 (d)	16
	15	20	35

Let Md_1 and Md_2 be median IQ of male and female respectively.

Problem to test

H_0: There is no significant difference between IQ of male and female ($Md_1 = Md_2$)

H_1: There is significant difference between IQ of male and female. ($Md_1 \neq Md_2$)

Test statistic

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)\,(b + d)\,(a + b)\,(c + d)} = \frac{35(7 \times 8 - 12 \times 8)^2}{15 \times 20 \times 19 \times 16} = 0.61$$

Critical value

Let $\alpha = 0.05$ be the level of significance , then critical value at 0.05 level of significance with 1 degree of freedom is $\chi^2_{(1)} = 3.84$.

Decision

$\chi^2 = 0.61 < \chi^2_{(1)} = 3.84$, accept H_0 at 0.05 level of significance.

Conclusion

IQ of male and female is same in the university.

Two Sample Kolmogorov Smirnov Test

It is non parametric test used to test whether two independent samples are from same population or not.

Let $x_1, x_2, x_3, \ldots\ldots x_{n_1}$ and $y_1, y_2, y_3, \ldots\ldots y_{n_2}$ be independent samples of size n_1 and n_2 drawn from continuous population having distribution function F(x) and F(y) respectively.

Different steps in the test are

Problem to test

H_0: F(x) = F(y)

H_1: F(x) ≠ F(y)(two tailed) or H_1: F(x) > F(y)one tail) or H_1: F(x) < F(y) (one tail)

Obtain cumulated distribution function of x and y separately after arranging in order.

$F(x) = \frac{k_1}{n_1}$, where k_1 is observed cumulative frequency of x

$F(y) = \frac{k_2}{n_2}$, where k_2 is observed cumulative frequency of y

Small sample test ($n_1 = n_2 < 40$, n_1 and $n_2 \leq 20$ for $n_1 \neq n_2$)

Test statistic

D_0 = maximum { | F(x) – F(y) |

Level of significance

Let α be level of significance

Critical value

At α level of significance critical value for n_1 and n_2 is

$D_{n_1, n_2, a}$

Decision

Reject H_0 at α level of significance if $D_0 \geq D_{n_1, n_2, \alpha}$

Accept otherwise

Large sample test ($n_1 = n_2 > 40$, n_1 and $n_2 > 20$ for $n_1 \neq n_2$)

Test statistic

D_0 = maximum { | F(x) – F(y) | for two tail test

$\chi^2 = 4D_0^2 \frac{n_1 n_2}{n_1 + n_2}$ follows chi square distribution with 2 degree of freedom. For one tail test

Level of significance

Let α be the level of significance

Critical value

At α level of significance critical value is

$D_\alpha = 1.36\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$ for two tail with α = 5%

$\chi^2_{\alpha, 2}$ for one tail.

Decision

Reject H_0 at α level of significance if $D_0 \geq D_\alpha$ for two tail

if $\chi^2 > \chi^2_{\alpha 2}$ for one tail

Accept otherwise.

Example 3: Life in years of two types of cells used in laptop are given below:

Cell X	4	6	5	6	3	4	5	3	5
Cell Y	2	5	4	3	4	2	4	3	5

Test whether life of two brands of cells are same or not? Use Kolmogorov Smirnov test at 0.05 level of significance.

Solution:

Problem to test

$H_0: F(x) = F(y)$

$H_1: F(x) \neq F(y)$

Combined ordered life	Frequency for x	Frequency for Y	F(x)	F(y)	|F(x) – F(y)|
2	0	2	0/9	2/9	2/9
3	2	2	2/9	4/9	2/9
4	2	3	4/9	7/9	3/9
5	3	2	7/9	1	2/9
6	2	0	1	1	0
Total	9	9			

Test statistics

$D_0 = \text{maximum } \{|F(x) - F(y)| = 3/9$

Critical value

At $\alpha = 0.05$ and $n_1 = 9, n_2 = 9$ from K – S table

$Dn_1, n_2, \alpha = 5/9$

Decision

$D = 3/9 < D\, n_1, n_2, \alpha = 5/9$

Accept H_0 at 0.05 level of significance.

Conclusion

Two brands of cell X and Y are same.

Example 4: Given below represents monthly income distribution of employees in a hardware company and software company

Income (000 rs)	Number of employees in hardware company	Number of employees in software company
20 – 30	6	12
30 – 40	10	18
40 – 50	11	16
50 – 60	13	12
60 – 70	25	10
70 – 80	15	12
80 – 90	10	10

Do the income distribution support that income of employees in Hardware Company is more than income of employees in Software Company? Use Kolmogorov Smirnov test.

Solution:

Let x be salary of employees in hardware company and y be salary of employees in software company

Problem to test:

$H_0 : H_0: F(x) = F(y)$

$H_1 : F(x) > F(y)$

Income (000 Rs)	Number of employees in hardware company (x)	Number of employees in software company (y)	F(x)	F(y)	|F(x) - F(y)|
20 – 30	6	12	6/100	12/90	66/900
30 – 40	10	18	16/100	30/90	156/900
40 – 50	11	16	27/100	46/90	57/900
50 – 60	13	12	40/100	58/90	220/900
60 – 70	25	10	65/100	68/90	95/900
70 – 80	20	12	85/100	80/90	35/900
80 – 90	15	10	1	1	0
Total	100	90			

Test statistic

$$D_0 = \text{maximum } \{ |F(x) - F(y)| = \frac{220}{900} = \frac{11}{45}$$

$$\chi^2 = 4D_0^2 \frac{n_1 n_2}{n_1 + n_2} = 4 \times \left(\frac{11}{45}\right)^2 \frac{100 \times 90}{100 + 90} = 11.32$$

Critical value

Let $\alpha = 0.05$ then critical value is $\chi^2_{\alpha,2} = 5.99$.

Decision

$\chi^2 = 11.32 > c^2_{a,\,2} = 5.99.$

Reject H_0 at $\alpha = 0.05$ level of significance.

Conclusion

Income of employees in Hardware Company is more than income of employees in Software Company.

Example 5: Life time of cell phones branded A and B recorded from a repair centre is given below, determine if there is any significant difference between two brand cell phones using Kolmogorov Smirnov test at 5% level of significance.

Life of cell phone in years	Cell phone A	Cell phone B
0 - 2	8	5
2 - 4	12	7
4 - 6	16	31
6 - 8	10	12
8 and above	4	5

Solution:

Let cell phone A = x and cell phone B = y

Problem to test

$H_0: F(x) = F(y)$

$H_1: F(x) \neq F(y)$

Life of cell phone in years	Cell phone A	Cell phone B	F(x)	F(y)	|F(x) – F(y)|
0 - 2	8	5	8/50	5/60	23/300
2 - 4	12	7	20/50	12/60	60/300
4 - 6	16	31	36/50	43/60	1/300
6 - 8	10	12	46/50	55/60	1/300
8 and above	4	5	1	1	0
Total	50	60			

Test statistic

$D_0 = \text{maximum }\{|F(x) - F(y)| = \frac{60}{300} = 0.2$

Critical value

At $\alpha = 0.05$ critical value is $D_a = 1.36\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36\sqrt{\frac{50 + 60}{50 \times 60}} = 0.26.$

Decision

$D_0 = 0.2 < D_a = 0.26$

Accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between two brand cell phones.

Mann Whitney U Test

It is non parametric test used to determine whether two independent samples have been drawn from populations with same distribution.

Let us consider two independent samples of sizes n_1 and n_2 are drawn from continuous populations with unknown medians Md_1 and Md_2 respectively.

Let $x_1, x_2, x_3, \ldots\ldots, x_{n_1}$ and $y_1, y_2, y_3, \ldots\ldots, y_{n_2}$ be independent samples of size n_1 and n_2.

Different steps in the test are

Problem to test

$H_0: Md_1 = Md_2$

$H_1: Md_1 \neq Md_2$ (two tailed) or $H_1: Md_1 > Md_2$ (one tailed right)

or $H_1: Md_1 < Md_2$ (one tailed left)

Combine n_1 and n_2 such that $n_1+n_2 = n$ and rank these n observations in ascending order. If two or more observations are equal then assign average rank and is called tied. Sum the ranks of sample of sizes n_1 and n_2 separately to get R_1 and R_2. If two sample sizes are unequal then smaller one is n_1. Obtain U_1 and U_2 as $U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1$ and $U_2 = n_1n_2 + \frac{n_2(n_2+1)}{2} - R_2$.

such that $n_1n_2 = U_1 + U_2$. Finally get $U_0 = \min\{U_1, U_2\}$

Small sample size ($n_1 \le 10$, $n_2 \le 10$)

Test statistic

$U_0 = \min\{U_1, U_2\}$

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

At α level of significance, we obtain critical value from Mann Whitney table as

$p = \text{Prob}(U \le U_0)$

Alternately

At α level of significance, we obtain critical value from Mann Whitney table as

$U_{tabulated} = U_{\alpha(n1,n2)}$ for two tail and $U_{\alpha/2(n1,n2)}$ for one tail test.

Decision

Accept H_0 at α level of significance if $p > \alpha$ for one tailed test and $2p > \alpha$ for two tailed test, reject otherwise.

Alternately

Accept H_0 at α level of significance if $U_0 > U_{tabulated}$ reject otherwise.
Large sample size ($n_1 > 10$, $n_2 > 10$)

For large sample size U_0 is approximately normally distributed with mean $\mu_u = \frac{n_1n_2}{2}$ and variance $\sigma_u^2 = \frac{n_1n_2(n_1+n_2+1)}{12}$.

Test statistic

$$Z = \frac{U_0 - \mu_\alpha}{\sigma_n} = \frac{U_0 - \frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}}$$

it is used even if tied occurs within sample. If tied occurs between samples then standard deviation is corrected as $\sigma_u = \frac{n_1 n_2}{n(n-1)}\left\{\frac{n^3-n}{12} - \frac{\Sigma t_i^3 - ti}{12}\right\}$

t_i = number of times i^{th} rank repeated between samples.

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

Critical value $Z_{tabulated}$ is obtained from table according to level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $Z > Z_{tabulated}$, Accept otherwise.

Example 6: The heart beating rate of 5 vegetarians and 5 non vegetarians are recorded below:

Vegetarians	56	67	82	60	75
Non vegetarians	53	42	75	58	65

Is the mean heart beating rate of non vegetarians significantly high. Use Mann Whitney U test.

Solution:

Vegetarians	Ranks	Non vegetarians	Ranks
56	3	53	2
67	7	42	1
82	10	75	8.5
60	5	58	4
75	8.5	65	6
	$R_1 = 33.5$		$R_2 = 21.5$

Here, Sample size of vegetarian $(n_1) = 5$

Sample size of Non vegetarian $(n_2) = 5$

Sum of ranks of vegetarian $(R_1) = 33.5$

Sum of ranks of non vegetarian $(R_2) = 21.5$

$$U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 5 \times 5 + \frac{5 \times 6}{2} - 33.5 = 6.5$$

$$U_2 = n_1n_2 - U_1 = 5 \times 5 - 6.5 = 18.5$$

$$U_0 = \min\{U_1, U_2\} = 6.5$$

Let Md_1 and Md_2 be median heart beat rate of vegetarians and non vegetarians respectively.

Problem to test

H_0: There is no significant difference between heart beating rate of vegetarian and non vegetarian $(Md_1 = Md_2)$

H_1: Heart beating rate of non vegetarian is significantly high than vegetarian $(Md_1 < Md_2)$

Test statistic

$U_0 = 6.5$

Critical value

Let $\alpha = 0.05$ be the level of significance then critical value is $p = 0.111$

Decision

$p = 0.111 > \alpha = 0.05$, accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between heart beating rate of vegetarians and non vegetarians.

Example 7: Test the hypothesis of no difference between the ages of male and female employees of a certain IT company, using the Mann- Whitney U test for the sample data below. Use $\alpha = 0.1$.

Male	35	43	26	44	40	42	33	38	25	26
Female	30	41	34	31	36	32	25	47	28	24

Solution:

Male	Ranks	Female	Ranks
35	12	30	7
43	18	41	16
26	4.5	34	11
44	19	31	8
40	15	36	13
42	17	32	9
33	10	25	2.5
38	14	47	20
25	2.5	28	6
26	4.5	24	1
	$R_1 = 116.5$		$R_2 = 93.5$

Here, Sample size of male $(n_1) = 10$

Sample size of female $(n_2) = 10$

Sum of ranks of male $(R_1) = 116.5$

Sum of ranks of female $(R_2) = 93.5$

$$U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 10 \times 10 + \frac{10 \times 11}{2} - 116.5 = 38.5$$

$$U_2 = n_1n_2 - U_1 = 10 \times 10 - 38.5 = 61.5$$

$$U_0 = \min\{U_1, U_2\} = 38.5$$

Let Md_1 and Md_2 be median age of male and female employees respectively.

Problem to test

H_0: There is no significant difference between age of male and female employee ($Md_1 = Md_2$)

H_1: There is significant difference between age of male and female employee ($Md_1 \neq Md_2$)

Test statistic

$U_0 = 38.5$

Critical value

Let $\alpha = 0.10$ be the level of significance then critical value is $U_{tabulated} = U_{\alpha(n_1, n_2)} = 27$

Decision

$U_0 = 38.5 > U_{\alpha(n_1, n_2)} = 27$, accept H_0 at 0.10 level of significance.

Conclusion:

There is no significant difference between ages of males and female employees.

Example 8: Comparing 2 kinds of emergency flares, a consumer testing service obtained the following burning times (rounded to the nearest tenth of a minute):

Brand C	19.4	21.5	15.3	17.4	16.8	16.6	20.3	22.5	21.3	23.4	19.7	21.0
Brand D	16.5	15.8	24.7	10.2	13.5	15.9	15.7	14.0	12.1	17.4	15.6	15.8

Use Mann Whitney U test at the 0.01 level of significance to check whether it is reasonable to say that the population of burning times of the two kinds of flares are identical.

Solution:

Brand C	Rank	Brand D	Rank
19.4	16	16.5	11
21.5	21	15.8	8.5
15.3	.5	24.7	24
17.4	14.5	10.2	1
16.8	13	13.5	3
16.6	12	15.9	10
20.3	18	15.7	7
22.5	22	14.0	4
21.3	20	12.1	2
23.4	23	17.4	14.5
19.7	17	15.6	6
21.0	19	15.8	8.5
	R_1=200.5		R_2= 99.5

Here,

Sample size of brand C (n_1) =12

Sample size of brand D (n_2) = 12

Sum of ranks of brand C (R_1) = 200.5

Sum of ranks of brand D (R_2) = 99.5

$$U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 12 \times 12 + \frac{12 \times 13}{2} - 200.5 = 21.5$$

$$U_2 = n_1n_2 - U_1 = 12 \times 12 - 21.5 = 122.5$$

$$U_0 = \min\{U_1, U_2\} = \text{Min }\{21.5, 200.5\} = 21.5$$

$$\mu_u = \frac{n_1n_2}{2} = \frac{12 \times 12}{2} = 72$$

Here tied occurs between groups, hence

$$\sigma_u = \frac{n_1 n_2}{n(n-1)}\left\{\frac{n^3-n}{12} - \frac{\Sigma t_i^3 - ti}{12}\right\} = \frac{12 \times 12}{24(24-1)}\left\{\frac{24^3-24}{12} - \frac{2^3-2}{12}\right\} = 0.2608 \times 1149.5 = 299.789$$

$\therefore \quad \sigma_u = 17.3$

Problem to test

H_0: Population burning time of two kinds of flares are identical.

H_1: Population burning time of two kinds of flares are not identical.

Test statistic

$$Z = \frac{U_0 - \mu_v}{\sigma_u} = \frac{21.5 - 72}{17.3} = -2.91$$

Critical value

At $\alpha = 0.01$ level of significance critical value is $Z_{tabulated} = Z_{\alpha/2} = 2.58$

Decision

$|Z| = 2.91 > Z_{tabulated} = 2.58$, reject H_0 at 0.01 level of significance.

Conclusion

Burning time of two kind of flares are different.

Example 9: The following are the scores which random samples of students from 2 groups obtained on a random coding test.

Group I	73	82	39	68	91	75	89	67	50	86	57	65	70
Group II	51	42	36	53	88	59	49	66	25	64	18	76	74

Use Mann Whitney U test at the 0.05 level of significance to test whether or not students from the two groups can be expected to score equally well on the test

Solution:

Group I	Rank	Group II	Rank
73	18	51	8
82	22	42	5
39	4	36	3
68	16	53	9
91	26	88	24
75	20	59	11
89	25	49	6
67	15	66	14
50	7	25	2
86	23	64	12
57	10	18	1
65	13	76	21
70	17	74	19
	$R_1 = 216$		$R_2 = 135$

Here, sample size for group I (n_1) = 13

Sample size for group II (n_2) = 13

Sum of ranks of group I (R_1) = 216

Sum of ranks of group II (R_2) = 135

$$U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 13 \times 13 + \frac{13 \times 14}{2} - 216 = 44$$

$$U_2 = n_1n_2 - U_1 = 13 \times 13 - 44 = 125$$

$$U_0 = \min\{U_1, U_2\} = \text{Min}\ \{44, 125\} = 44$$

$$\mu_u = \frac{n_1n_2}{2} = \frac{13\times13}{2} = 84.5$$

$$\sigma_u = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}} = \sqrt{\frac{13 \times 13(13+13+1)}{12}} = \sqrt{380.25} = 19.5$$

Let Md_1 and Md_2 be median scores of group I and group II respectively.

Problem to test

H_0 : Students from two groups score equally($Md_1 = Md_2$)

H_1: Students from two groups do not score equally ($Md_1 \neq Md_2$)

Test statistic

$$Z = \frac{U_0 - \mu_u}{\sigma_u} = \frac{44 - 84.5}{19.5} = -2.07$$

Critical value

At $\alpha = 0.05$ level of significance, critical value is $Z_{tabulated} = Z_{\alpha/2} = 1.96$.

Decision

$|Z| = 2.07 > Z_{0.05} = 1.96$, reject H_0 at 0.05 level of significance.

Conclusion

Students from two minority groups do not score equally well

Chi Square Test for Goodness of Fit

It is test used to test the significant difference between observed frequencies (O_i) and expected frequencies (E_i).

Let us consider n observations of random variable x is classified into k classes with their respective frequencies.

Different steps in the test are;

Problem to test

H_0: $O_i = E_i$

H_1 : $O_i \neq E_i$

Test statistic

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

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

Critical value $\chi^2_{\alpha\,(k-1)}$ is obtained according to the level of significance α and degree of freedom (k-1).

Decision

Reject H_0 at α level of significance if $\chi^2 > \chi^2_{\alpha(k-1)}$, accept otherwise.

Note

If any cell frequencies are less than 5 then combine the frequencies of adjoining cells till the resulting cell frequency is greater than or equal to 5.

Example 10: The following table gives the number of air crafts accidents that occurred during seven days of the week. Find whether the accidents are uniformly distributed over week.

Days	Sun	Mon	Tue	Wed	Thu	Fri	Sat
No. of accidents	14	16	8	12	11	9	14

Solution:

Here, $n = 14 + 16 + 8 + 12 + 11 + 9 + 14 = 84$, $k = 7$

Days	No. of accidents(O_i)	P_i	$E_i = NP_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Sun	14	1/7	12	4	0.333
Mon	16	1/7	12	16	1.333
Tue	8	1/7	12	16	1.333
Wed	12	1/7	12	0	0
Thu	11	1/7	12	1	0.0833
Fri	9	1/7	12	9	0.75
Sat	14	1/7	12	4	0.333
Total	$N = \Sigma O_i = 84$				$\sum \frac{(O_i - E_i)^2}{E_i} = 4.166$

Problem to test

H_0: Accidents are uniformly distributed over the week.

H_1: Accidents are not uniformly distributed over the week.

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 4.166$$

Critical value

Let $\alpha = 0.05$ be the level of significance then critical value for k-1 = 6 degree of freedom at 0.05 level of significance is $\chi^2_{0.05(6)} = 12.59$

Decision:

$\chi^2 = 4.166 < \chi^2_{0.05(6)} = 12.59$, accept H_0 at 0.05 level of significance.

Conclusion

Accidents are uniformly distributed over the week.

Example 11: In an experiment on pea breeding, Mendal obtained the following frequencies of seeds: 315 round and yellow, 101 wrinkle and yellow, 108 round and green, 32 wrinkle and green in total of 556. Theory predicts that the frequencies should be in ratio 9:3:3:1 respectively. Set up the proper hypothesis and test it at 10% level of significance.

Solution:

Seeds	Frequencies (o_i)	P_i	$E_i = N P_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Round and yellow	315	9/16	312.75	5.0625	0.01618
Wrinkle and yellow	101	3/16	104.25	10.5625	0.10131
Round and green	108	3/16	104.25	14.0625	0.13489
Wrinkle and green	32	1/16	34.75	2.75	0.07913
Total	$N = \Sigma O_i = 556$				0.3315

Here k = 4

Problem to test

H_0: There is no significant difference between theory and experiment.

H_1: There is significant difference between theory and experiment.

Test statistic

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

$$= 0.3315$$

Critical value

At 0.10 level of significance and 3 degree of freedom, critical value is $\chi^2_{0.05(3)} = 4.166$

Decision: $\chi^2 = 0.3315 < \chi^2_{0.05(3)} = 4.166$, accept H_0 at 10% level of significance.

Conclusion: There is no significant difference between theory and experiment.

Example 12: A die is thrown 132 times with the following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution:

No. turned up	Observed freq.(O_i)	P_i	Expected freq(E_i) $= NP_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	1/6	22	36	36/22
2	20	1/6	22	4	4/22
3	25	1/6	22	9	9/22
4	14	1/6	22	64	64/22
5	29	1/6	22	49	49/22
6	28	1/6	22	36	36/22
Total	$N = \Sigma O_i = 132$				9

Here k = 6

Problem to test

H_0: Die is unbiased ($O_i = E_i$)

H_1: Die is biased ($O_i \neq E_i$)

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 9$$

Critical value

Let $\alpha = 5\%$ be the level of significance, then critical value for 5 degree of freedom at 0.05 level of significance is $\chi^2_{0.05(5)} = 11.07$.

Decision

$\chi^2 = 9 < \chi^2_{0.05(5)} = 11.07$, accept H_0 at 0.05 level of significance.

Conclusion

The die is unbiased.

Example 13: Test whether binomial distribution fits the following data:

x	0	1	2	3	4
f	28	62	46	10	4

Solution:

Problem to test

H_0: Binomial distribution fits the data

H_1: Binomial distribution does not fit the data.

x	f	fx
0	28	0
1	62	62
2	46	92
3	10	30
4	4	16
	$\Sigma f = 150$	$\Sigma fx = 200$

$N = 150$, $n = 4$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{200}{150} = \frac{4}{3}$$

$$\bar{x} = np$$

or $$\frac{4}{3} = 4p$$

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

<table>
<tr><th>x</th><th colspan="2">O_i</th><th>$P_i = c(n,x)p^xq^{n-x}$</th><th colspan="2">$E_i=Np_i$</th><th>$(O_i-E_i)^2$</th><th>$(O_i-E_i)^2/E_i$</th></tr>
<tr><td>0</td><td colspan="2">28</td><td>$C(4,0)(1/3)^0(2/3)^4 = 0.1975$</td><td colspan="2">$29.63 \approx 30$</td><td>4</td><td>0.13</td></tr>
<tr><td>1</td><td colspan="2">62</td><td>$C(4,1)(1/3)^1(2/3)^3 = 0.3951$</td><td colspan="2">$59.26 \approx 59$</td><td>9</td><td>0.15</td></tr>
<tr><td>2</td><td colspan="2">46</td><td>$C(4,2)(1/3)^2(2/3)^2 = 0.2963$</td><td colspan="2">$44.44 \approx 44$</td><td>4</td><td>0.09</td></tr>
<tr><td>3</td><td>10</td><td rowspan="2">14</td><td>$C(4,3)(1/3)^3(2/3)^1 = 0.0988$</td><td>$14.81 \approx 15$</td><td rowspan="2">17</td><td>9</td><td>0.529</td></tr>
<tr><td>4</td><td>4</td><td>$C(4,4)(1/3)^4(2/3)^0 = 0.0123$</td><td>$1.85 \approx 2$</td><td></td><td></td></tr>
<tr><td></td><td colspan="2"></td><td></td><td colspan="2"></td><td></td><td>$\Sigma(O_i - E_i)^2/E_i$
$= 0.899$</td></tr>
</table>

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.899$$

Critical value

Let 0.05 be the level of significance then critical value for 2 degree of freedom is $\chi^2_{0.05(2)} = 5.911$ (here 2 more df is reduced because of estimation of parameter p and combination of two classes)

Decision

$\chi^2 = 0.899 < \chi^2_{0.05(3)} = 5.911$, accept H_0 at 0.05 level of significance.

Conclusion

Binomial distribution fits the data.

Example 14: The number of cyber attack per month at a certain websites was checked for 50 months and the results shown are as follows:

X (no of cyber attack)	0	1	2
f	22	18	10

Assuming the observations are independent, test the hypothesis that the random variable has a poisson distribution.

Solution:

x	f	fx
0	22	0
1	18	18
2	10	20
	$\Sigma f = 50$	$\Sigma fx = 38$

$N = 50$

$\bar{X} = \frac{\Sigma fx}{N} = \frac{38}{50} = 0.76 \qquad \lambda = \bar{X} = 0.76$

Problem to test

H_0: The random variable has poisson distribution.

H_1: The random variable has not poisson distribution.

x	O_i	$P_i = e^{-\lambda}\lambda^x/x!$	$E_i = Np_i$	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
0	22	$e^{-0.76}(0.76)^0/0! = 0.4676$	$23.38 \approx 23$	1	0.0434
1	18	$e^{-0.76}(0.76)^1/1! = 0.3554$	$17.77 \approx 18$	0	0
2	10	$e^{-0.76}(0.76)^2/2! = 0.1350$	$6.75 \approx 7$	9	1.285
					$\Sigma(O_i-E_i)^2/E_i = 1.3284$

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 1.3284$$

Critical value

Let 0.05 be the level of significance then critical value for 1 degree of freedom is $\chi^2_{0.05(1)} = 3.841$ (Here one more df is reduced due to estimation of parameter λ)

Decision

$\chi^2 = 1.3284 < \chi^2_{0.05(1)} = 3.841$, accept H_0 at 0.05 level of significance.

Conclusion

Random variable has poisson distribution.

Chi Square Test for Independence of Attributes

The characteristics which are capable of being measured qualitatively but not quantitatively are called attributes. This test is used to find any association or not between the attributes.

Let us consider a sample of size n is taken from population of unknown distribution. The observations are classified into attributes say A and B into $A_1, A_2, A_3 \ldots\ldots A_r$ and $B_1, B_2, B_3 \ldots\ldots B_c$ classes respectively. Let O_{ij} be the observed frequency of $(i, j)^{th}$ class.

The arrangement of observed frequencies O_{ij} in r x c contingency table is present as

A_i \ B_j	B_1	B_2	B_j	B_c	Total(O_i)
A_1	O_{11}	O_{12}	O_{1j}	O_{1c}	$O_{1.}$
A_2	O_{21}	O_{22}	O_{2j}	O_{2c}	$O_{2.}$
A_i	O_{i1}	O_{i2}	O_{ij}	O_{ic}	$O_{i.}$
A_r	O_{r1}	O_{r2}	O_{rj}	O_{rc}	$O_{r.}$
Total($O_{.j}$)	$O_{.1}$	$O_{.2}$	$O_{.j}$	$O_{.c}$	N

Different steps in the test are

Problem to test

H_0: Attributes A and B are independent.

H_1: Attributes A and B are dependent.

Test statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

$$E_{ij} = (O_{i.} \times O_{.j})/N$$

Level of significance

Let α be the level of significance. Generally fix α = 0.05 unless we are given.

Critical value

At α level of significance and (r – 1)(c – 1) degree of freedom the critical value is $\chi^2_{\alpha\,(r-1)(c-1)}$

Decision

Reject H_0 at α level of significance if $\chi^2 > \chi^2_{\alpha\,(r-1)(c-1)}$

When attributes A and B are arranged into two subgroups then the 2 x 2 contingency table will be

A_i \ B_j	B_1	B_2	Total
A_1	a	b	a+b
A_2	c	d	c+d
Total	a+c	b+d	N

Different steps in the test are

Problem to test

H_0: Attributes A and B are independent.

H_1: Attributes A and B are dependent.

Test statistic

$$\chi^2 = \frac{N(ad - bc)^2}{(a + c)\,(b + d)\,(a + b)\,(c + d)} \sim \chi^2_{(1)}$$

If any cell frequency is less than 5 then the test statistic is

$$\chi^2_{corrected} = \frac{N(|ad - bc| - \frac{N}{2})^2}{(a + c)\,(b + d)\,(a + b)\,(c + d)} \sim \chi^2_{(1)}$$

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

At α level of significance and 1 degree of freedom the critical value is $\chi^2_{\alpha\,(1)}$

Decision

Reject H_0 at α level of significance if $\chi^2 > \chi^2_{\alpha\,(1)}$

Example 15: In an experiment to study the dependence of hypertension on smoking habits, the following data were taken on 180 individuals:

	No smokers	Moderate smokers	Heavy smokers
Hypertension	21	36	36
No hypertension	48	26	19

Test the hypothesis that presence or absence of hypertension is independent of smoking habit.

Solution:

	No smokers (B_1)	Moderate smokers (B_2)	Heavy smokers (B_3)	$O_{i.}$
Hypertension (A_1)	21	36	36	93
No hypertension(A_2)	48	26	19	93
$O_{.j}$	69	62	55	N=186

Now

Group	O_{ij}	$E_{ij} = (O_{i.} \times O_{.j})/N$	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2/E_{ij}$
A_1B_1	21	$(93 \times 69)/186 = 34.5$	-13.5	5.2826
A_1B_2	36	$(93 \times 62)/186 = 31$	5	0.8064
A_1B_3	36	$(93 \times 55)/186 = 27.5$	8.5	2.6272
A_2B_1	48	$(93 \times 69)/186 = 34.5$	13.5	5.2826
A_2B_2	26	$(93 \times 62)/186 = 31$	-5	0.8064
A_2B_3	19	$(93 \times 55)/186 = 27.5$	-8.5	2.6272
Total				17.4324

Problem to test

H_0: Hypertension is independent of smoking habit.

H_1: Hypertension is dependent of smoking habit.

Test statistic

$$\chi^2 = \Sigma\Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 17.4324$$

Critical value

Let $\alpha = 0.05$ be the level of significance, then critical value for (2-1)(3-1) =2 degree of freedom at 0.05 level of significance is $\chi^2_{\alpha\,(2)} = 5.99$

Decision

$\chi^2 = 17.4324 > \chi^2_{\alpha\,(2)} = 5.99$, reject H_0 at 5% level of significance.

Conclusion

Presence or absence of hypertension depends upon the smoking habit.

Example 16: A tobacco company claims that there is no relationship between smoking and lung ailments. To investigate the claims random sample of 300 males in age group of 40 to 50 is given medical test. The observed sample result are tabulated below:

	Lung ailment	No lung ailment	Total
Smokers	75	105	180
No smokers	25	95	120
Total	100	200	300

On the basis of this information, can it be concluded that smoking and lung ailment are independent?

Solution:

	Lung ailment(B)	No lung ailment(β)	Total
Smokers (A)	75 =a	105 =b	180=a+b
No smokers(α)	25 =c	95=d	120=c+d
Total	100 =a+c	200= b+d	300=N

Problem to test

H_0: Smoking and lung ailment are independent

H_1: Smoking and lung ailment are dependent

Test statistic

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

$$= \frac{300(75 \times 95 - 105 \times 25)^2}{100 \times 200 \times 180 \times 120} = 14.063$$

Critical value

Let $\alpha = 0.05$ be the level of significance, then critical value for 1 degree of freedom at 0.05 level of significance is $\chi^2_{0.05(1)} = 3.84$.

Decision

$\chi^2 = 14.063 > \chi^2_{0.05(1)} = 3.84$, reject H_0 at 0.05 level of significance.

Conclusion

Smoking and lung ailment are dependent.

Example 17: In an experiment on immunization of cattle from tuberculosis, the following results were obtained.

	Affected	Not affected
Inoculated	2	10
Not inoculated	6	6

Examine whether the vaccine has an effect in controlling the disease.

Solution:

	Affected	Not affected	Total
Inoculated	2 =a	10 =b	12 =a+b
Not inoculated	6 =c	6 =d	12 =c+d
Total	8= a+c	16= b+d	N=24

Problem to test

H_0: The vaccine has no effect in controlling the disease.

H_1: The vaccine has effect on controlling the disease.

Test statistic

$$\chi^2_{corrected} = \frac{N(|ad - bc| - \frac{N}{2})^2}{(a + c)(b + d)(a + b)(c + d)}$$

$$= \frac{24(|12 - 60| - 12)^2}{8 \times 16 \times 12 \times 12} = 1.687$$

Critical value

Let $\alpha = 0.05$ be the level of significance, then for 1 degree of freedom at 0.05 level of significance critical value is $\chi^2_{0.05(1)} = 3.84$

Decision

$\chi^2 = 1.687 < \chi^2_{0.05(1)} = 3.84$, accept H_0 at 0.05 level of significance.

Conclusion

The vaccine has no effect in controlling the disease.

Example 18: The table given below shows the data obtained during outbreak of smallpox:

	Attacked	Not attacked
Vaccinated	31	469
Not vaccinated	185	1315

Test the effectiveness of vaccination in preventing the attack from smallpox.

Solution:

Here,

	Attacked(B)	Not attacked (β)	Total ($O_{i.}$)
Vaccinated (A)	31	469	500
Not vaccinated (α)	185	1315	1500
Total ($O_{.j}$)	216	1784	2000

Now,

Group	O_{ij}	$E_{ij} = (O_{i.} \times O_{.j})/N$	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2/E_{ij}$
AB	31	54	-23	9.796
Aβ	469	446	23	1.186
αB	185	162	23	3.265
αβ	1315	1338	-23	0.395
Total				14.642

Problem to test

H_0: Vaccination is not effective in preventing the attack from smallpox

H_1: Vaccination is effective in preventing the attack from smallpox.

Test statistic

$$\chi^2 = \Sigma\Sigma\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 14.642$$

Critical value

Let α = 0.05 be the level of significance, then for (r-1)(c-1) =1 degree of freedom at 0.05 level of significance, the critical value is $\chi^2_{0.05(1)} = 3.84$

Decision

$\chi^2 = 14.642 > \chi^2_{0.05(1)} = 3.84$, reject H_0 at 0.05 level of significance.

Conclusion

Vaccination is effective in preventing the attack from smallpox.

EXERCISE

1. What do you mean by median test? Describe the procedure of median test.
2. Describe Kolmogorov Smirnov two sample test.
3. Describe the function and procedure of Mann Whitney U test.
4. Describe function and procedure of chi square test for goodness of fit.
5. Discuss the rationale and method of chi square test for independence of attributes.
6. Differentiate between Chi square test for goodness of fit and Kolmogorov Smirnov one sample test.

7. The following are the performance score of 10 person under two software training X and Y.

X	46	45	32	42	39	48	49	30	51	34
Y	44	40	59	47	55	50	47	71	43	55

Use median test to test the effectiveness of two training. **Ans:** (0.328, insig.)

8. The length of life in kilowatt hours of some type of electronic Neon tube and Helium tube made by two manufacturers were as follows.

Tube	Length of life															
Neon	96	238	24	200	7	108	76	140	39	165	61	25	41	99		
Helium	11	125	47	20	34	101	25	68	17	59	178	30	83	75	45	28

Compare by using Median test at 5% level of significance, the median lives of the electronic tubes made by manufacturers Neon and Helium. **Ans:** (2.14, insig.)

9. The same C programming papers were marked by two teachers A and B. The final marks were recorded as follows:

Teacher A	73	89	82	43	80	73	66	45	93	36	77	60
Teacher B	88	78	91	48	85	74	77	31	78	62	76	77

Using median test at 5% level of significance to determine if the marks distributions of two teachers differ significantly. **Ans:** (0.66, insig.)

10. A quality controller wishes to determine whether there is a difference in outcome between two different tools of software I and II. The following data shows the outcome of two different tools. Can the controller conclude that a difference exists? Use median test at 5% level of significance.

Software I	24.0	16.7	22.8	19.8	18.9	
Software II	23.2	19.8	18.1	17.6	20.2	17.8

Ans: (p=0.6, insig.)

11. Life in years of two types of electric motor used in irrigation of farms are given below;

Motor M	3	6	7	4	3	5	5	3	4
Motor N	4	3	5	6	3	2	4	7	6

Is there any significant difference between two types of motor? Use Kolmogorov Smirnov test at 0.05 level of significance. **Ans:** $D_0 = 0.15$ insig.

12. Amount of time required to design website by software developers A and B are found as follows;

Time (hrs)	0 – 4	4 – 8	8 – 12	12 – 16	16- 20
Number of website designed by A	2	7	12	5	4
Number of web sires designed by B	6	9	8	4	3

Does A take more time than B to design website? Use Kolmogorov Smirnov test at 5% level of significance. **Ans:** $D_0 = 0.2$ insig.

13. Two independent samples of 26 junior programming and 25senior programming smokers selected from a software company used to smoke following number of cigarettes per day

Number of cigarettes	0 - 2	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
Number of junior programmer	7	6	4	2	2	3	2
Number of senior programmer	5	4	6	3	4	2	1

Using Kolmogorov Smirnov test identify if there is any significance between junior and senior programmer? Use 0.05 level of significance. **Ans:** $D_0 = 0.14$, insig.

14. Two groups of date managers, one group consisting of trained ones, another groups not trained have the following number of correction required.

Trained	78	64	75	45	82
Untrained	110	70	53	51	

Use Mann Whitney U test to test if there is a difference between the two average number of correction of trained and untrained data manger. **Ans:** Uo = 9, insig.

15. The nicotine contents of two brands of cigarettes, measured in milligrams was found to be as follows:

Brand A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

If there any significance difference between two brands of cigarettes. Use Mann-whitney U-test. **Ans:** p=0.0729, insig.

16. A farmer wishes to determine whether there is a difference in yields between two different varieties of wheat I and II. The following data shows the production of wheat per unit area using the two varieties. Can the farmer conclude at significance level 0.01 that a difference exists?

Wheat I	15.9	15.3	16.4	14.9	15.3	16.0	14.6	15.3	14.5	16.6	16.0
Wheat II	16.4	16.8	17.1	16.9	18.0	15.6	18.1	17.2	15.4		

Use Mann Whitney U test. **Ans:** z=-2.89, sig.

17. Two independent random samples of unemployed men and women are drawn and the ages of 4 unemployed women and 5 unemployed men are recorded as follows:

Women	60	63	36	44	
Men	53	39	22	23	24

Do the data present sufficient evidence to conclude that there is a difference in the average age of unemployed men and women? Use Mann Whitney U test at $\alpha = 0.05$. (p = 0.055, insig.)

18. The following are the number of minutes it took a sample of 13 men and 12 women to complete the application form for a position:

Men	16.5	20.0	17.0	19.8	18.5	19.2	19.0	18.2	20.8	18.7	16.7	18.1	17.9
Women	18.6	17.8	18.3	16.6	20.5	16.3	19.3	18.4	19.7	18.8	19.9	17.6	

Use Mann Whitney U test at the 0.05 level of significance to test the hypothesis that the two samples come from identical population against the alternative that two populations are not identical.
Ans: $z = -0.108$, insig.

19. A die was rolled for 60 times and observed the following outcomes:

Side	1	2	3	4	5	6	Total
Number of times observed	8	9	13	7	15	8	60

Is the die fair? Test the hypothesis at 5% level of significance.
Ans: (5.2, insig.)

20. A certain chemical plant processes sea water to collect sodium chloride and magnesium. From scientific analysis, sea water is known to contain sodium chloride, magnesium and other elements in the ratio of 62:4:34. A sample of 200 tons of sea water resulted in 130 tons of sodium chloride and 6 tons of magnesium. Are these data consistent with the scientific model at 5% level of significance?
Ans: (1.025, insig)

21. A group of 150 college students were asked to indicate their most liked mobile brand among six different well known film actors viz. , A, B, C, D, E and F in order to ascertain their relative popularity. The observed frequency data were as follows:

Brrand	A	B	C	D	E	F	Total
Frequencies	24	20	32	25	28	21	150

Test at 5% level of significance whether all brands are equally popular.
Ans: (4, insig)

22. Genetic theory states that children having one parent of blood type A and other of blood type B will always be one of three types; A, AB, B and the proportion of three types will on an average be as 1:2:1. A report states that out of 300 children having one A parent and other B parent, 30% were found to be type A, 45% type AB and remainder type B. Test the hypothesis at 5% level of significance.
Ans: (4.5, insig.)

23. A publishing house got a 500 page book composed for printing. Before final printing the first draft was sent for proof reading. The proof reader detected the number of misprints X on each pages tabulated as follows. Test whether poisson distribution fits the data.

No. of misprint page	0	1	2	3	4	5
No. of pages	221	167	70	30	7	5

Ans: (13.32, sig.)

24. In 50 random sample of a manufactured mice, the number of samples containing defective mice is noted below:

No. of defective mice	0	1	2	3	4	5
Frequencies	4	13	17	12	3	1

Can the binomial distribution be a good with $p = 0.30$?
Ans: (11.11, sig.)

25. The number of telephone calls received per day over a period of 100 days is shown in the table below.

Number of calls	0	1	2	3
Number of days	45	25	20	10

Carry out chi square test to test the hypothesis that the number of calls received per day has a Poisson distribution at 5% level of significance. **Ans:** (8.01, sig)

26. The distribution of persons according to sex and blood groups are given below:

	Blood group			
Sex	O	A	B	AB
Male	100	40	45	10
Female	110	35	55	5

Is there any association between sex and blood group? **Ans:** (3.607, insig.)

27. A random sample of 200 married men, all retired was classified according to education and number of children.

Education	Number of children		
	0 – 1	2 - 3	Over 3
Elementary	14	37	32
Secondary	19	42	17
College	12	17	10

Test the hypothesis at the 0.05 level of significance, that the number of children is independent of the level of education attained by the father. **Ans:** (8.04, insig.)

28. Test whether the color of son's eyes is associated with that of the father's at 5% level of significance using the data available in the following table.

Father's eye color	Son's eye color	
	Not light	Light
Not light	230	148
Light	151	471

Ans: (133.32 sig.)

29. 88 workers of a IT company were interviewed during a sample survey for their smoking habit. Classification of respondents according to their gender and their smoking habit are found as

	Sex	
Habit	Male	Female
Smoker	40	33
Non-smoker	3	12

Do the smoking tea habit is associated with gender. **Ans:** 4.71, sig.

30. Out of a sample of 120 persons in a village, 76 were administered a new drug for preventing influenza and out of them 24 persons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza. Is the new drug effective in controlling influenza, test at 5% level of significance. **Ans:** 19.54, sig.

31. The following table gives the result of flower color and type of leaf.

Flower color	Type of leaf	
	Flat	Curled
Pink	3	22
Red	9	11

Test whether the flower color is independent of flatness of leaf or not.

Ans: 4.61, sig.

Paired Sample Test

Wilcoxon Matched Pair Signed Rank Test

It is non parametric test used to compare two populations for which observations are paired. Itis based upon magnitude and direction of difference between observations within each pair of related random samples.

Let us consider two related samples of size n drawn from continuous populations with unknown medians Md_1 and Md_2 respectively.

Let $x_1, x_2, x_3, \ldots\ldots x_n$ and $y_1, y_2, y_3, \ldots\ldots y_n$ be two related samples of size n

Different steps in the test are

Problem to test

H_0: $Md_1 = Md_2$

H1: $Md_1 \neq Md_2$ (two tailed) or H_1:$Md_1 > Md_2$ (one tailed right) or H_1: $Md_1 < Md_2$ (one tailed left)

Find $d_i = y_i - x_i$ or $x_i - y_i$ for each pair of observations(x_i, y_i) , $i = 1,2,3\ldots\ldots n$. Rank di irrespective of sign in ascending order but omit $d_i = 0$. If two or more d_i are equal then assign the average rank and is called tied. In such case corrected sample size $n_c = n - t$, t is number of tied occurred .Assign sign to the ranks with respect to the sign of d_i. Sum the ranks of + sign and - sign separately to get S(+) and S(-) respectively. Finally get T = min {S(+) , S(-)}.

Small sample size ($n \leq 25$)

Test statistic

T = min {S(+) , S(-)}.

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

At α level of significance, we obtain critical value from Wilcoxon Matched pair signed rank test table

$T_{\alpha, n}$ where n is corrected sample size after omitting $d_i = 0$.

Decision

Reject H_0 at α level of significance if $T \leq T_{\alpha, n}$ accept otherwise.

Large sample size (n> 25)

For large sample size sampling distribution of T is approximately normally distributed with mean μ_T and variance σ_T^2.

$\mu_T = \dfrac{n(n+1)}{4}$ and $\sigma_T^2 = \dfrac{n(n+1)(2n+1)}{24}$ where n is corrected sample size if $d_i = 0$.

Test statistic

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1) \qquad (\text{Here } n = n_c)$$

Level of Significance

Let α be the level of significance. Generally fix α= 0.05 unless we are given.

Critical Value

Critical value $Z_{tabulated}$ is obtained from table according to level of significance and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $Z > Z_{tabulated}$, Accept otherwise.

Example 1: The number of server manufactured from each of two companies A and B was recorded daily for a period of 10 days with the following results;

Days	1	2	3	4	5	6	7	8	9	10
A	172	165	206	184	174	142	190	169	161	200
B	201	179	159	192	177	170	182	179	169	210

Assuming both companies produced same daily output, test the hypothesis that there is no difference between median defectives for two companies against the alternative hypothesis that company B produced more defective than company A using Wilkcoxon Matched pairs signed rank test.

Solution:

Days	Company A(x_i)	Company B(y_i)	$d_i = y_i - x_i$	R_i
1	172	201	29	9
2	165	179	14	7
3	206	159	-47	-10
4	184	192	8	3
5	174	177	3	1
6	142	170	28	8
7	190	182	-8	-3
8	169	179	10	5.5
9	161	169	8	3
10	200	210	10	5.5

Here, n = 10,Sum of ranks of + sign [S(+)] = 9+7+3+1+8+5.5+3+5.5 = 42 , Sum of ranks of – sign [S(-)]= 10 + 3 = 13

Let Md_A and Md_B be median defective of company A and B respectively.

Problem to test

H_0: There is no significant difference in median defect of company A and company B($Md_A = Md_B$)

H_1: There is significant increase in median defective in company B than company A ($Md_A < Md_B$)

Test statistic

$T = \min\{S(+), S(-)\}$

$= \min\{42, 13\} = 13$

Critical value

Let $\alpha = 5\%$ be the level of significance the critical value is $T_{0.05(10)} = 8$

Decision

$T = 13 > T_{0.05(10)} = 11$

Accept H_0 at 5% level of significance.

Conclusion

There is no difference between median defectives for two companies A and B.

Example 2: The scores under two conditions X and Y obtained by the respondents are given below:

X	12	16	8	6	4	8
Y	7	12	17	5	12	11

Use Wilcoxon Matched pair signed rank test to test the difference between scores under condition X and Y.

Solution:

X	Y	d = Y - X	R
12	7	-5	-4
16	12	-4	-3
8	17	9	6
6	5	-1	-1
4	12	8	5
8	11	3	2

Here, $n = 6$,Sum of ranks of + sign [S(+)]= 6+5+2 = 13 , Sum of ranks of – sign [S(-)]= 4+3+1 = 8

Problem to test

H_0 : Distribution of X = Distribution of Y

H_1: Distribution of X $\neq$ Distribution of Y

Test statistic

$T = \min\{S(+), S(-)\} = \min\{13, 8\} = 8$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value is $T_{0.05(6)} = 0$

Decision

$T = 8 > T_{0.05(6)} = 1$

Accept H_0 at 5% level of significance.

Conclusion

There is no difference between scores under condition X and Y.

Example 3: The following data gives the additional hours of sleep gained by 31 patients in an experiment to test the effect of a drug. Do these data give evidence that the drug produces additional hours of sleep?

Sleep hours gained by patients are; 0.5 , 0.7 , 0.1 , -0.2 , 1.2 , 1.5 , -2 , 4 , 0.1 , 3.4 , 3.7, -1.1, 0.8 , - 0.8 , 1.3 , 2.7 , -3.4 , -1.9 , 3.4 ,0.1 , 0.6 , 2.3 , 0.1 , 2.7 , -0.9 , 3.1 , 2.0 , 1.2 , 1.2, -1.8 , 1.0

Solution:

Additional hours sleep (d_i)	Rank
0.5	6
0.7	8
0.1	2.5
-0.2	-5
1.2	15
1.5	18
-2	-21.5
4	31
0.1	2.5
3.4	28
3.7	30
-1.1	-13
0.8	9.5
-0.8	-9.5
1.3	17
2.7	24.5
-3.4	-28
-1.9	-20
3.4	28
0.1	2.5
0.6	7
2.3	23
0.1	2.5
2.7	24.5
-0.9	-11
3.1	26
2.0	21.5
1.2	15
1.2	15
-1.8	-19
1.0	12

Here, n = 31

Sum of - sign ranks [S(-)] = 5+21.5+13+9.5+28+20+11+19 = 127

Sum of + sign ranks [S(+)] = $S_n - S(-) = \frac{n(n+1)}{2} - S(-) = \frac{31(1+31)}{2} - 127 = 369$

T = min {S(+) , S(-) } = min { 369 , 127} = 127

Problem to test

H_0 : drug does not produce additional hours of sleep

H_1: drug produces additional hours of sleep

Test statistic

$T = 127$

$$\mu_T = \frac{n(n+1)}{4} = \frac{31(31+1)}{4} = 248$$

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24} = \frac{31(31+1)(62+1)}{24} = 2604$$

$$\Rightarrow \sigma_T = 51.029$$

Hence, $Z = \frac{T - \mu_T}{\sigma_T} = \frac{127 - 248}{51.029} = -2.37$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value is $Z_{0.05} = 1.96$

Decision

$|Z| = 2.37 > Z_{0.05} = 1.96$, reject H_0 at 5% level of significance.

Conclusion

Drug produces additional hours of sleep to the patients.

Cochran Q Test

It is non parametric test used for more than two related samples. It is used to test significant difference in frequencies or proportions of three or more related samples.

Let us consider k treatments (k>2) applied to the set of n objects dichotomized as Yes(Y) or No(N), Presence(P) or Absence(A), Pass(P) or Fail(F), Accept(A) or Reject (R), Increase (I) or Decrease (D) etc.

Treatments	Objects				
	1	2	3	J	N
T_1 T_2 T_3	Y N Y	N Y N	Y N Y	Y Y N	N Y N
. . .					
T_i	N	Y	N	Y	Y
T_k	Y	N	Y	N	Y

Different steps in the test are

Problem to test

H_0: All the treatments are equally effective

H_1: All the treatments are not equally effective.

Sum all the Y(Positive) according to treatment to get R_i (Row wise) and according to objects to get C_i (Column wise), $i = 1, 2, 3, \ldots\ldots k$ and $j = 1, 2, 3, \ldots\ldots n$. Then get

$$\sum_{i=1}^{k} R_i,\ \sum_{i=1}^{k} R_i^2\ \sum_{j=1}^{n} C_j,\ \sum_{j=1}^{n} C_j^2$$

Test statistic

$$Q = \frac{(k-1)\left[K\Sigma R_i^2 - (\Sigma Ri)^2\right]}{K\Sigma C_j - \Sigma C_j^2} \sim \chi^2$$

Level of significance

Let α be the level of significance .Generally fix $\alpha = 0.05$ unless we are given.

Critical value

$\chi^2_{\alpha\,(k-1)}$ is the critical value obtained according to level of significance and degree of freedom.

Decision

Reject H_0 at α% level of significance if $Q > \chi^2_{\alpha(k-1)}$, accept otherwise.

Example 4: Five laptop users were asked for the acceptability of four brands for daily use. The response of acceptability (A) and rejection (R) are given below;

Users	Brands			
	Alfa	Beta	Gamma	Delta
H_1	A	R	A	R
H_2	R	A	A	R
H_3	R	A	R	A
H_4	A	R	R	R
H_5	A	A	R	A

Test whether there is any significant difference between brands with respect to acceptability

Solution:

Lipstick Brands	Users					R_i	R_i^2
	H_1	H_2	H_3	H_4	H_5		
Alfa	A	R	R	A	A	3	9
Beta	R	A	A	R	A	3	9
Gamma	A	A	R	R	R	2	4
Delta	R	R	A	R	A	2	4
C_j	2	2	2	1	3	$\Sigma R_i = \Sigma C_j = 10$	$\Sigma R_i^2 = 26$
C_j^2	4	4	4	1	9	$\Sigma C_j^2 = 22$	

Here,

Number of brands (k) = 4 , Number of housewives (n) = 5

Problem to test

H_0 : There is no significant difference between brands

H_1: There is at least one significant difference between brands.

Test statistic

$$Q = \frac{(k-1)\left\{k\sum_{i=1}^{k} R_i^2 - \left(\sum_{i=1}^{k} R_i\right)^2\right\}}{k\sum_{j=1}^{n} C_j - \sum_{j=1}^{n} C_j^2} = \frac{(4-1)\{4 \times 26 - (10)^2\}}{4 \times 10 - 22} = \frac{12}{18} = 0.666$$

Critical value

Let 5% be the level of significance then critical value is $\chi^2_{0.05\,(3)} = 7.81$.

Decision

$Q = 0.666 < \chi^2_{0.05\,(3)} = 7.81$, accept H_0 at 5% level of significance.

Conclusion:

There is no significant difference between brands according to acceptability.

Kruskal Wallis H Test

It is also called Kruskal Wallis one way ANOVA test

It is test used to test the significant difference between location of three or more independent populations.

Let us consider k independent samples of size n_i such that $\Sigma n_i = n$ drawn from continuous population with unknown medians $Md_1, Md_2 \ldots\ldots\ldots Md_k$ respectively.

Let $x_{11}, x_{12}, x_{13}, \ldots\ldots x_1n_1$ be samples of size n_1. $x_{21}, x_{22}, x_{23}, \ldots\ldots\ldots x_2n_2$ be sample of size n_2. $x_{31}, x_{32}, x_{33}, \ldots\ldots\ldots x_3n_3$ be sample of size n_3 .... $xk_1, xk_2, xk_3, \ldots\ldots\ldots x_kn_k$ be sample of size n_k respectively.

Samples	Blocks				
	1	2	3	j	n
1	x_{11}	x_{12}	x_{13}	$x1j$	x_1n_1
2	x_{21}	x_{22}	x_{23}	x_{2j}	x_2n_2
3	x_{31}	x_{32}	x_{33}	x_{3j}	x_3n_3
.					
.					
i	x_{i1}	x_{i2}	x_{i3}	x_{ij} x_in_i	
.					
.					
k	x_{k1}	x_{k2}	$x_{k3}x_{kj}$	x_kn_k	

Different steps in the test are

Problem to test

H_0: $Md_1 = Md_2 = Md_3 = \ldots\ldots\ldots\ldots Md_k$

H_1: At least one Md_i is different $i = 1, 2, 3, \ldots\ldots\ldots k$.

Combine $n_1, n_2, n_3 \ldots\ldots\ldots$ and n_k such that $n_1 + n_2 + n_3 + \ldots\ldots\ldots + nk = n$ and rank these n observations in ascending order .If two or more observations are equal then assign average rank and is called tied. Sum the ranks of sample of sizes $n_1, n_2, n_3, \ldots\ldots\ldots$ and n_k separately to get $R_1, R_2, R_3, \ldots\ldots\ldots R_k$.

Test statistic

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) \sim \chi^2(k-1)$$

If tied occurs then corrected test statistic is

$$H = \frac{\frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n^3 - n}}$$, t_i = number of times i^{th} rank is repeated.

Level of significance

Let α be the level of significance .Generally fix $\alpha = 0.05$ unless we are given.

Critical value

For ni ≤ 5 and $k = 3$, critical value p is obtained from Kruskal Wallis table.

For $n_i > 5$ and $k > 3$, critical value is $\chi^2_{\alpha\,(k-1)}$.

Decision

Accept H_0 at α level of significance if $p > \alpha$, reject otherwise for $n_i \leq 5$ and k=3.

Reject H0 at α level of significance if $H > \chi^2\alpha\ (k-1)$, accept otherwise for $n_i > 5$ and $k > 3$.

Example 5: A bacteriologist was interested to study the number of plankton organism inhabiting the lake water. He made hauls of water from three lakes each and the following results were obtained.

Lake	Number of plankton organism				
Phewa	12	19	16		
Rara	4	8	3	2	3
Taudaha	14	12	20	12	

Do the data provide substantial evidence to conclude significant variation between lake water? Use Kruskal Wallis test at 0.05 level of significance.

Solution:

Lake	Number of plankton organism					R_i	R_i^2/n_i
Phewa	12	19	16				
Rank	7	11	10			28	261.33
Rara	4	8	3	2	3		
Rank	4	5	2.5	1	2.5	15	45
Taudaha	14	12	20	12			
Rank	9	7	12	7		35	306.25
							612.58

Sample size of Phewa (n_1) = 3

Sample size of Rara (n_2) = 5

Sample size of Taudha (n_3) = 4

Total sample size (n) = $n_1 + n_2 + n_3 = 3 + 5 + 4 = 12$

No of times rank 2.5 is repeated (t_1) = 2

No of times rank 7 is repeated (t_2) =3

$\Sigma (t_i^3 - t_i) = (2^3-2)+(3^3-3) = 6+24 = 30$

R_1, R_2, R_3 be sum of ranks for Phewa , Rara and Taudaha respectively.

Let Md_1, Md_2 and Md_3 be median of Phewa, Rara and Taudaha lake respectively.

Problem to test

H_0: There is no significant variation between lake water.($Md_1 = Md_2 = Md_3$)

H_1: There is at least one significant variation between lake water. (At least one Mdi is different, i = 1 , 2 , 3)

Test statistic

$$H = \frac{\frac{12}{n(n+1)} \Sigma \frac{R_i^2}{n_i} - 3(n+1)}{1 - \Sigma \frac{(t_i^3 - t_i)}{n^3 - n}} = \frac{\frac{12}{12\,(12+1)} \times 612.5 - 3(12+1)}{1 - \frac{30}{12^3 - 12}} = \frac{8.12}{0.98} = 8.285$$

Level of significance

$\alpha = 0.05$ is the level of significance.

Critical value

From Kruskal Wallis table critical value is p = 0.01

Decision

$P = 0.01 < \alpha = 0.05$, reject H_0 at 0.05 level of significance.

Conclusion

There is at least one significant variation between lake water.

Example 6: The following are the numbers of misprints counted on pages selected at random from three editions of a book

Edition I	4	10	2	6	4	12
Edition II	8	5	13	8	8	10
Edition III	7	9	11	2	14	7

Use Kruskal Wallis H test at the 0.05 level of significance to test the null hypothesis that the samples come from identical populations.

Solution:

Date	No. of misprints						R_i	R_i^2/n_i
April 11	4	10	2	6	4	12		
Rank	3.5	13.5	1.5	6	3.5	16	44	322.66
April 18	8	5	13	8	8	10		
Rank	10	5	17	10	10	13.5	65.5	715.041
April 25	7	9.	11	2	14	7		
Rank	7.5	12	15	1.5	18	7.5	61.5	630.375
Total								1668.076

Sample size of April 11 $(n_1) = 6$

Sample size of April 18 $(n_2) = 6$

Sample size of April 25 $(n_3) = 6$

Total sample size (n) $n = n_1 + n_2 + n_3 = 6 + 6 + 6 = 18$

No of times rank 1.5 is repeated $(t_1) = 2$, No of times rank 3.5 is repeated $(t_2) = 2$

No of times rank 7.5 is repeated $(t_3) = 2$, No of times rank 10 is repeated $(t_4) = 3$

No of times rank 13.5 is repeated $(t_5) = 2$

$\Sigma(t_i^3 - t_i) = (2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2) = 48$

R_1, R_2, R_3 be sum of ranks for April 11, April 18 and April 25 respectively.

Let Md_1, Md_2 and Md_3 be median misprint of April 11, April 18 and April 25 respectively.

Problem to test

H_0: Populations are identical. $(Md_1 = Md_2 = Md_3)$

H_1: Populations are not identical.(At least one Md_i is different, $i = 1, 2, 3$)

Test statistic

$$H = \frac{\frac{12}{n(n+1)}\Sigma\frac{R_i^2}{n_i} - 3(n+1)}{1 - \Sigma\frac{(t_i^3 - t_i)}{n^3 - n}}$$

$$= \frac{\frac{12}{18(18+1)} \times 1668.076 - 3(18+1)}{1 - \frac{48}{18^3 - 18}}$$

$$= \frac{58.528 - 57}{0.9917} = 1.5407$$

Critical value

Critical value at 0.05 level of significance for 2 degree of freedom is $\chi^2_{0.05\,(2)} = 5.99$.

Decision

$H = 1.54 < \chi^2_{0.05\,(2)} = 5.99$, accept H_0 at 0.05 level of significance.

Conclusion

The samples come from identical population.

Friedman F Test

It is also called Friedman two way ANOVA test.

It is test used to test the significant difference between location of three or more independent populations.

Let us consider k independent samples of size n each.

Let x_{ij} be the observations classified into k rows (treatments) and n columns(blocks) such that there are $N = n \times k$ observations.

Samples	Blocks				
	1	2	3	j	n
1	x_{11}	x_{12}	x_{13}	x_{1j}	x_{1n}
2	x_{21}	x_{22}	x_{23}	x_{2j}	x_{2n}
3	x_{31}	x_{32}	x_{33}	x_{3j}	x_{3n}
.					
.					
i	x_{i1}	x_{i2}	$x_{i3} x_{ij}$ x_{in}		
.					
.					
k	x_{k1}	x_{k2}	$x_{k3} x_{kj}$ x_{kn}		

Different steps in the test are

Problem to test

H_0: $Md_1 = Md_2 = Md_3 = \ldots\ldots\ldots\ldots\ldots Md_k$

H_1: At least one Md_i is different $i = 1, 2, 3, \ldots\ldots\ldots\ldots k$.

Rank k sample observations for each block separately from 1 to k. in ascending order. If two or more observations are same then assign average rank which is also called tied. Obtain sum of ranks for each sample to get R_i, $i = 1, 2, 3, \ldots\ldots\ldots k$

Test statistic

$$F_r = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

If tied occurs then corrected test statistic is

$$F_r = \frac{\dfrac{12}{nk(k+1)} \sum\limits_{i=1}^{k} R_j^2 - 3n(k+1)}{1 - \sum \dfrac{(t_i^3 - t_i)}{n(k^3 - k)}}$$ t_i = number of times ith rank is repeated.

Level of significance

Let α be the level of significance .Generally fix $\alpha = 0.05$ unless we are given.

Critical value

For $2 \le n \le 9$ and $k = 3$, also $2 \le n \le 4$ and $k = 4$ critical value p is obtained from Friedman probability table.

For $n \ge 5$ and $k > 3$, critical value is $\chi^2_{\alpha\,(k-1)}$

Decision

Accept H_0 at α level of significance if $p > \alpha$, reject otherwise for $2 \le n \le 9$ and $k = 3$, also $2 \le n \le 4$ and $k = 4$.

Reject H_0 at α level of significance if $H > \chi^2\alpha\,(k-1)$, accept otherwise for $n \ge 5$ and $k > 3$.

Example 7: A survey was conducted in four hospitals in a particular city to obtain the number of babies born over a 12 months period. This time period was divided into four seasons to test the hypothesis that the birth rate is constant over all the four seasons. The results of the survey were as follows:

Hospital	No of births			
	Winter	Spring	Summer	Fall
A	92	72	94	77
B	15	16	10	17
C	58	71	51	62
D	19	26	20	18

Analyze the data using Friedman two way ANOVA test.

Solution:

Hospital	No of births in Hospital								R_i	R_i^2
	A	Rank	B	Rank	C	Rank	D	Rank		
Winter	92	3	15	2	58	2	19	2	9	81
Spring	72	1	16	3	71	4	26	4	12	144
Summer	94	4	10	1	51	1	20	3	9	81
Fall	77	2	17	4	62	3	18	1	10	100
$\Sigma R_i^2 = 406$										

Here,

Number of hospitals (n) = 4

Number of seasons (k) = 4

Let Md_1, Md_2, Md_3, Md_4 be median of winter, spring, summer and fall seasons respectively.

Problem to test

H_0: The birth rate is constant over all four seasons.($Md_1 = Md_2 = Md_3 = Md_4$)

H_1: The birth rate is not constant over all four seasons (At least one Md_i is different, i = 1, 2, 3, 4)

Test statistic

$$F_r = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

$$= \frac{12 \times 406}{4 \times 4 \times 5} - 3 \times 4 \times 5 = 0.9$$

Level of significance

Let $\alpha = 5\%$ be the level of significance

Critical value

The critical value is $p = P(F_r > 0.9) = 0.9$ for $n = 4$ and $k = 4$

Decision

$P = 0.9 > \alpha = 0.05$, accept H_0 at 5% level of significance.

Conclusion: Birth rate is constant over all four seasons.

Example 8: An investigator wants to study the scores of 3 matched groups under 5 conditions, each group contains five subjects, one being assigned to each of the five conditions. Let the score obtained are given in the following table.Is there any significant difference between three groups.

Group	Conditions				
	I	II	III	IV	V
A	9	8	5	1	7
B	6	7	5	2	8
C	9	7	5	2	6

Solution:

Ranking the different conditions separately

Group	Conditions					R_i	R_i^2
	I	II	III	IV	V		
A	2.5	3	2	1	2	10.5	110.25
B	1	1.5	2	2.5	3	10	100
C	2.5	1.5	2	2.5	1	9.5	90.25
							300.5

Here,

Number of conditions $(n) = 5$

Number of groups $(k) = 3$

Number of times rank 2.5 is repeated in I $(t_1) = 2$

Number of times rank 1.5 is repeated in II $(t_2) = 2$

Number of times rank 2 is repeated in III $(t_3) = 3$

Number of times rank 2.5 is repeated in IV $(t_4) = 2$

Let Md_A, Md_B, Md_c be median of group A, B and C respectively.

Problem to test

H_0: Scores of 3 matched groups is same ($Md_A = Md_B = Md_c$)

H_1: Scores of 3 matched groups is different (At least one Md_i is different, i= A, B, C)

Test statistic

$$F_r = \frac{\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_j^2 - 3n(k+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n(k^3 - k)}} = \frac{\frac{12 \times 300.5}{5 \times 3 \times 4} - 3 \times 5 \times 4}{1 - \frac{\{(2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2)\}}{5(3^3 - 3)}} = \frac{60.1 - 60}{1 - 0.35} = 0.153$$

Level of significance

Let $\alpha = 5\%$ be the level of significance

Critical value

Critical value is $p = P(F_r > 0.153) = 0.4$ for n=5 and k=3

Decision

$p = 0.4 > \alpha = 0.05$, accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between scores of three matched group.

Example 9: Three different advertising media T.V., Radio and News paper are being compared to study their effectiveness in promoting sales of WaiWai noodles. Each advertising media is exposed for specified period of time and sales (000 package) from 10 stores located at different areas are recorded.

Advertising Media	Stores									
	A	B	C	D	E	F	G	H	I	J
T.V.	20	21	15	12	14	17	21	16	20	18
Radio	7	9	11	12	10	10	14	12	8	7
News Paper	8	6	11	12	9	6	8	10	8	6

Are three advertising media equally effective, use Friedman two way ANOVA test.

Solution:

Ranking the sales from different stores separately

Advertising Media	Stores										R_i	R_i^2
	A	B	C	D	E	F	G	H	I	J		
T.V.	3	3	3	2	3	3	3	3	3	3	29	841
Radio	1	2	1.5	2	2	2	2	2	1.5	2	18	324
News Paper	2	1	1.5	2	1	1	1	1	1.5	1	13	169
$\Sigma R_i^2 = 1334$												

Here,

Number of stores (n) = 10

Number of advertising media (k) = 3

Number of times rank 1.5 is repeated in store C (t_1) = 2

Number of times rank 2 is repeated in store D (t_2) = 3

Number of times rank 1.5 is repeated in store I (t_3) = 2

Let Md_A, Md_B, Md_c be medians of advertising media TV, Radio and Newspaper respectively.

Problem to test

H_0: Three advertising media are equally effective ($Md_A = Md_B = Md_c$)

H_1: Three advertising media are not equally effective (At least one Md_i is different, i = A, B, C)

Test statistic

$$F_r = \frac{\frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n(k^3 - k)}}$$

$$= \frac{\frac{12 \times 1334}{10 \times 3 \times 4} - 3 \times 10 \times 4}{1 - \frac{\{(2^3-2) + (3^3-3) + (2^3-2)\}}{10(3^3-3)}} = \frac{133.4 - 120}{1 - 0.15} = 15.76$$

Critical value

Let $\alpha = 5\%$ be the level of significance then critical value is $\chi^2_{0.05(2)} = 5.99$

Decision

$F_r = 15.76 > \chi^2_{0.05(2)} = 5.99$, reject H_0 at 5% level of significance

Conclusion

Three advertising media T.V., Radio and News Paper are not equally effective in sales of WaiWai noodle.

EXERCISE

1. What is wilcoxon matched pair signed rank test? Why is it superior to sign test?
2. Discuss the rationale and method of Wilcoxon Matched pair signed rank test.
3. Describe rationale and procedure of Cochran Q test.
4. Discuss the function and procedure of Kruskal Wallis H test.
5. What do you mean by Friedman two way ANOVA test? Describe process of the test.
6. Differentiate between Kruskal Wallis One way ANOVA test and Friedman Two way ANOVA test.
7. The weight (kg) of 5 people before they stopped smoking are as follows:

Before	66	80	69	52	75
After	71	82	68	56	73

Use wilcoxon Matched pairs signed rank test for paired observations to test the hypothesis at 0.05 level of significance that giving up smoking has no effect on a person's weight against the alternative hypothesis that one's weight increases if he or she quits smoking. **Ans:** T=3.5, insig.

8. To evaluate a speed of reading course, a group of 10 subjects were asked to read two comparable articles one before the course and one after the course. Their scores on reading test are as follows;

Before course(X)	57	80	64	70	90	59	76	98	70	83
After course(Y)	60	90	62	70	95	58	80	99	75	94

Test whether the course is beneficial using the Wilcoxon Matched pairs signed rank test at 5% level of significance. **Ans:** T=4.5, sig.

9. Seven prospective graduate students took a test twice with the following scores.

First attempt	470	530	610	440	600	590	580
Second attempt	510	550	600	490	585	620	598

Test whether there is significant difference between scores in first attempt and second attempt using Wilcoxon Matched pair signed rank test. **Ans:** T=3, insig.

10. The following are the number of artifacts dug up by two archaeologists at an ancient cliff dwelling on 30 days.

Arch. X	1	0	2	3	1	0	2	2	3	0	1	1	4	1	2
Arch. Y	0	0	1	0	2	0	0	1	1	2	0	1	2	1	1
Arch . X	1	3	5	2	1	3	2	4	1	3	2	0	2	4	1
Arch. Y	0	2	2	6	0	2	3	0	2	1	0	1	0	1	5

Do these data present sufficient evidence to support the research hypothesis that archaeologist X is better than Y. Use Wilcoxon Matched pairs signed rank test. **Ans:** z = 1.86, sig.

11. Four objective questions are given to 5 students and the results of correct answer(1) and wrong answers(0) are arranged in the following table.

Objective Questions	Students				
	1	2	3	4	5
Q_1	1	0	0	1	1
Q_2	0	1	1	0	1
Q_3	1	1	1	0	0
Q_4	0	0	1	0	1

Apply Cochran q test for testing the hypothesis that there is no significant difference between four objective questions with respect to the correct answers. **Ans:** Q=0.529, insig.

12. Three diets P, Q and R are fed to 9 buffaloes, each diet for one month and the result of Increasing (I) and decreasing (D) of milk given by different buffaloes are given in the following table.

Diet	Buffaloes								
	1	2	3	4	5	6	7	8	9
P	I	D	I	D	I	I	D	D	I
Q	I	I	I	D	I	I	D	I	D
R	I	D	D	D	I	D	I	I	I

Test whether the three diets are equally effective or not using Cochran Q test at $\alpha = 5\%$. **Ans:** Q=0.33, insig.

13. Following are the final examination marks of three group of students who were taught computer by three different methods;

Method I	94	88	91	74	87	97	
Method II	85	82	79	84	61	72	80
Method III	89	67	72	76	69		

Are all three methods equally effective? Use H test at 0.05 level of significance. **Ans:** H=6.67, sig.

14. An agricultural experiment was conducted to compare the yield of wheat by using three types of chemical fertilizer nitrogen(N), phosphorous (P) and potash (K). Twelve plots of equal size were selected at random and divided into three groups of four each and planted wheat. Each group was randomly selected and the fertilizer was applied in the plots under the identical conditions. Then the yield of wheat recorded were given in the following table.

Chemical fertilizer		
N	P	K
122	81	80
80	80	82
138	79	65
121	65	58

Test whether the three types of fertilizers are equally effective or not. Use Kruskal Wallis test at 0.05 level of significance. **Ans:** P=0.008, insig.

15. For the following scores of 3 groups, apply Kruskal Wallis H test to test the hypothesis that the three groups are not significantly different:

Group	Scores				
A	96	128	83	61	101
B	82	124	132	135	109
C	115	149	166	147	

Ans: p=0.01, sig.

16. An experiment designed to compare three preventive methods against corrosion yielded the following maximum depths of pits (in thousandths of an inch)in pieces of wire subjected to the respective treatments:

Method I	77	54	67	74	71	66	
Method II	60	41	59	65	62	64	52
Method III	49	52	69	47	56		

Use the 0.05 level of significance to test the hypothesis that the three samples come from identical population. Use Kruskal Wallis H test.

Ans: H=6.66, sig.

17. The following data represent the operating times in hours for 3 types of scientific pocket calculators before a charge is required.

Calculator A	4.9	6.1	4.3	4.6	5.3		
Calculator B	5.5	5.4	6.2	5.8	5.5	5.2	4.8
Calculator C	6.4	6.8	5.6	6.5	6.3	6.6	

Use Kruskal Wallis test, at the 0.0 level of significance to test the hypothesis that the operating times for all three calculators are equal.

Ans: H=10.4, sig.

18. A researcher wants to compare the teaching standard of three English medium schools on the basis of performance of the student's final examination scores. The percentage of passers in I to IV grade in the schools are presented in the following table.

	Grade			
	I	II	III	IV
Alpha	89	98	70	80
Sigma	45	76	40	55
Gamma	20	58	35	67

Test the performances of the schools with respect to pass percentage using Friedman's test.

Ans: p=0.042, sig.

19. A survey was conducted in four Hospitals in a particular city to obtain the number of babies born over a 12 months period divided into four seasons to test the hypothesis that the birth rate is constant over all four seasons. The result of the survey are as follows:

Hospitals	Number of births			
	Winter	Spring	Summer	Fall
A	92	112	94	77
B	9	11	10	15
C	58	71	51	62
D	19	26	19	18

Analyze the data using Friedman's test.

Ans: p = 0.928, insig.

20. The scores of 3 matched groups under the six conditions are given below

Group	Condition					
	I	II	III	IV	V	VI
A	9	5	2	5	6	7
B	6	4	3	4	6	5
C	5	1	3	3	6	5

Apply the Friedman two way ANOVA test to identify if there is significantly difference in variation between matched groups. Use 5% level of significance.

Ans: 0.184, insig.

Using Software for Non parametric test:

Run test

In 30 toss of a coin the following sequence of heads (H) and tails (T) is obtained.

H T T H T H HH T H H TT H T H T H H T H T T H T H H H T H T

Test at 0.05 level of significance level whether the sequence is random.

Using Excel

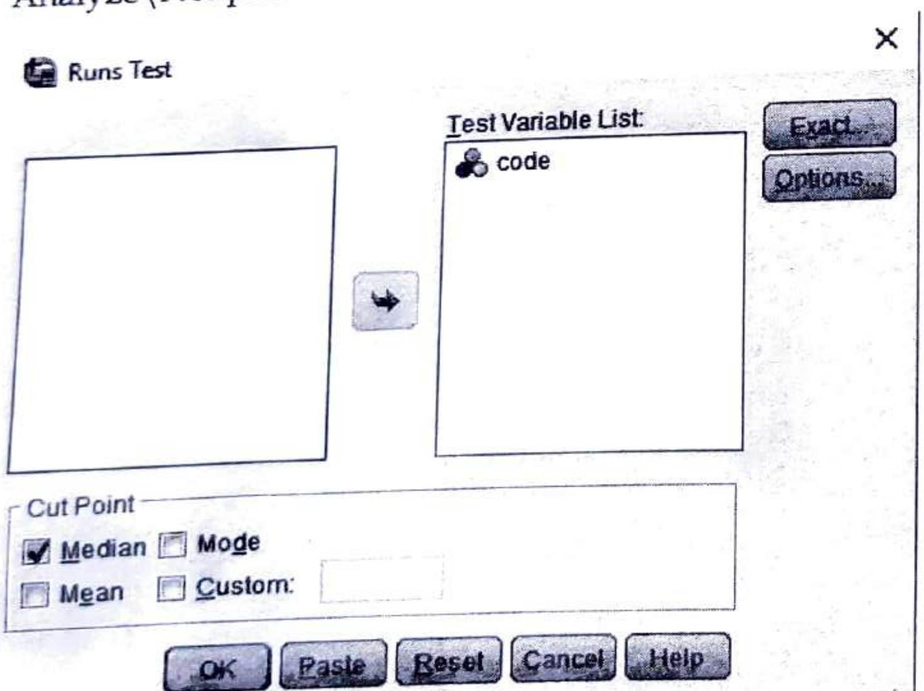
	A	B	C
1	outcome	code	runs
2	H	1	1
3	T	0	2
4	T	0	2
5	H	1	3
6	T	0	4
7	H	1	5
8	H	1	5
9	H	1	5
10	T	0	6
11	H	1	7
12	H	1	7
13	T	0	8
14	T	0	8
15	H	1	9
16	T	0	10
17	H	1	11
18	T	0	12
19	H	1	13
20	H	1	13
21	T	0	14
22	H	1	15
23	T	0	16
24	T	0	16
25	H	1	17
26	T	0	18
27	H	1	19
28	H	1	19
29	T	0	20
30	H	1	21
31	T	0	22
32			

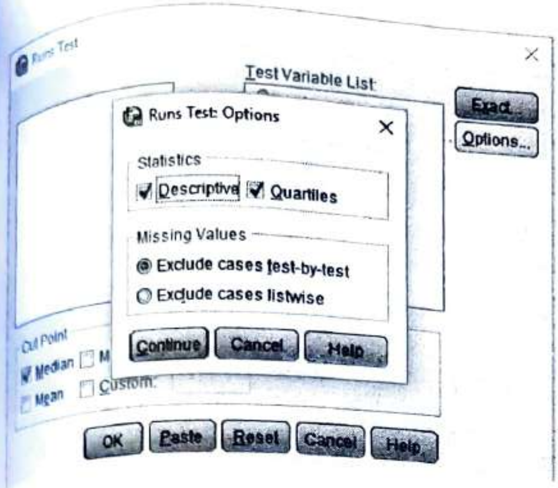
	A	B	C	D
34	cases	symbol	value	formula
35	no of observation	n	30	=COUNT(B2:B31)
36	no of head	n1	16	=COUNTIF(B2:B31,1)
37	no of tail	n2	14	=COUNTIF(B2:B31,0)
38	no of runs	R	22	=MAX(C2:C31)
39	mean	E(R)	15.93333	=((2*C36*C37)/(C36+C37))+1
40	$\mu_r = \frac{2n_1n_2}{n_1+n_2}+1$	$\sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2\,(n_1+n_2-1)}$		
41				
42				
43	variance	V(R)	7.174866	=(2*C36*C37*(2*C36*C37-C36-C37))/(((C36+C37)^2)*(C36+C37-1))
44				
45	test hypothesis			
46	Null Hypothesis	H0: sample observation is random.		
47	Alternative hypothesis	H1: sample observation isnot random.		
48				
49	test statistic			
50	z statistic	z	2.26487	=(C38-C39)/SQRT(C43)
51	level of significance	α	0.05	
52	tabulated z for two tailed test	$z_{\alpha/2}$	1.959964	=NORMSINV(1-C51/2)
53	p value	p	0.01176	=1-NORMSDIST(C50)

	A	B	C	D
54				
55	Decision:			
56	*Significant approach*			=IF(ABS(C50)<C52,"There is no reason to reject Null Hypothesis H0","H0 is rejected")
57	H0 is rejected			
58	*p-value approach*			=IF(C51<C53,"It is insignificant","It is significant")
59	It is significant			

Using SPSS

Analyze\Nonparametric tests\Legacy Dialogs\Runs





Outputs

NPar Tests

activate

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
code	30	.53	.507	0	1	.00	1.00	1.00

Runs Test

	code
Test Value[a]	1
Cases < Test Value	14
Cases >= Test Value	16
Total Cases	30
Number of Runs	22
Z	2.078
Asymp. Sig. (2-tailed)	.038

a. Median

Using STATA

Run test using stata

Recode the data such that the head = 1 and tail = 2 in variable n1code.

Then use syntax

Runtest n1code

Output will be as follows;

```
. runtest n1code
N(n1code <= 1) = 16
N(n1code >  1) = 14
           obs = 30
       N(runs) = 22
             z = 2.26
      Prob>|z| = .02
```

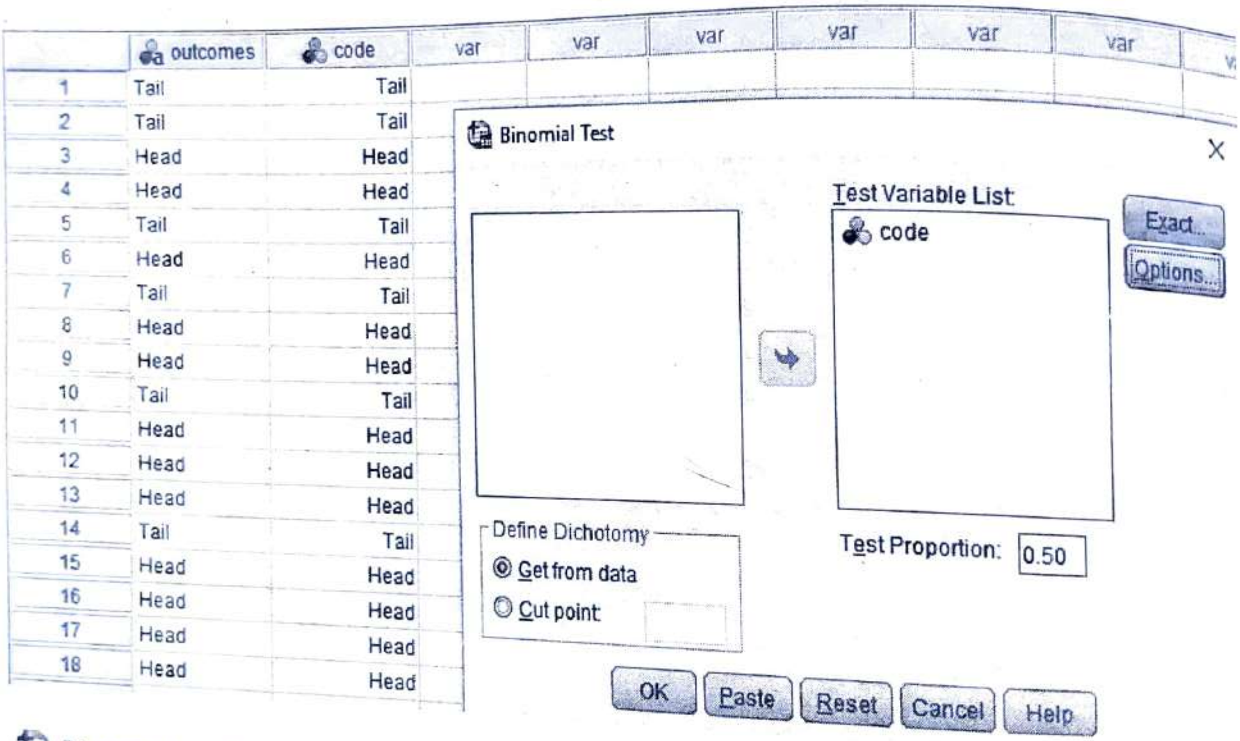
Binomial test

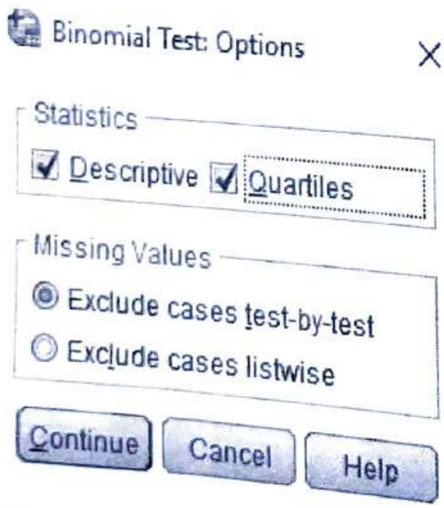
Test whether the coin is unbiased from following observations.

Tail	Tail	Head	Head	Tail	Head	Tail	Head	Head	Tail
Head	Head	Head	Tail	Head	Head	Head	Head	Head	Head
Tail	Tail	Tail	Tail	Head	Tail	Tail	Tail	Tail	Tail
Tail	Tail	Head	Tail	Tail	Tail	Head	Tail	Tail	Tail
Tail	Head	Tail	Tail	Head	Tail	Head	Tail	Tail	Tail

Using SPSS

Analyze\Nonparametric tests\Legacy Dialogs\Binomial





Outputs

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
code	50	.40	.495	0	1	.00	.00	1.00

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
code	Group 1	0 Tail	30	.60	.50	.203
	Group 2	1 Head	20	.40		
	Total		50	1.00		

Since p=0.203> 0.05, the null hypothesis of equality of proportion of head and tail can be accepted.

Using STATA

bitest code == 0.5

```
. bitest code == 0.5

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        code |       50         20           25      0.50000      0.40000

Pr(k >= 20)            = 0.940540  (one-sided test)
Pr(k <= 20)            = 0.101319  (one-sided test)
Pr(k <= 20 or k >= 30) = 0.202639  (two-sided test)
```

Since p=0.203> 0.05, the null hypothesis of equality of proportion of head and tail can be accepted.

Kolmogorov Smirnov one sample

The number of disease infected tomato plants in 10 different plots of equal size are given below. Test whether the disease infected plants are uniformly distributed over the entire area use Kolmogorov Smirnov test.

Plot no.	1	2	3	4	5	6	7	8	9	10
No of infected plants	8	10	9	12	15	7	5	12	13	9

Using excel

	A	B	C	D	E	F	G	H
4	Plot no.	No of infected plants						
5	1	8						
6	2	10						
7	3	9						
8	4	12						
9	5	15						
10	6	7						
11	7	5						
12	8	12						
13	9	13						
14	10	9						
15								
16								
17	Problem to test							
18	H0: The disease infected plants are uniformly distributed over the entire area.							
19	H1: The disease infected plants are not uniformly distributed over the entire area.							

	A	B	C	D	E	F	G	H
21	Plot no.	No of infected plants	Observed cf (cf0)	Observed relative freq (F0)	Expected freq (fe)	Expected cf (cfe)	Expected relative freq(Fe)	$\lvert F_e - F_0 \rvert$ difference
22	1	8	8	0.08	10	10	0.1	0.02
23	2	10	18	0.18	10	20	0.2	0.02
24	3	9	27	0.27	10	30	0.3	0.03
25	4	12	39	0.39	10	40	0.4	0.01
26	5	15	54	0.54	10	50	0.5	0.04
27	6	7	61	0.61	10	60	0.6	0.01
28	7	5	66	0.66	10	70	0.7	0.04
29	8	12	78	0.78	10	80	0.8	0.02
30	9	13	91	0.91	10	90	0.9	0.01
31	10	9	100	1.00	10	100	1	0
32								
33	*Test statistic*		value	formula				
34		D_0 =	0.04	=MAX(H22:H31)				
35		n	=COUNT(B22:B31)					
36		α	0.05					
37	Critical value							
38	Let 5% be the level of significance then critical value is							
39		$D_{(n,\alpha)}$	0.40925	from Ks test D table				

	A	B	C	D	E
41	decision				
42					
43	*Significant approach*				
44	There is no reason to reject Null Hypothesis H0				

use Kolmogorov Smirnov test.

Test for normal distribution

weight									
19.5	20	26.9	27.1	28.1	30	31.6	32.7	34.4	37.2
37.5	37.9	38	38.4	38.6	38.8	38.9	40.1	41.6	42.6
42.9	45	45.2	45.5	46.5	46.8	47.3	48.1	48.3	48.4
48.8	49	49.1	49.3	49.4	49.5	49.9	50.4	51.8	54.4
54.9	55.3	55.6	57.3	57.4	57.5	58.7	58.8	59.7	59.9

Using excel

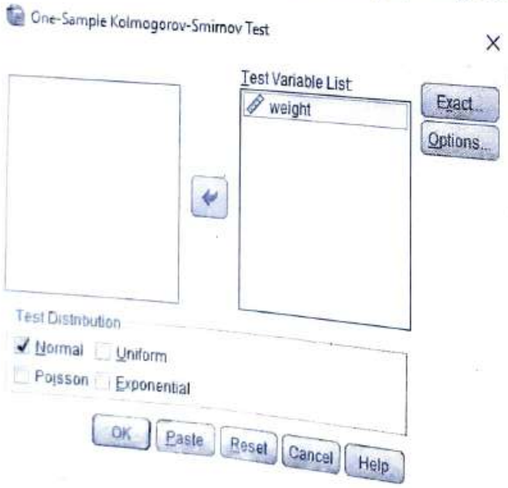
	A	B	C	D	E	F
1	weight	rank	expected	(rank-1)/n	actual	difference
2	19.5	1	0.02	0	0.007859	0.007859
3	20	2	0.04	0.02	0.008969	0.011031
4	26.9	3	0.06	0.04	0.044763	0.004763
5	27.1	4	0.08	0.06	0.046624	0.013376
6	28.1	5	0.1	0.08	0.056876	0.023124
7	30	6	0.12	0.1	0.081156	0.018844
8	31.6	7	0.14	0.12	0.10708	0.01292
9	32.7	8	0.16	0.14	0.128072	0.011928
10	34.4	9	0.18	0.16	0.165842	0.005842
11	37.2	10	0.2	0.18	0.242198	0.062198
12	37.5	11	0.22	0.2	0.251377	0.051377
13	37.9	12	0.24	0.22	0.263896	0.043896
14	38	13	0.26	0.24	0.267074	0.027074
15	38.4	14	0.28	0.26	0.279979	0.019979
16	38.6	15	0.3	0.28	0.286543	0.006543
17	38.8	16	0.32	0.3	0.293179	0.006821
18	38.9	17	0.34	0.32	0.296524	0.023476
19	40.1	18	0.36	0.34	0.337946	0.002054
20	41.6	19	0.38	0.36	0.392564	0.032564
21	42.6	20	0.4	0.38	0.430271	0.050271
22	42.9	21	0.42	0.4	0.441725	0.041725
23	45	22	0.44	0.42	0.522732	0.102732
24	45.2	23	0.46	0.44	0.53045	0.09045
25	45.5	24	0.48	0.46	0.542006	0.082006
26	46.5	25	0.5	0.48	0.580216	0.100216

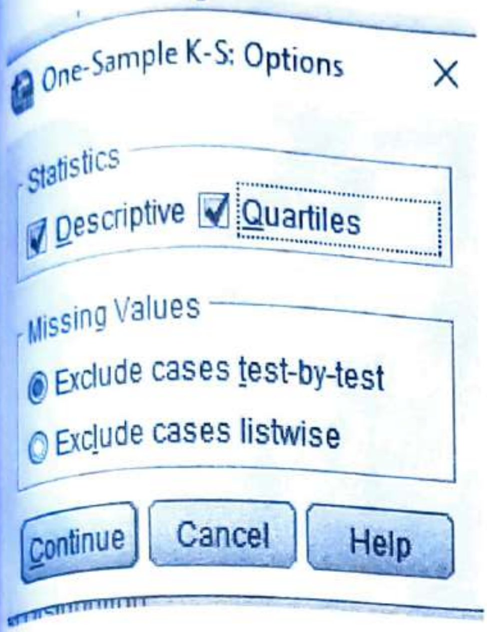
[part of data is shown]

	A	B	C	D	E	F	G
55	Problem to test						
56	H0: weights are normally distributed over the entire area.						
57	H1: weights are not normally uniformly distributed over the entire area.						
58							
59		n	50				
60		mean	44.412				
61		st dev	10.3139				
62							
63	Test Statistics						
64			value	formula			
65		D	0.102732	=MAX(F2:F51)			
66		n	50	=COUNT(A2:A51)			
67		α	0.05				
68	Critical value						
69	Let 5% be the level of significance then critical value is						
70		$D_{(n,\alpha)}$	0.18845	From table			
71							
72	decision						
73							
74	*Significant approach*						
75	There is no reason to reject Null Hypothesis H0						

Using SPSS

Analyze\Nonparametric tests\Legacy Dialogs\1 sample K-S





NPar Tests

[DataSet1]

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
weight	50	44.4120	10.31390	19.50	59.90	37.9750	46.6500	50.7500

One-Sample Kolmogorov-Smirnov Test

		weight
N		50
Normal Parameters[a,b]	Mean	44.4120
	Std. Deviation	10.31390
Most Extreme Differences	Absolute	.103
	Positive	.067
	Negative	-.103
Test Statistic		.103
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

Mann- Whitney U test

Small sample

Test the hypothesis of no difference between the ages of male and female employees of a certain company, using the Mann- Whitney U test for the sample data below. Use $\alpha = 0.1$

Male	35	43	26	44	40	42	33	38	25	26
Female	30	41	34	31	36	32	25	47	28	24

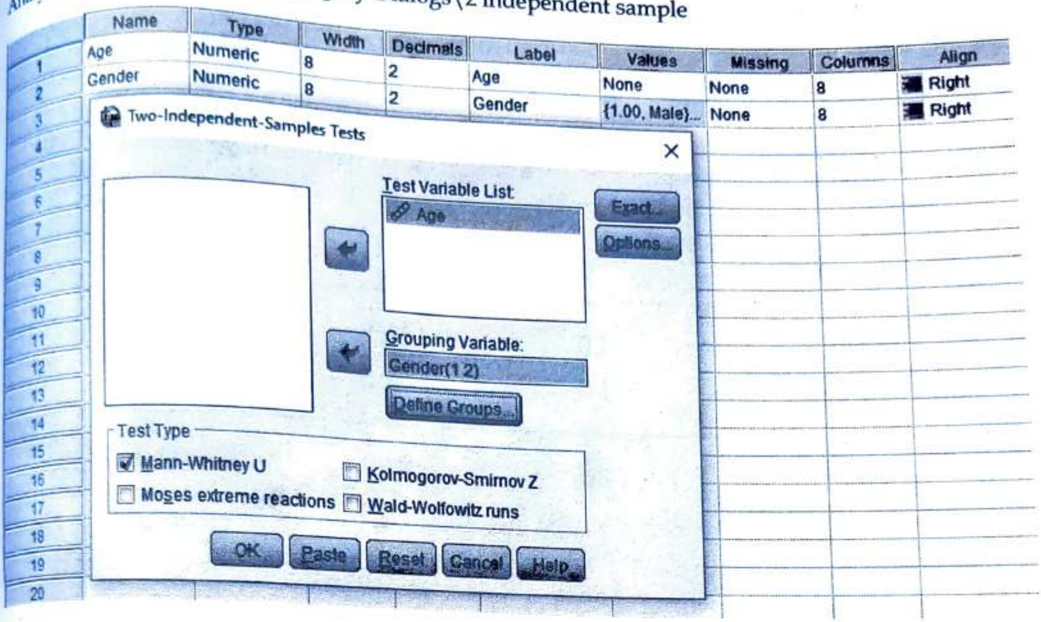
Using excel

	A	B	C
4			
5	Age	Gender	rank
6	24	2	1
7	25	1	2.5
8	25	2	2.5
9	26	1	4.5
10	26	1	4.5
11	28	2	6
12	30	2	7
13	31	2	8
14	32	2	9
15	33	1	10
16	34	2	11
17	35	1	12
18	36	2	13
19	38	1	14
20	40	1	15
21	41	2	16
22	42	1	17
23	43	1	18
24	44	1	19
25	47	2	20
26			

	A	B	C	D
27		symbol	value	formula
28	Sample size of male	n_1	10	=COUNTIF(B6:B25,1)
29	Sample size of female	n_2	10	=COUNTIF(B6:B25,2)
30	Sum of ranks of male	R_1	116.5	=SUMIF(B6:B25,1,C6:C25)
31	Sum of ranks of female	R_2	93.5	=SUMIF(B6:B25,2,C6:C25)
32	level of significance	α	0.01	
33–34	U_1 $\quad n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1$		38.5	=C28*C29+(C28*(C28+1)/2)-C30
35	U_2	$n_1n_2 - U_1$	61.5	=C28*C29-C33
36				
37	U0	$\min\{U_1,U_2\}$	38.5	=MIN(C33,C35)
38				
39	Let Md_1 and Md_2 be median age of male and female employees respectively.			
40	Problem to test			
41	H0: There is no significant difference between age of male and female employee ($Md_1 = Md_2$)			
42	H_1: There is significant difference between age of male and female employee ($Md_1 \neq Md_2$)			
43				
44	Test statistic			
45	U0	38.5		
46	Critical value			
47	Let $\alpha = 0.10$ be the level of significance then critical value is			
48	$U_{\alpha\,(n_1,n_2)}$	27	From table	
49				
50	Decision			
51	$U0 = 38.5 > U_{\alpha\,(n_1,n_2)} = 27$,		accept H_0 at 0.10 level of significance.	

Using SPSS

Analyze\Nonparametric tests\Legacy Dialogs\2 independent sample



Outputs

Mann-Whitney Test

Ranks

	Gender Gender	N	Mean Rank	Sum of Ranks
Age Age	1.00 Male	10	11.65	116.50
	2.00 Female	10	9.35	93.50
	Total	20		

Test Statistics[a]

	Age Age
Mann-Whitney U	38.500
Wilcoxon W	93.500
Z	-.870
Asymp. Sig. (2-tailed)	.384
Exact Sig. [2*(1-tailed Sig.)]	.393[b]

a. Grouping Variable: Gender Gender

b. Not corrected for ties.

Using STATA

ranksum Age, by(Gender)

```
. ranksum Age, by(Gender)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      Gender |      obs    rank sum    expected
-------------+---------------------------------
        Male |       10       116.5         105
      Female |       10        93.5         105
-------------+---------------------------------
    combined |       20         210         210

unadjusted variance      175.00
adjustment for ties       -0.26
                       ----------
adjusted variance        174.74

Ho: Age(Gender==Male) = Age(Gender==Female)
             z =   0.870
    Prob > |z| =   0.3843
```

Result: Analysis result shows the p-value for Z test is 0.3843. This indicates that our null hypothesis, of no difference in male female, is accepted.

Large sample

The following are the scores which random samples of students from 2 minority groups obtained on a current event test:

Minority Group I	73	82	39	68	91	75	89	67	50	86	57	65	70
Minority Group II	51	42	36	53	88	59	49	66	25	64	18	76	74

Use Mann Whitney U test at the 0.05 level of significance to test whether or not students from the two minority groups can be expected to score equally well on the test.

Using excel

A	B	C	D	E
scores	Group	rank =RANK.AVG(A7:A32,A7:A32,1)		
73	1	18		
82	1	22		
39	1	4		
68	1	16		
91	1	26		
75	1	20		
89	1	25		
67	1	15		
50	1	7		
86	1	23		
57	1	10		
65	1	13		
70	1	17		
51	2	8		
42	2	5		
36	2	3		
53	2	9		
88	2	24		
59	2	11		
49	2	6		
66	2	14		
25	2	2		
64	2	12		
18	2	1		
76	2	21		
74	2	19		

	A	B	C	D
35		symbol	value	formula
36	Sample size of male	n_1	13	=COUNTIF(B7:B32,1)
37	Sample size of female	n_2	13	=COUNTIF(B7:B32,2)
38	Sum of ranks of male	R_1	216	=SUMIF(B7:B33,1,C7:C33)
39	Sum of ranks of female	R_2	135	=SUMIF(B7:B33,2,C7:C33)
40	level of significance	α	0.05	
41–42	U_1	$n_1n_2+\dfrac{n_1(n_1+1)}{2}-R_1$	44	=C28*C29+(C28*(C28+1)/2)-C30
43	U_2	$n_1n_2-U_1$	125	=C28*C29-C33
44				
45	U0	$\min\{U_1,U_2\}$	44	=MIN(C33,C35)
46				

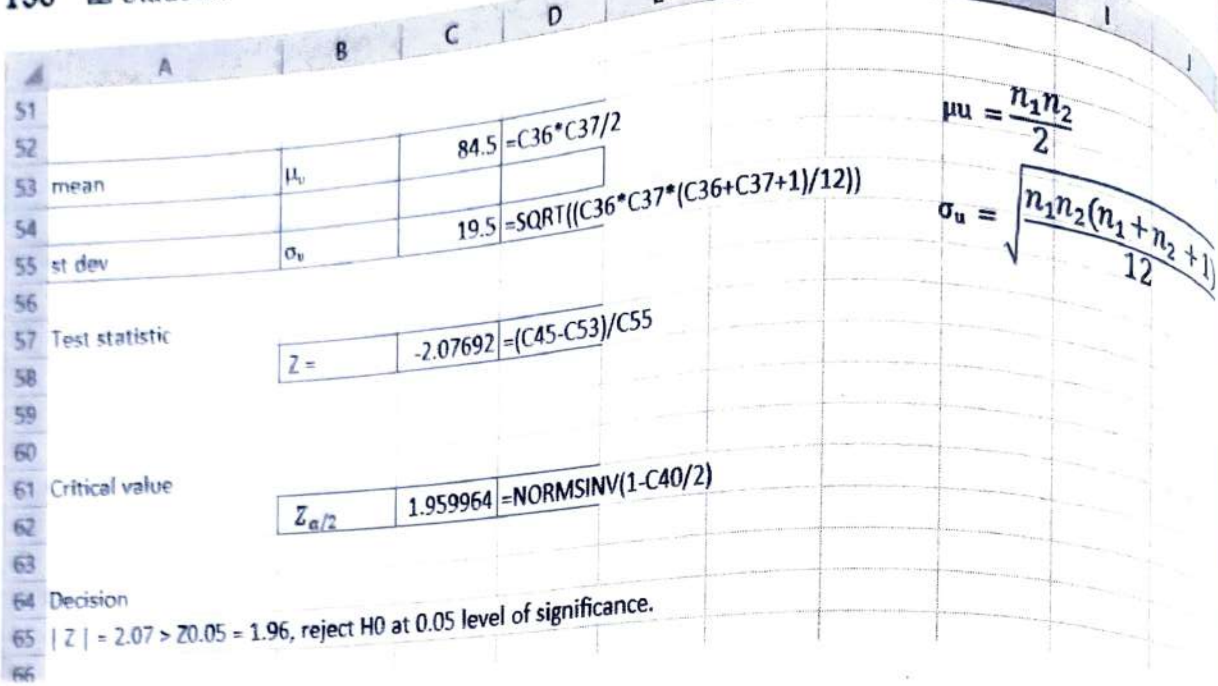
47 Let Md_1 and Md_2 be median age of male and female employees respectively.

48 Problem to test

49 H0: There is no significant difference between age of male and female employee ($Md_1 = Md_2$)

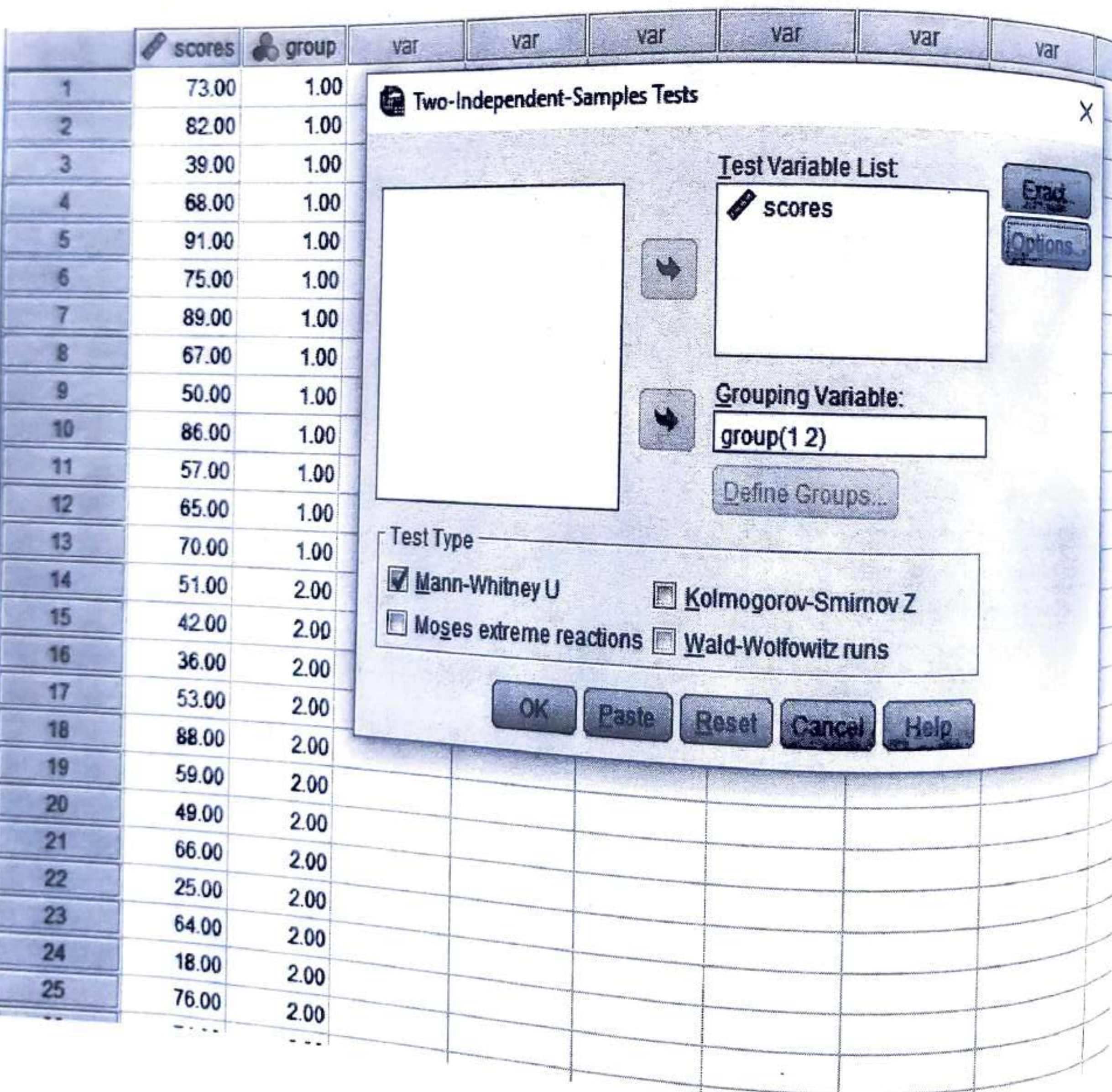
50 H_1: There is significant difference between age of male and female employee ($Md_1 \neq Md_2$)

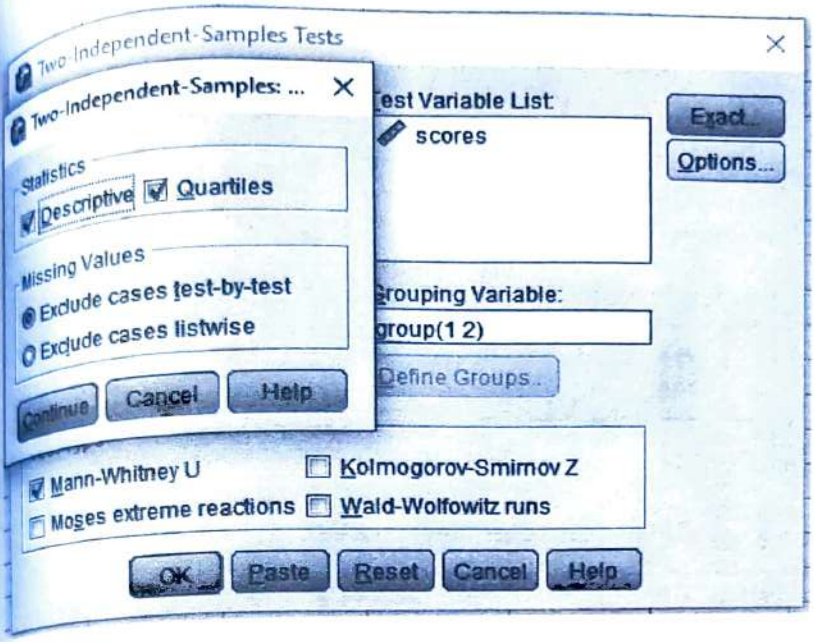
51



Using SPSS

Analyze\Nonparametric tests\Legacy Dialogs\





Outputs

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	Percentiles 50th (Median)	Percentiles 75th
scores	26	62.0385	19.45761	18.00	91.00	49.7500	65.5000	75.2500
group	26	1.5000	.50990	1.00	2.00	1.0000	1.5000	2.0000

Mann-Whitney Test

Ranks

	group	N	Mean Rank	Sum of Ranks
scores	1.00	13	16.62	216.00
	2.00	13	10.38	135.00
	Total	26		

Test Statistics[a]

	scores
Mann-Whitney U	44.000
Wilcoxon W	135.000
Z	-2.077
Asymp. Sig. (2-tailed)	.038
Exact Sig. [2*(1-tailed Sig.)]	.039[b]

a. Grouping Variable: group

b. Not corrected for ties.

Using STATA

ranksum scores, by(group)

```
. ranksum scores, by(group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
           1 |       13         216       175.5
           2 |       13         135       175.5
-------------+---------------------------------
    combined |       26         351         351

unadjusted variance        380.25
adjustment for ties          0.00
                          ----------
adjusted variance          380.25

Ho: scores(group==1) = scores(group==2)
             z =   2.077
    Prob > |z| =   0.0378
```

Result: Analysis result shows the p-value for Z test is 0.0378. This indicates that our null hypothesis, of no difference in male female, is rejected.

Use chi-square test of independence to determine association between type of place of place of residence (V102) with source of drinking water(V113) and type of toilet facility(V116) using NPBR7HFL data.

Using Excel

	A	B	C	D	E
1		Observed Frequency			
2			V102 Type of place of residence		Total
3			1 Urban	2 Rural	
4	V113 Source of drinking water	11 Piped into dwelling	958	124	1082
5		12 Piped to yard/plot	3386	2523	5909
6		13 Piped to neighbor	283	173	456
7		14 Public tap/standpipe	2730	2818	5548
8		21 Tube well or borehole	6051	3937	9988
9		31 Protected well	148	38	186
10		32 Unprotected well	269	32	301
11		41 Protected spring	254	136	390
12		42 Unprotected spring	233	209	442
13		43 River/dam/lake/ponds/stream/canal/irrigation channel	458	283	741
14		51 Rainwater	9	0	9
15		61 Tanker truck	39	0	39
16		62 Cart with small tank	3	0	3
17		71 Bottled water	265	6	271
18		97 Not a dejure resident	423	240	663
19	Total		15509	10519	26028

Expected Frequency

=RT*CT/GT

V112 Source of drinking water	V102 Type of place of residence: 1 Urban	2 Rural	Total	=((ABS(C4-C24)-0.5)^2)/C24		=((ABS(C4-C24))^2)/C24	
11 Piped into dwelling	645	437	1082	151.7439	223.7282	152.2295	224.444
12 Piped to yard/plot	3521	2388	5909	5.132327	7.566998	5.170577	7.623394
13 Piped to neighbor	272	184	456	0.428374	0.631586	0.469	0.691485
14 Public tap/standpipe	3306	2242	5548	100.125	147.6222	100.2991	147.8789
21 Tube well or borehole	5951	4037	9988	1.649068	2.431352	1.665756	2.455957
31 Protected well	111	75	186	12.13317	17.8889	12.4663	18.38006
32 Unprotected well	179	122	301	44.30988	65.32959	44.80832	66.06447
41 Protected spring	232	158	390	1.918604	2.82875	2.010543	2.964304
42 Unprotected spring	263	179	442	3.387559	4.994549	3.501921	5.163161
43 River/dam/lake/ponds/stream/canal/irrigation channel	442	299	741	0.577556	0.851537	0.614289	0.905696
51 Rainwater	5	4	9	1.835354	2.706008	2.466987	3.637275
61 Tanker truck	23	16	39	10.02278	14.77739	10.69028	15.76153
62 Cart with small tank	2	1	3	0.283932	0.418623	0.822329	1.212425
71 Bottled water	161	110	271	65.72809	96.90816	66.36764	97.8511
97 Not a dejure resident	395	268	663	1.906776	2.811312	1.976883	2.914676
Total	15509	10519	26028				

Here	Symbol	Value	Formula
No of rows	r	15	=COUNT(C24:C38)
No of columns	c	2	=COUNT(C24:D24)
No of observation	N=rc	30	=COUNT(C24:D38)
Level of significance	α	0.05	
Degrees of freedom	(r-1)*(c-1)	14	=(D42-1)*(D43-1)

Test Hypothesis:

H0: Type of residence and source of Drinking water are independent

H1: Type of residence and source of Drinking water are not independent

Test Statistics

$$\chi^2 = \frac{\left\{|O_i - E_i| - \frac{1}{2}\right\}^2}{E_i}$$

for Yates corrected cases

$$\chi^2 = \frac{\{|O_i - E_i|\}^2}{E_i}$$

for no Yates correction

	Yates corrected	Formula	No Yates correction	Formula
calculated chi square	992.6774914	=SUM(H24:I38)	1003.508	=SUM(K24:L38)
			2.8E-205	=CHISQ.TEST(C4:D18,C24:D38)
p value	5.8268E-203	=CHISQ.DIST.RT(C56,D46)	2.8E-205	=CHISQ.DIST.RT(F56,D46)

Tabulated Chi square	23.68479
decision:	
significant approach	=IF(F56<F60,"there is no reason to reject null hypothesis H0","Reject H0")
Reject H0	
p value approach	=IF(F59>D45,"it is insignificant","it is significant")
It is significant	

Using SPSS

Crosstab

Count		V102 Type of place of residence		
		1 Urban	2 Rural	Total
V113 Source of drinking water	11 Piped into dwelling	958	124	1082
	12 Piped to yard/plot	3386	2523	5909
	13 Piped to neighbor	283	173	456
	14 Public tap/standpipe	2730	2818	5548
	21 Tube well or borehole	6051	3937	9988
	31 Protected well	148	38	186
	32 Unprotected well	269	32	301
	41 Protected spring	254	136	390
	42 Unprotected spring	233	209	442
	43 River/dam/lake/ponds/stream/canal/irrigation channel	458	283	741
	51 Rainwater	9	0	9
	61 Tanker truck	39	0	39
	62 Cart with small tank	3	0	3
	71 Bottled water	265	6	271
	97 Not a dejure resident	423	240	663
Total		15509	10519	26028

Chi Square test of type of residence vs source of drinking water

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1003.508[a]	14	.000
Likelihood Ratio	1174.554	14	.000
Linear-by-Linear Association	90.020	1	.000
N of Valid Cases	26028		

a. 3 cells (10.0%) have expected count le Double-click to activate
minimum expected count is 1.21.

V102 Type of place of residence * V116 Type of toilet facility

Crosstab

Count

		V102 Type of place of residence		
		1 Urban	2 Rural	Total
V116 Type of toilet facility	11 Flush to piped sewer system	595	30	625
	12 Flush to septic tank	8919	4766	13685
	13 Flush to pit latrine	2226	2094	4320
	14 Flush to somewhere else	44	21	65
	15 Flush, don't know where	3	0	3
	21 Ventilated Improved Pit latrine (VIP)	569	663	1232
	22 Pit latrine with slab	479	533	1012
	23 Pit latrine without slab/open pit	168	138	306
	31 No facility/bush/field	1817	1870	3687
	41 Composting toilet	259	162	421
	96 Other	7	2	9
	97 Not a dejure resident	423	240	663
Total		15509	10519	26028

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	954.042[a]	11	.000
Likelihood Ratio	1050.446	11	.000
Linear-by-Linear Association	45.574	1	.000
N of Valid Cases	26028		

a. 3 cells (12.5%) have expected count less than 5. The minimum expected count is 1.21.

Using STATA

Chi-square type of place of residence (V102) vs Source of drinking water (V113)

tabulate V113 V102, chi2

```
. tabulate V113 V102,chi2
```

Source of drinking water	Type of place of residence Urban	Rural	Total
Piped into dwelling	958	124	1,082
Piped to yard/plot	3,386	2,523	5,909
Piped to neighbor	283	173	456
Public tap/standpipe	2,730	2,818	5,548
Tube well or borehole	6,051	3,937	9,988
Protected well	148	38	186
Unprotected well	269	32	301
Protected spring	254	136	390
Unprotected spring	233	209	442
River/dam/lake/ponds/	458	283	741
Rainwater	9	0	9
Tanker truck	39	0	39
Cart with small tank	3	0	3
Bottled water	265	6	271
Not a dejure resident	423	240	663
Total	15,509	10,519	26,028

Pearson chi2(14) = 1.0e+03 Pr = 0.000

Chi-square type of place of residence (V102) vs Type of toilet facility (V116)

tab V116 V102, chi2

```
. tab V116 V102, chi2
```

Type of toilet facility	Type of place of residence Urban	Rural	Total
Flush to piped sewer	595	30	625
Flush to septic tank	8,919	4,766	13,685
Flush to pit latrine	2,226	2,094	4,320
Flush to somewhere el	44	21	65
Flush, don't know whe	3	0	3
Ventilated Improved P	569	663	1,232
Pit latrine with slab	479	533	1,012
Pit latrine without s	168	138	306
No facility/bush/fiel	1,817	1,870	3,687
Composting toilet	259	162	421
Other	7	2	9
Not a dejure resident	423	240	663
Total	15,509	10,519	26,028

Pearson chi2(11) = 954.0416 Pr = 0.000

Median Test

For Small sample size ($n_1 \leq 10, n_2 \leq 10$)

Data below shows one week growth(in cm) of maize plant from two different localities (Sample I &Sampel II);

Sample I	10	11	8	8	14		
Sample II	9	12	13	9	15	9	17

Test whether the two samples have come from the same population with respect to their medians. Use median test at 0.05 level of significance.

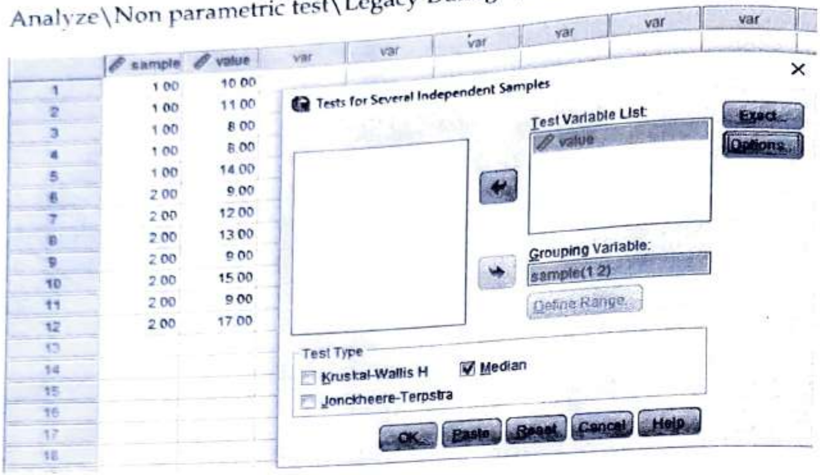
Using Excel

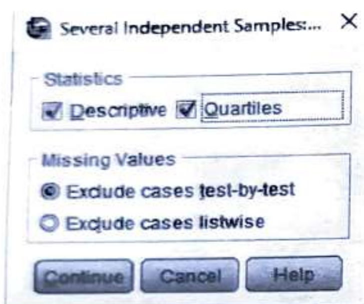
	A	B	C	D	E	F	G	H
1	Sample I	10	11	8	8	14		
2	Sample II	9	12	13	9	15	9	17
3	sample		value					
4	1		10					
5	1		11					
6	1		8					
7	1		8					
8	1		14					
9	2		9					
10	2		12					
11	2		13					
12	2		9					
13	2		15					
14	2		9					
15	2		17					
16								
17		symbol	value	formula				
18	no of observation	n	12					
19	median	Md	10.5					
20	sample size 1	n1	5					
21	sample size 2	n2	13					
22		k	9					

	A	B	C	D	E	F	G	H	I	J
24	no of obs less than median									
25	in 1st sample	a	3							
26										
27	Let Let Md_1 and Md_2 be median of I population and II population respectively.									
28										
29	Problem to test									
30	H0: There is no significant difference between median of I population and median of II population ($Md_1 = Md_2$)									
31	H_1: There is significant difference between median of I population and II population ($Md_1 \neq Md_2$)									
32	Test statistic									
33	$P(A=a) =$	$\frac{c(n_1,a)c(n_2,k-a)}{c(n_1+n_2,k)}$								
34										
35										
36	level ov significance	0.05								
37	Critical value									
38	$P = P(A \geq a) = P(A \geq 3)$	0.5								
39										
40	decision									
41	since 2*p=	1								
42	it is insignificant									
43										

Using SPSS

Analyze\Non parametric test\Legacy Dailogs\k-independent samples





Outputs

```
[DataSet1] D:\Desktop\1111111111111111\stat2_pra\md small.sav
```

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	Percentiles 50th (Median)	Percentiles 75th
value value	12	11.2500	2.95804	8.00	17.00	9.0000	10.5000	13.7500
sample sample	12	1.5833	.51493	1.00	2.00	1.0000	2.0000	2.0000

Median Test

Frequencies

		sample sample 1.00	sample sample 2.00
value value	> Median	2	4
	<= Median	3	3

Test Statistics[a]

	value value
N	12
Median	10.5000
Exact Sig.	1.000

a. Grouping Variable: sample sample

Answer as the p value associated with the median test is about 1 it indicates that null hypothesis of no difference is accepted for the given data.

using STATA

median value, by(sample) medianties(below)

```
. median value, by(sample) medianties(below)

Median test

 Greater  |
 than the |        sample
  median  |         1          2 |     Total
----------+----------------------+----------
       no |         3          3 |         6
      yes |         2          4 |         6
----------+----------------------+----------
    Total |         5          7 |        12

      Pearson chi2(1) =   0.3429   Pr = 0.558

 Continuity corrected:
      Pearson chi2(1) =   0.0000   Pr = 1.000
```

Large sample size ($n_1 > 10$, $n_2 > 10$)

An IQ test was given to a randomly selected 15 male and 20 female students of a university. Their scores were recorded as follows;

Male: 56, 66, 62, 81, 75, 73, 83, 68, 48, 70, 60, 77, 86, 44, 72

Female: 63, 77, 65, 71, 74, 60, 76, 61, 67, 72, 64, 65, 55, 89, 45, 53, 68, 73, 50, 81

Use median test to determine whether IQ of male and female students is same in the university.

(Given that the median of combined sample = 68)

Using EXCEL

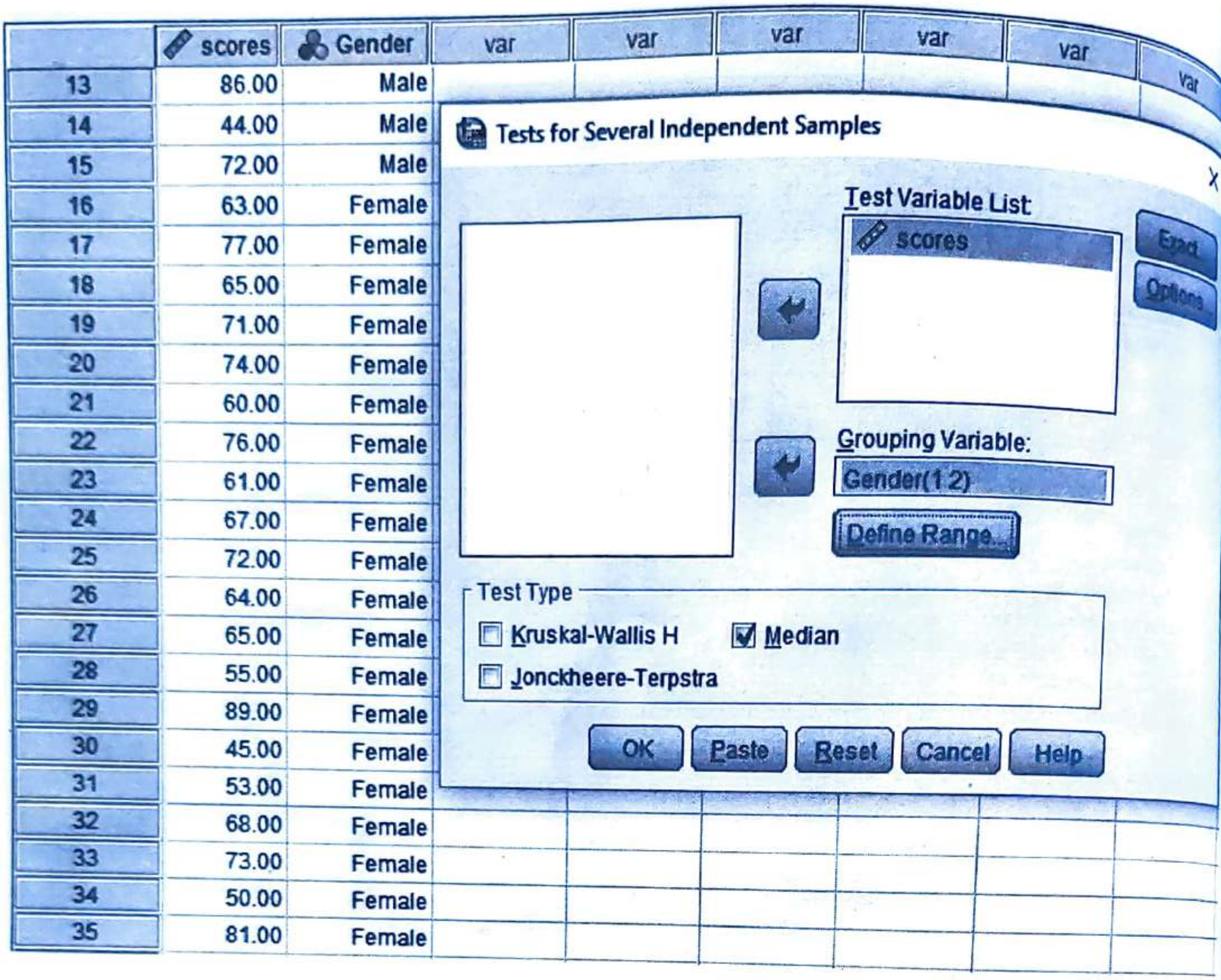
	A	B	C	D	E
5	scores	Gender			
6	56	Male			
7	66	Male			
8	62	Male			
9	81	Male			
10	75	Male			
11	73	Male			
12	83	Male			
13	68	Male			
14	48	Male			
15	70	Male			
16	60	Male			
17	77	Male			
18	86	Male			
19	44	Male			
20	72	Male			
21	63	Female			
22	77	Female			
23	65	Female			
24	71	Female			
25	74	Female			
26	60	Female			
27	76	Female			
28	61	Female			
29	67	Female			
30	72	Female			
31	64	Female			
32	65	Female			
33	55	Female			
34	89	Female			
35	45	Female			
36	53	Female			
37	68	Female			
38	73	Female			
39	50	Female			
40	81	Female			

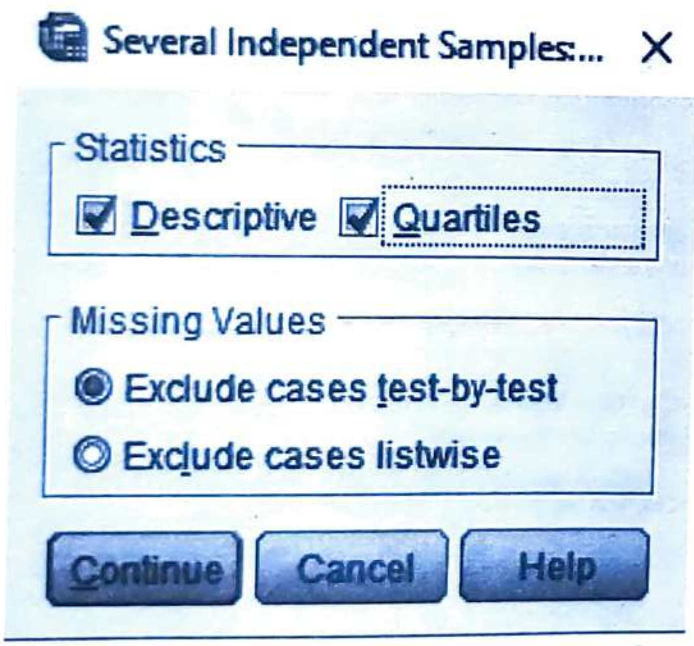
	A	B	C	D	E
41	Table using pivot table and grouping by 25 since minimum value is 44				
42	Count of scores	Column Labels			
43	Row Labels	Male	Female	Grand Total	
44	44-68	7	12	19	
45	69-93	8	8	16	
46	Grand Total	15	20	35	
47					
48	2X2 contingency table for below and above median values from pivot table				
49	Row Labels	Male	Female	Grand Total	
50	44-68	7	12	19	
51	69-93	8	8	16	
52	Grand Total	15	20	35	
53					
54	Here	Symbol	value	fomula	
55	no of male	n1	15	=B52	
56	no of female	n2	20	=C52	
57	Total sample size	N=n1+n2	35	=D52	
58	no. of obs. of male ≤ Md	a	7	=B50	
59	no. of obs. of female ≤ Md	b	12	=C50	
60	no. of obs. of male > Md	c	8	=B51	
61	no. of obs. of female> Md	d	8	=C51	
62					
63	Let Md_1 and Md_2 be median IQ of male and female respectively.				

	A	B	C	D
65	Problem to test			
66	H0: There is no significant difference between IQ of male and female ($Md_1 = Md_2$)			
67	H_1: There is significant difference between IQ of male and female. ($Md_1 \neq Md_2$)			
68	Test statistic			
69	$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$			
70				=(C57*(C58*C61-
71	χ^2	0.614035088		C59*C60)^2)/(D50*D51*B52*C52)
72				
73	level of significance	α	0.05	
74	degrees of freedom	(r-1)(c-1)	1	=(2-1)*(2-1)
75	Critical value			
76	Tabulated chi-square	tab χ^2 0.05,(1)	3.8415	=CHISQ.INV.RT(C73,C74)
77	p value	p	0.4333	=CHISQ.DIST.RT(B71,C74)
78				
79	Decision			
80	significant approach			=IF(B71<C76,"There is no reason to reject
81	There is no reason to reject null hypothesis H0			null hypothesis H0", "Reject H0.")
82				
83	pvalue approach			
84	It is insignificant.			=IF(C77>C73,"It is insignificant.", "It is significant.")

Using SPSS

Analyze\Non parametric test\Legacy Dailogs\k-independent samples


scores
Gender
var
13 86.00 Male
14 44.00 Male
15 72.00 Male
16 63.00 Female
17 77.00 Female
18 65.00 Female
19 71.00 Female
20 74.00 Female
21 60.00 Female
22 76.00 Female
23 61.00 Female
24 67.00 Female
25 72.00 Female
26 64.00 Female
27 65.00 Female
28 55.00 Female
29 89.00 Female
30 45.00 Female
31 53.00 Female
32 68.00 Female
33 73.00 Female
34 50.00 Female
35 81.00 Female
Tests for Several Independent Samples
Test Variable List:
scores
Exact
Options
Grouping Variable:
Gender(1 2)
Define Range...
Test Type
Kruskal-Wallis H
Median
Jonckheere-Terpstra
OK
Paste
Reset
Cancel
Help


Several Independent Samples:...
Statistics
Descriptive
Quartiles
Missing Values
Exclude cases test-by-test
Exclude cases listwise
Continue
Cancel
Help

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
scores scores	35	67.1429	11.31408	44.00	89.00	60.0000	68.0000	75.0000
Gender Gender	35	1.57	.502	1	2	1.00	2.00	2.00

Median Test

Frequencies

		Gender Gender 1 Male	2 Female
scores scores	> Median	8	8
	<= Median	7	12

Test Statistics[a]

		scores scores
N		35
Median		68.0000
Chi-Square		.614
df		1
Asymp. Sig.		.433
Yates' Continuity Correction	Chi-Square	.194
	df	1
	Asymp. Sig.	.659

a. Grouping Variable: Gender Gender

Answer: As the asymptotic significance related with the Chi-square test is found to be 0.659 this indicates that the null hypothesis of no difference is accepted for the given data.

Using STATA

median scores, by(Gender) medianties(below)

```
. median scores, by(Gender) medianties(below)

Median test

 Greater
 than the          Gender
   median      Male    Female     Total

       no         7        12        19
      yes         8         8        16

    Total        15        20        35

   Pearson chi2(1) =   0.6140   Pr = 0.433

 Continuity corrected:
   Pearson chi2(1) =   0.1943   Pr = 0.659
```

Wilcoxon Matched pair signed rank test

Small Sample (n<25)

Use Wilcoxon Matched pair signed rank test to determine the equality of effectiveness of two types of drugs in suppressing pain from following data.

Patient No.	Drug A	Drug B	Patient No.	Drug A	Drug B
1	6.5	3.5	11	5.4	5.5
2	3.7	3.7	12	4.0	4.1
3	3.9	4.7	13	5.7	4.1
4	6.7	5.0	14	3.9	4.2
5	6.2	5.6	15	3.6	3.7
6	6.7	4.3	16	4.9	4.1
7	6.1	5.4	17	3.9	5.4
8	4.3	5.8	18	5.8	3.7
9	5.5	4.3	19	4.9	4.1
10	6.8	4.3	20	4.9	4.1

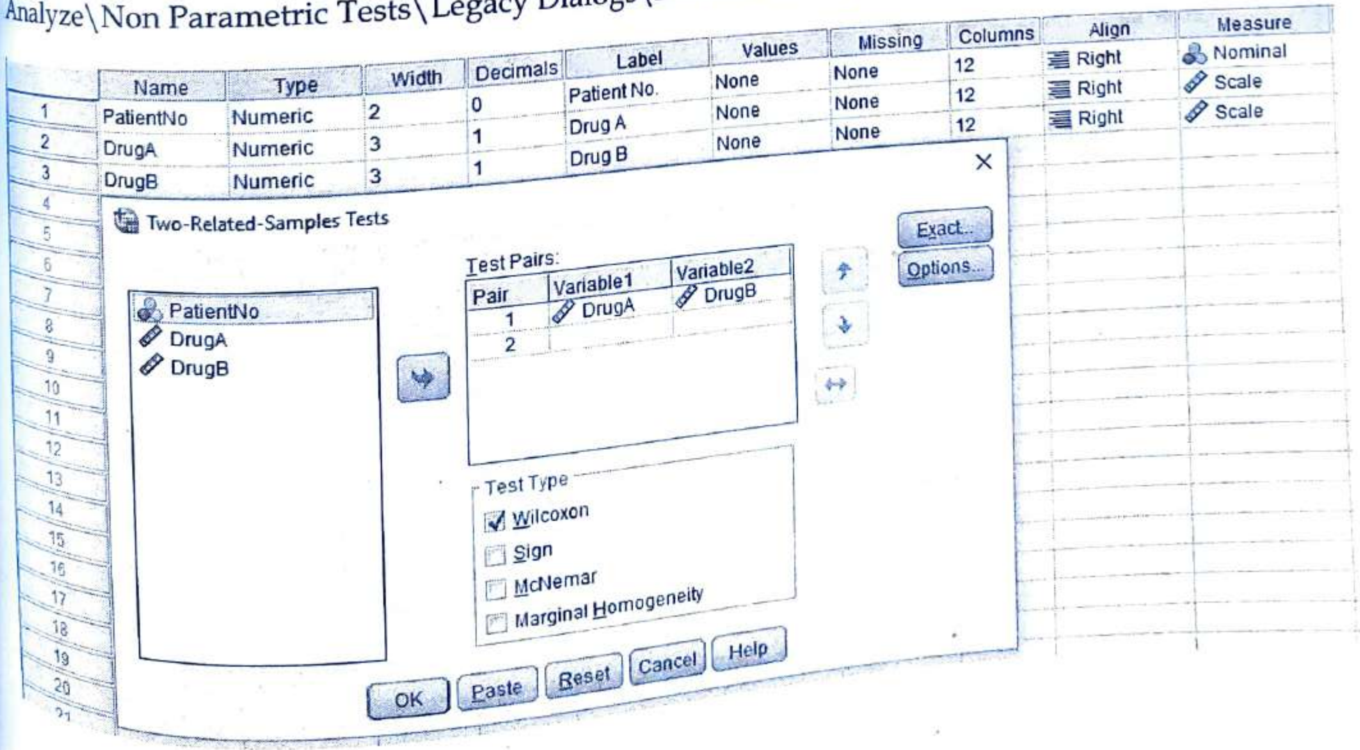
Using Excel

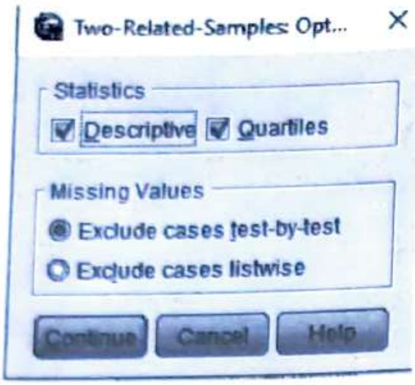
	A	B	C	D	E	F	G	H	I
1	Patient No.	Drug A	Drug B	difference	Sign	|difference|	Rank	Signed Rank	tied count list
2				=B3-C3	=IF(D3>0, 1,IF(D3<0 ,-1,""))	=IF(OR(B3="" ,C3="",D3=0), "",ABS(D3))	=IF(F3="","" ,RANK.AVG (F3,$F$3:$F $22,1))	=IF(F3="","", E3*G3)	=IF(H3="","",IF(CO UNTIF($G$2:G2,G3) >0,"",COUNTIF(G3: G22,G3)))
3	1	6.5	3.5	3.0	1	3	19	19	1
4	2	3.7	3.7	0.0					
5	3	3.9	4.7	-0.8	-1	0.8	7	-7	1
6	4	6.7	5.0	1.7	1	1.7	15	15	1
7	5	6.2	5.6	0.6	1	0.6	5	5	1
8	6	6.7	4.3	2.4	1	2.4	17	17	1
9	7	6.1	5.4	0.7	1	0.7	6	6	1
10	8	4.3	5.8	-1.5	-1	1.5	12	-12	1
11	9	5.5	4.3	1.2	1	1.2	11	11	1
12	10	6.8	4.3	2.5	1	2.5	18	18	1
13	11	5.4	5.5	-0.1	-1	0.1	1.5	-1.5	2
14	12	4.0	4.1	-0.1	-1	0.1	1.5	-1.5	
15	13	5.7	4.1	1.6	1	1.6	14	14	1
16	14	3.9	4.2	-0.3	-1	0.3	4	-4	1
17	15	3.6	3.7	-0.1	-1	0.1	3	-3	1
18	16	4.9	4.1	0.8	1	0.8	9	9	3
19	17	3.9	5.4	-1.5	-1	1.5	13	-13	1
20	18	5.8	3.7	2.1	1	2.1	16	16	1
21	19	4.9	4.1	0.8	1	0.8	9	9	
22	20	4.9	4.1	0.8	1	0.8	9	9	
23									

	A	B	C	D
		Symbol	Value	Formula
29	Sample Size of Drug A	n1	20	=COUNT(B3:B22)
30	Sample Size of Drug B	n2	20	=COUNT(C3:C22)
31	no of case with zero difference		1	=COUNTIF(D3:D22,0)
32	sum of positive ranks	S(+)	148	=SUMIF(E3:E22,">"&0,H3:H22)
33	sum of negative ranks	S(-)	42	=ABS(SUMIF(E3:E22,"<"&0,H3:H22))
34	count of postive ranks	n+	12	=COUNTIF(H3:H22,">"&0)
35	count of negative ranks	n-	7	=COUNTIF(H3:H22,"<"&0)
36	effective sample size	n'	19	=C31-C32
37	**Since sample size, n<25**			
38	**Test statistic**			
39	T = min {S(+) , S(-)}.	T	42	=MIN(C33:C34)
40	**Level of significance**	α	0.05	
41				
42	**Critical value**			
43	At α level of significance, we obtain critical value from Wilcoxon Matched pair signed rank test table	Tα , n = T0.05,19	46	From Wilcoxon Matched pair signed rank test table
44	Note:	Tα , n where n is corrected sample size after omitting di = 0.		
45	**Decision**			
46	Ho is rejected			
47	Note:			
48	Reject H0 at α level of significance if T < Tα , n, accept otherwise			

Using SPSS

Analyze\Non Parametric Tests\Legacy Dialogs\2 Related Samples





Outputs

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
DrugA Drug A	20	5.170	1.1164	3.6	6.8	3.925	5.150	6.175
DrugB Drug B	20	4.480	.7157	3.5	5.8	4.100	4.250	5.300

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
DrugB Drug B - DrugA Drug A	Negative Ranks	12[a]	12.21	146.50
	Positive Ranks	7[b]	6.21	43.50
	Ties	1[c]		
	Total	20		

a. DrugB Drug B < DrugA Drug A

b. DrugB Drug B > DrugA Drug A

c. DrugB Drug B = DrugA Drug A

Test Statistics[a]

	DrugB Drug B - DrugA Drug A
Z	-2.076[b]
Asymp. Sig. (2-tailed)	.038

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Since, $p<0.05$, it is significant.

Using STATA

signrankDrugA=DrugB

```
. signrank DrugA=DrugB

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       12         160       104.5
    negative |        7          49       104.5
        zero |        1           1           1
-------------+---------------------------------
         all |       20         210         210

unadjusted variance          717.50
adjustment for ties           -0.63
adjustment for zeros          -0.25
                         ----------
adjusted variance            716.63

Ho: DrugA = DrugB
             z =   2.073
    Prob > |z| =   0.0382

.
```

Since, $p<0.05$, it is significant.

Large sample size (n> 25)

For large sample size sampling distribution of T is approximately normally distributed with mean μ_T and variance σ^2_T

$$\therefore \mu T \frac{n(n+1)}{4}$$

$$\mu_T = \frac{n(n+1)}{4} \text{ and } \sigma_T^2 = \frac{n(n+1)\,(2n+1)}{24}$$, where n is corrected sample size if $d_i = 0$.

Test statistic

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)\,(2n+1)}{24}}} \sim N(0, 1) \text{ (Here } n = n_c)$$

Level of significance

Let α be the level of significance. Generally fix $\alpha = 0.05$ unless we are given.

Critical value

Critical value $Z_{tabulated}$ is obtained from table according to level of significance and alternative hypothesis.

Decision

Reject H_0 at a level of significance if $Z > Z_{tabulated}$, Accept otherwise.

Use Wilcoxon Matched pair signed rank test to determine the equality of oxygen level of patients in ICU on the day of admission and 7 days after admission from following data.

Patient No	Day1	Day7	Patient No	Day1	Day7	Patient No	Day1	Day7
1	39.32	64.88	21	61.68	62.28	41	64.08	65.03
2	70.59	67.20	22	71.56	72.80	42	60.34	75.71
3	53.92	74.77	23	75.84	44.88	43	37.35	66.13
4	73.46	72.39	24	46.03	46.65	44	60.93	58.75
5	45.16	58.17	25	32.74	58.93	45	66.48	79.03
6	72.21	61.86	26	69.85	55.90	46	46.74	64.87
7	47.35	82.36	27	42.24	59.98	47	58.38	85.63
8	72.32	52.09	28	45.49	83.25	48	74.70	83.81
9	43.60	65.38	29	58.86	81.93	49	49.50	64.98
10	35.69	76.27	30	72.18	75.00	50	75.66	69.47
11	73.25	71.53	31	48.12	56.99	51	73.74	47.54
12	54.22	80.13	32	50.11	78.41	52	63.07	69.22
13	42.50	50.80	33	71.18	69.54	53	64.88	44.26
14	46.85	69.35	34	64.92	67.39	54	51.60	73.08
15	46.17	73.07	35	71.45	59.49	55	39.14	66.76
16	46.89	84.32	36	51.67	78.51	56	59.49	68.01
17	71.03	67.15	37	75.67	67.74	57	51.75	55.39
18	64.79	68.41	38	40.13	82.34	58	39.06	58.86
19	34.68	59.60	39	52.48	52.68	59	75.43	84.15
20	57.02	54.47	40	53.50	43.92			

Using EXCEL

	A	B	C	D	E	F	G	H	I
1	Patient No	Day1	Day7	difference	Sign	|difference|	Rank	Signed Rank	tied count list
				=B3-C3	=IF(D3>0, 1,IF(D3<0 ,-1,""))	=IF(OR(B3="" ,C3="",D3=0), "",ABS(D3))	=IF(F3="" ,"",RANK. AVG(F3,$ F$3:$F$6 1,1))	=IF(F3="","", E3*G3)	=IF(H3="","",I F(COUNTIF($G $2:G2,G3)>0," ",COUNTIF(G3 :G22,G3)))
2	1	39.32	64.88	-25.56	-1	25.56	44	-44	1
3	2	70.59	67.20	3.39	1	3.39	13	13	1
4	3	53.92	74.77	-20.85	-1	20.85	38	-38	1
5	4	73.46	72.39	1.07	1	1.07	5	5	1
6	5	45.16	58.17	-13.01	-1	13.01	29	-29	1
7	6	72.21	61.86	10.35	1	10.35	26	26	1
8	7	47.35	82.36	-35.01	-1	35.01	55	-55	1
9	8	72.32	52.09	20.23	1	20.23	36	36	1
10	9	43.60	65.38	-21.78	-1	21.78	40	-40	1
11	10	35.69	76.27	-40.58	-1	40.58	58	-58	1
12	11	73.25	71.53	1.72	1	1.72	8	8	1
13	12	54.22	80.13	-25.91	-1	25.91	45	-45	1
14	13	42.50	50.80	-8.30	-1	8.3	20	-20	1
15	14	46.85	69.35	-22.50	-1	22.5	41	-41	1
16	15	46.17	73.07	-26.90	-1	26.9	49	-49	1
17	16	46.89	84.32	-37.43	-1	37.43	56	-56	1
18	17	71.03	67.15	3.88	1	3.88	16	16	1
19	18	64.79	68.41	-3.62	-1	3.62	14	-14	1
20	19	34.68	59.60	-24.92	-1	24.92	43	-43	1
21	20	57.02	54.47	2.55	1	2.55	11	11	1

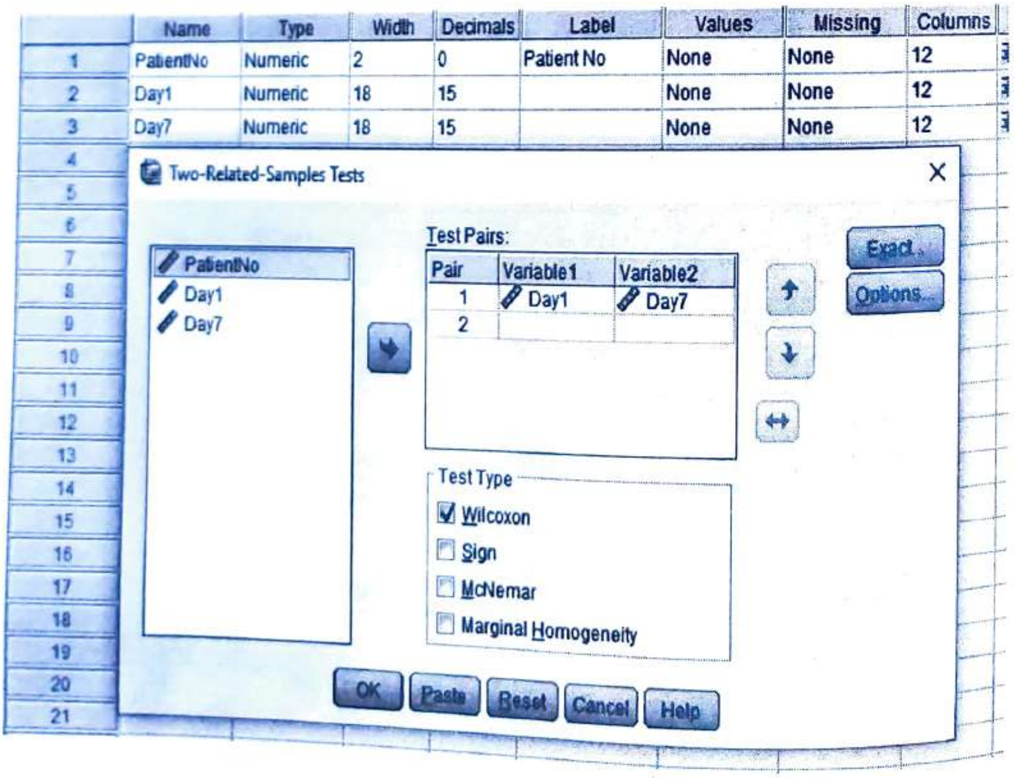
Only part of data is shown

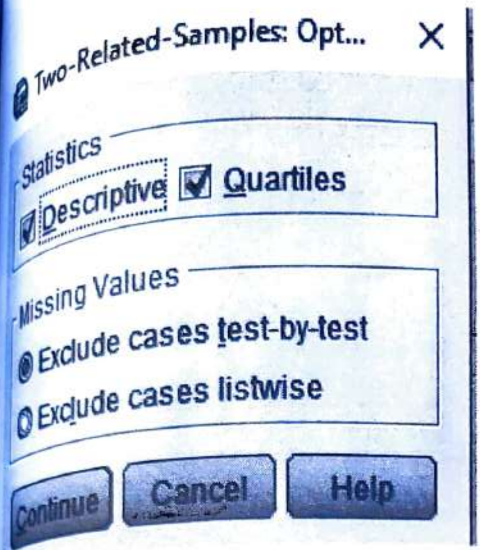
	A	B	C	D	E	F	G	H
64	H0: there is no significant difference between oxygen level							
65	H1: there is significant difference between osygen level							
66								
67		Symbol	Value	Formula				
68	Sample Size of Drug A	n1	59	=COUNT(B3:B22)				
69	Sample Size of Drug B	n2	59	=COUNT(C3:C22)				
70	no of case with zero difference		0	=COUNTIF(D3:D22,0)				
71	sum of positive ranks	S(+)	388	=SUMIF(E3:E22,">"&0,H3:H22)				
72	sum of negative ranks	S(-)	1382	=ABS(SUMIF(E3:E22,"<"&0,H3:H22))				
73	count of postive ranks	n+	17	=COUNTIF(H3:H22,">"&0)				
74	count of negative ranks	n-	42	=COUNTIF(H3:H22,"<"&0)				
75	effective sample size	n'	59	=C31-C32				
76		T	388	=MIN(C71:C72)				
77	Since Sample Size n>25							
78	For large sample size sampling distribution of T is approximately normally distributed with mean							
79	μ_T *and variance* σ_T^2							
80								
81								
82	mean	Symbol	Value	Formula				
83		μ_T	885	=C75*(C75+1)/4	$\frac{n(n+1)}{4}$			
84	Variance	σ_T^2	17552.5	=(C75*(C75+1)*(2*C75+1))/(24)			$\frac{n(n+1)(2n+1)}{24}$	

	A	B	C	D
85				
86	Test Statistics	$\frac{T-\mu_T}{\sigma_T}$		$\frac{T-\frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$
87	Z =			
88				
89				
90	cal Z		-3.75134406	=(C76-C82)/SQRT(C84)
91				
92	level of significance	α	0.05	
93	tabulated Z for Two tailed test			
94		$z_{\alpha/2}$	1.959963985	=NORM.S.INV(1-C92/2)
95	tabulated Z for one tailed test			
96		z_α	1.644853627	=NORM.S.INV(1-C92)
97	pvalue			
98	Two tailed		0.000175889	=2*NORMSDIST(C90)
99	One tailed		8.79446E-05	=NORMSDIST(C90)
100				
101	Decision:			
102	Significant approach for two taied test			
103	Null hypothesis H0 is rejected			
104				
105	p value approach			
106	It is significant			
107				

Using SPSS

Analyze\Non Parametric Tests\Legacy Dialogs\2 Related Samples





Outputs

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
Day1	59	56.93288136	13.00609659	32.74000000	75.84000000	46.17000000	57.02000000	71.03000000
Day7	59	66.66932203	11.19511514	43.92000000	85.63000000	58.86000000	67.20000000	75.00000000

Wilcoxon Signed Ranks Test

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Day7 - Day1	Negative Ranks	17[a]	22.82	388.00
	Positive Ranks	42[b]	32.90	1382.00
	Ties	0[c]		
	Total	59		

a. Day7 < Day1

b. Day7 > Day1

c. Day7 = Day1

Test Statistics[a]

	Day7 - Day1
Z	-3.751[b]
Asymp. Sig. (2-tailed)	.000

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Since, $p<0.05$, it is significant.

Using STATA

signrank Day1=Day7

```
. signrank Day1=Day7

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+---------------------------------
    positive |       17         388         885
    negative |       42        1382         885
        zero |        0           0           0
-------------+---------------------------------
         all |       59        1770        1770

unadjusted variance     17552.50
adjustment for ties         0.00
adjustment for zeros        0.00
                        --------
adjusted variance       17552.50

Ho: Day1 = Day7
             z =  -3.751
    Prob > |z| =   0.0002
```

Since, $p<0.05$, it is significant.

Cochran Q test

Five housewives were asked for the acceptability of four brands of lipsticks for daily use. The response of acceptability (A) and rejection (R) are given below;

House wives	Lipstick Brands			
	Alfa	Beta	Gamma	Delta
H_1	A	R	A	R
H_2	R	A	A	R
H_3	R	A	R	A
H_4	A	R	R	R
H_5	A	A	R	A

Test whether there is any significant difference between brands with respect to acceptability

Solution: Using Excel

Cochran's Q test is a non-parametric statistical test to verify whether k treatments have identical effects. It is the nonparametric counterpart of the two-way ANOVA in RBD. The data for the test are binary that needed to be coded 0, 1 for analysis. In this test the null hypothesis is acceptability of four different lipstick brands are same among housewives.

	A	B	C	D	E	F	G
1		Lipstick Brands					
2	House wives	Alfa	Beta	Gamma	Delta		
3	H_1	A	R	A	R		
4	H_2	R	A	A	R		
5	H_3	R	A	R	A		
6	H_4	A	R	R	R		
7	H_5	A	A	R	A		
8							
9		Lipstick Brands					
10	House wives	Alfa	Beta	Gamma	Delta	C_j	C_j^2
11	H_1	1	0	1	0	2	4
12	H_2	0	1	1	0	2	4
13	H_3	0	1	0	1	2	4
14	H_4	1	0	0	0	1	1
15	H_5	1	1	0	1	3	9
16	Ri	3	3	2	2	10	22
17	Ri^2	9	9	4	4	26	

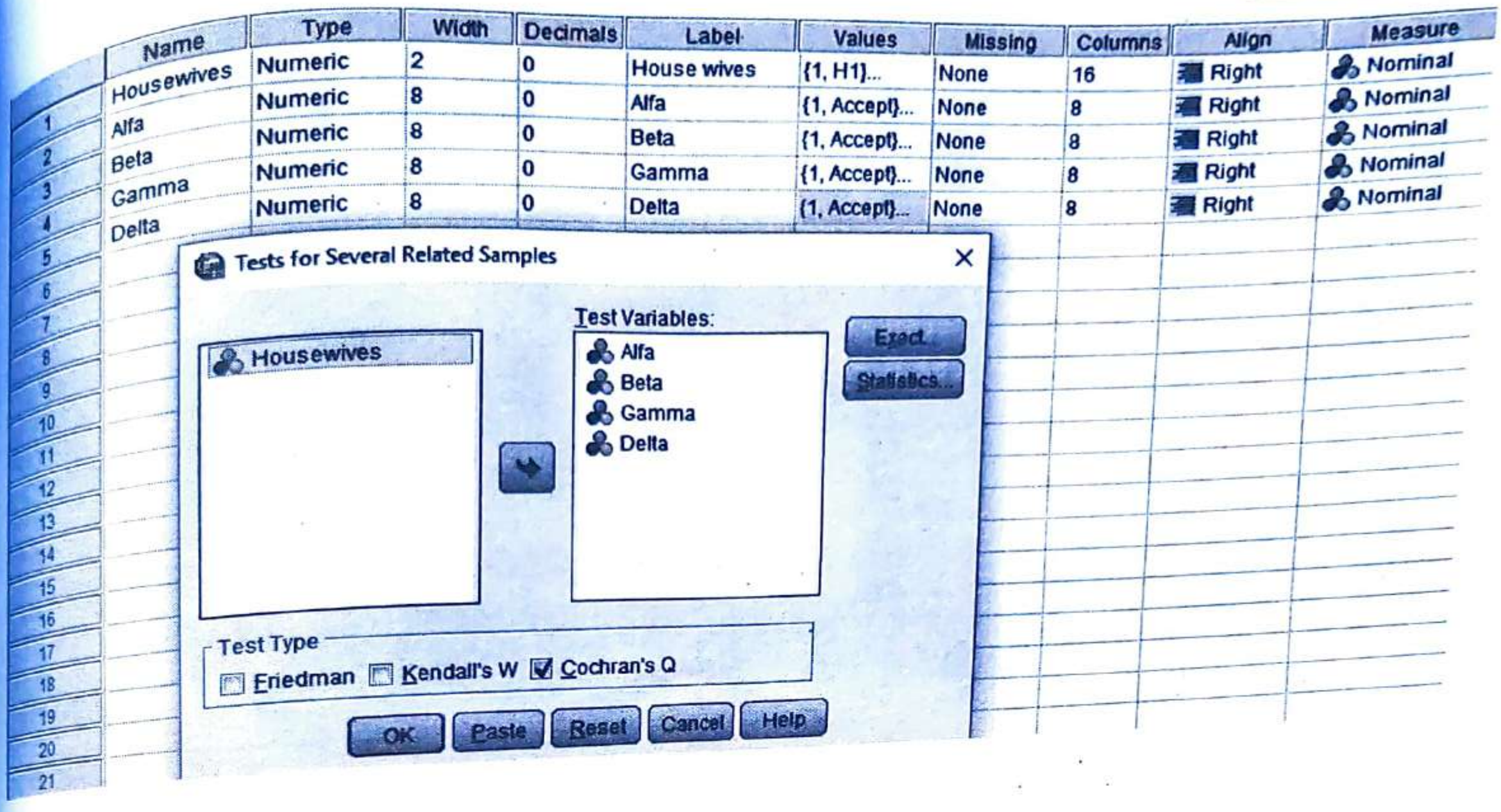
	A	B	C	D	E	F	G	H
19								
20		symbol	value	formula				
21	Number of brands	k	4	=COUNT(B11:E11)				
22	success	1						
23	Number of housewives	n	5	=COUNT(B11:B15)				
24		ΣR_i	10					
25		ΣR_i^2	26					
26		ΣC_j	10					
27		ΣC_j^2	22					
28								
29	**Problem to test**							
30	H_0 : There is no significant difference between brands							
31	H_1: There is at least one significant difference between brands.							
32	**Test statistic**							
33–35	$Q=\frac{(k-1)\left\{k\sum_{i=1}^{k}R_i^2-(\sum_{i=1}^{k}R_i)^2\right\}}{k\sum_{j=1}^{n}C_j-\sum_{j=1}^{n}C_j^2}\sim\chi^2_{k-1}$							
36								
37		Q	0.6666667	=((C21-1)*(C21*C25-(C24)^2))/(C21*C26-C27)				
38	level of significance	α	0.05					
39	degrees of freedom	k-1	3					
40	Critical value	$\chi^2_{\alpha,(k-1)}$	7.8147279					
41	p value	p	0.8810148					
42								

	A	B	C	D	E	F	G	H
43	Decision							
44	Significant Approcah							
45	There if no reason to reject null hypothesis H0.			=IF(C37<C40,"There if no reason to reject null hypothesis H0.","Reject H0")				
46								
47	p value Approach							
48	It is in significant.							
49	Conclusion:			=IF(C41>C38,"It is in significant."," It is significant.")				
50	There is no significant difference between brands according to acceptability.							
51								

Answer: Here the analysis shows the p-value associated with the test statistics Q is 0.88 that indicates acceptance of the null hypothesis of no difference.

Solution: Using SPSS

Analyze\Nonparametric test\Legacy Dialogs\k Related Samples



Outputs

Cochran Test

Frequencies

	Value 0	Value 1
Alfa Alfa	2	3
Beta Beta	2	3
Gamma Gamma	3	2
Delta Delta	3	2

Test Statistics

N	5
Cochran's Q	.667[a]
df	3
Asymp. Sig.	.881

a. 1 is treated as a success.

Kruskal Wallis H test

For non-repeated ranks

Following are the scores obtained by trainees in 3 different categories. Test whether 3 categories have performed equally.

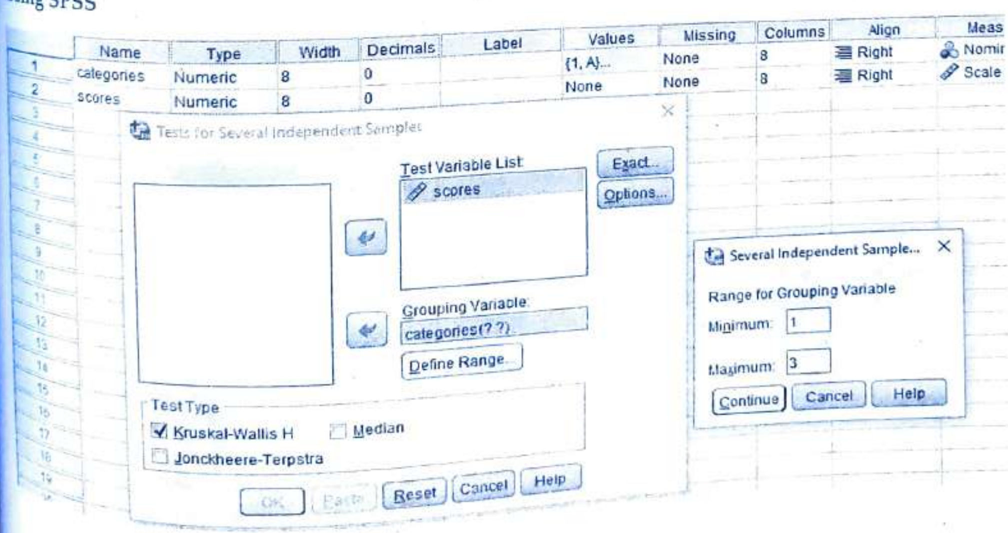
categories	scores									
A	68	65	92	82	62	64	68	92	86	64
B	93	86	73	87	76	85	67	79	75	75
C	95	72	85	70	80	80	78	85	72	90

Using Excel

	A	B	C	D	E	F	G
6	categories	scores	rank				
7			=RANK.AVG(B8:B22,B8:B22,1)				
8	1	67	9				
9	1	77	14				
10	1	79	15				
11	1	72	13				
12	1	64	7				
13	2	55	3				
14	2	50	1				
15	2	58	5				
16	2	57	4				
17	2	54	2				
18	3	66	8				
19	3	70	12				
20	3	69	11				
21	3	62	6				
22	3	68	10				
23			120				
24			=SUM(C8:C22)				
25							
26	test statistics						
27	H0: there is no significant difference betweenmedain score of three categories						
28	H1: there is significant difference betweenmedain score of three categories						

	A	B	C	D	E
	cases	code	symbol	value	formula
32	no of categories		k	3	=MAX(A8:A22)
33	no of categories	1	n1	5	=COUNTIF($A$8:$A$22,B34)
34	no of categories	2	n2	5	=COUNTIF($A$8:$A$22,B35)
35	no of categories	3	n3	5	=COUNTIF($A$8:$A$22,B36)
36	no of observations		n	15	=COUNT(A8:A22)
37	sum of rank in	1	R1	58	=SUMIF($A$8:$A$22,B38,$C$8:$C$22)
38	sum of rank in	2	R2	15	=SUMIF($A$8:$A$22,B39,$C$8:$C$22)
39	sum of rank in	3	R3	47	=SUMIF($A$8:$A$22,B40,$C$8:$C$22)
40					
41					
42	For not tied observations				
43	H	$=\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(n+1)$		9.98	=((12/(D37^2+D37))*((D38^2/D34)+(D39^2/D35)+(D40^2/D36))-3*(D37+1))
44					
45					
46					
47					
48	level of significance		α	0.05	
49	degrees of freedom		k-1	2	=D33-1
50	critical value		$\chi^2_{\alpha,k-1}$	5.99146	=CHISQ.INV(1-D49,D50)
51	p value		p	0.99319	=CHISQ.DIST(D45,D50,TRUE)
52					
53	Decision				
54	Significant approach				
55	Reject Null Hypothesis HO				=IF(D45<D51,"There is no reason to reject null hypothesis HO","Reject Null Hypothesis HO")
56					
57	p value approach				
58	it is insignificant				=IF(D52>D49,"it is insignificant","it is significant")
59					
60					

Using SPSS



Outputs

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
scores	15	64.53	8.493	50	79	57.00	66.00	70.00
categories	15	2.00	.845	1	3	1.00	2.00	3.00

Kruskal-Wallis Test

Ranks

	categories	N	Mean Rank
scores	1 A	5	11.60
	2 B	5	3.00
	3 C	5	9.40
	Total	15	

Test Statistics[a,b]

	scores
Kruskal-Wallis H	9.980
df	2
Asymp. Sig.	.007

a. Kruskal Wallis Test

Using STATA

kwallis scores, by(categories)

```
. kwallis scores, by(categories)

Kruskal-Wallis equality-of-populations rank test
```

catego~s	Obs	Rank Sum
A	5	58.00
B	5	15.00
C	5	47.00

```
chi-squared =      9.980 with 2 d.f.
probability =      0.0068

chi-squared with ties =     9.980 with 2 d.f.
probability =      0.0068
```

Answer: As we found that P value for Chi-square test with tie is 0.0068 which is smaller than 0.05 we can conclude that the sores among three categories is significantly different as null is rejected.

For repeated ranks

Following are the scores obtained by trainees in 3 different categories. Test whether 3 categories are equally performed.

categories	scores									
A	68	65	92	82	62	64	68	92	86	64
B	93	86	73	87	76	85	67	79	75	75
C	95	72	85	70	80	80	78	85	72	90

Using Excel

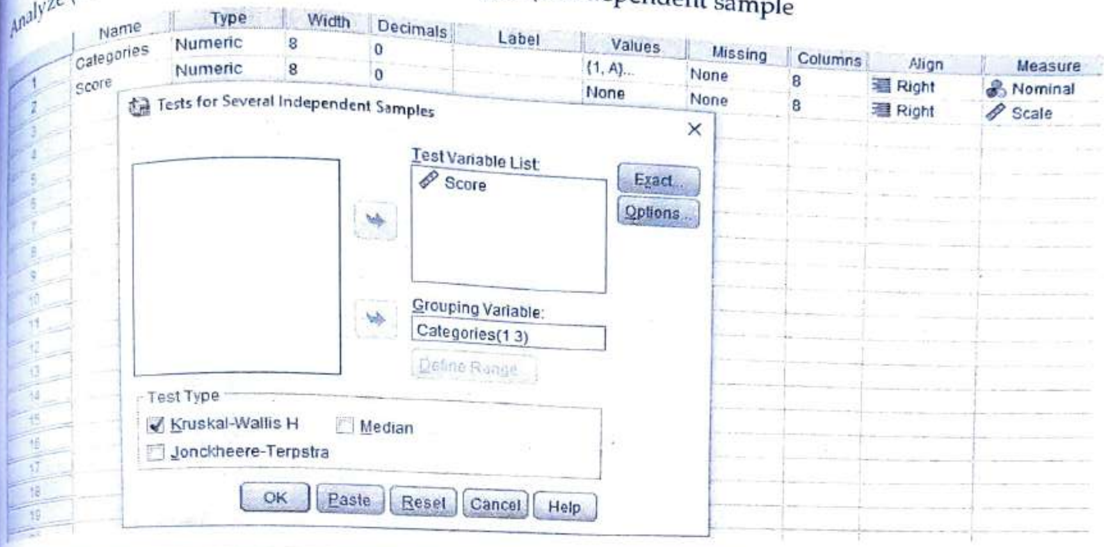
A	B	C	D	E
categories	scores	rank		
		=RANK.AVG(B8:B37,B8:B37,1)		
1	68	6.5		
1	65	4		
1	92	27.5		
1	82	19		
1	62	1		
1	64	2.5		
1	68	6.5		
1	92	27.5		
1	86	23.5		
1	64	2.5		
2	93	29		
2	86	23.5		
2	73	11		
2	87	25		
2	76	14		
2	85	21		
2	67	5		
2	79	16		
2	75	12.5		
2	75	12.5		
3	95	30		
3	72	9.5		
3	85	21		
3	70	8		
3	80	17.5		
3	80	17.5		
3	78	15		
3	85	21		
3	72	9.5		
3	90	26		
		465		
		=SUM(C8:C37)		

	A	B	C	D	E
41	test statistics				
42	H0: there is no significant difference betweenmedain score of three categories				
43	H1: there is significant difference betweenmedain score of three categories				
44					
45					formula
46		code	symbol	value	
47	cases		k	3	=MAX(A8:A37)
48	no of categories	1	n1	10	=COUNTIF($A$8:$A$37,B49)
49	no of categories	2	n2	10	=COUNTIF($A$8:$A$37,B50)
50	no of categories	3	n3	10	=COUNTIF($A$8:$A$37,B51)
51	no of categories		n	30	=COUNT(A8:A37)
52	no of observations	1	R1	120.5	=SUMIF($A$8:$A$37,B53,$C$8:$C$37)
53	sum of rank in	2	R2	169.5	=SUMIF($A$8:$A$37,B54,$C$8:$C$37)
54	sum of rank in	3	R3	175	=SUMIF($A$8:$A$37,B55,$C$8:$C$37)
55	sum of rank in				
56		2.5	t1	2	=COUNTIF($C$8:$C$37,B57)
57	no of repeated ranks	6.5	t2	2	=COUNTIF($C$8:$C$37,B58)
58	no of repeated ranks	9.5	t3	2	=COUNTIF($C$8:$C$37,B59)
59	no of repeated ranks	12.5	t4	2	=COUNTIF($C$8:$C$37,B60)
60	no of repeated ranks	17.5	t5	2	=COUNTIF($C$8:$C$37,B61)
61	no of repeated ranks	23.5	t6	2	=COUNTIF($C$8:$C$37,B62)
62	no of repeated ranks	27.5	t7	2	=COUNTIF($C$8:$C$37,B63)
63	no of repeated ranks	21	t8	3	=COUNTIF($C$8:$C$37,B64)
64	no of repeated ranks				

	A	B	C	D	E
65	$\sum T = \sum_{p=1}^{8} T_p = \sum_{i=1}^{8}(t_i^3 - t)$			66	=SUMPRODUCT(D57:D64^3)-SUM(D57:D64)
66					
67					
68					
69	For ties observations				
70	$H_{corrected}$	$= \dfrac{\dfrac{12}{n(n+1)}\sum_{i=1}^{k}\dfrac{R_i^2}{n_i} - 3(n+1)}{1 - \dfrac{\sum T}{n^3 - n}}$		2.328925	=((12/(D52^2+D52))*((D53^2/D49)+(D54^2/D50)+(D55^2/D51))-3*(D52+1))/(1-(D66/(D52^3-D52)))
71					
72					
73					
74					
75	level of significance		α	0.05	
76	degrees of freedom		k-1	2	=D48-1
77	critical value		$\chi^2_{\alpha,k-1}$	5.991465	=CHISQ.INV(1-D75,D76)
78	p value		p	0.68791	=CHISQ.DIST(D71,D76,TRUE)
79					
80	Decision				
81	Significant approach				
82	There is no reason to reject null hypothesis H0				=IF(D71<D77,"There is no reason to reject null hypothesis H0","Reject Null Hypothesis H0")
83					
84	p value approach				
85	it is insignificant				=IF(D78>D75,"it is insignificant","it is significant")

Using SPSS

Analyze\Non parametric test\Legacy dialogue\k independent sample



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
Score	30	78.20	9.739	62	95	69.50	78.50	86.00
Categories	30	2.00	.830	1	3	1.00	2.00	3.00

Kruskal-Wallis Test

Ranks

	Categories	N	Mean Rank
Score	1 A	10	12.05
	2 B	10	16.95
	3 C	10	17.50
	Total	30	

Test Statistics[a,b]

	Score
Kruskal-Wallis H	2.329
df	2
Asymp. Sig.	.312

a. Kruskal Wallis Test

b. Grouping Variable: Categories

Using STATA

kwallis Score, by(Categories)

```
kwallis Score, by(Categories)

Kruskal-Wallis equality-of-populations rank test

  +------------------------+
  | Catego~s | Obs | Rank Sum |
  |----------+-----+----------|
  |        A |  10 |   120.50 |
  |        B |  10 |   169.50 |
  |        C |  10 |   175.00 |
  +------------------------+

chi-squared =     2.323 with 2 d.f.
probability =     0.3130

chi-squared with ties =     2.329 with 2 d.f.
probability =     0.3121
```

Answer: As we found that P value with tie for the Chi-square test is 0.3121 which is larger than 0.05 we can conclude that the sores among three categories is not significantly different as null is accepted.

Friedman F test

A survey was conducted in four hospitals in a particular city to obtain the number of babies born over a 12 months' period. This time period was divided into four seasons to test the hypothesis that the birth rate is constant over all the four seasons. The results of the survey were as follows:

Hospital	No of births			
	Winter	Spring	Summer	Fall
A	92	72	94	77
B	15	16	10	17
C	58	71	51	62
D	19	26	20	18

Analyze the data using Friedman two way ANOVA test.

Using EXCEL

Here we assume that the depended variable (here birth) is normally distributed, which is the needed assumption for this test. Also, we perform Friedman test when we have one within-subjects independent variable(hospital) with two or more levels(seasons) and a dependent variable. We use this test to determine if there is a difference in the birth by season, so the null hypothesis tested will be H0: birth is distributed homogeneously in all seasons vs H1: at least in some season it is different from other.

Hospital	No of births			
	Winter	Spring	Summer	Fall
A	92	72	94	77
B	15	16	10	17
C	58	71	51	62
D	19	26	20	18

ranks						
Hospital	Winter	Spring	Summer	Fall		
A	3	1	4	2		formula
B	2	3	1	4		=RANK.AVG(B4:E4,B4:E4,1)
C	2	4	1	3		=RANK.AVG(B5:E5,B5:E5,1)
D	2	4	3	1		=RANK.AVG(B6:E6,B6:E6,1)
Ri	9	12	9	10		=RANK.AVG(B7:E7,B7:E7,1)
Ri^2	81	144	81	100		=SUM(E11:E14)
					406	=E15^2
						=SUM(B16:E16)

Test Hypothesis

H0: The birth rate is constant over all four seasons.

H1: The birth rate is not constant over all four seasons.

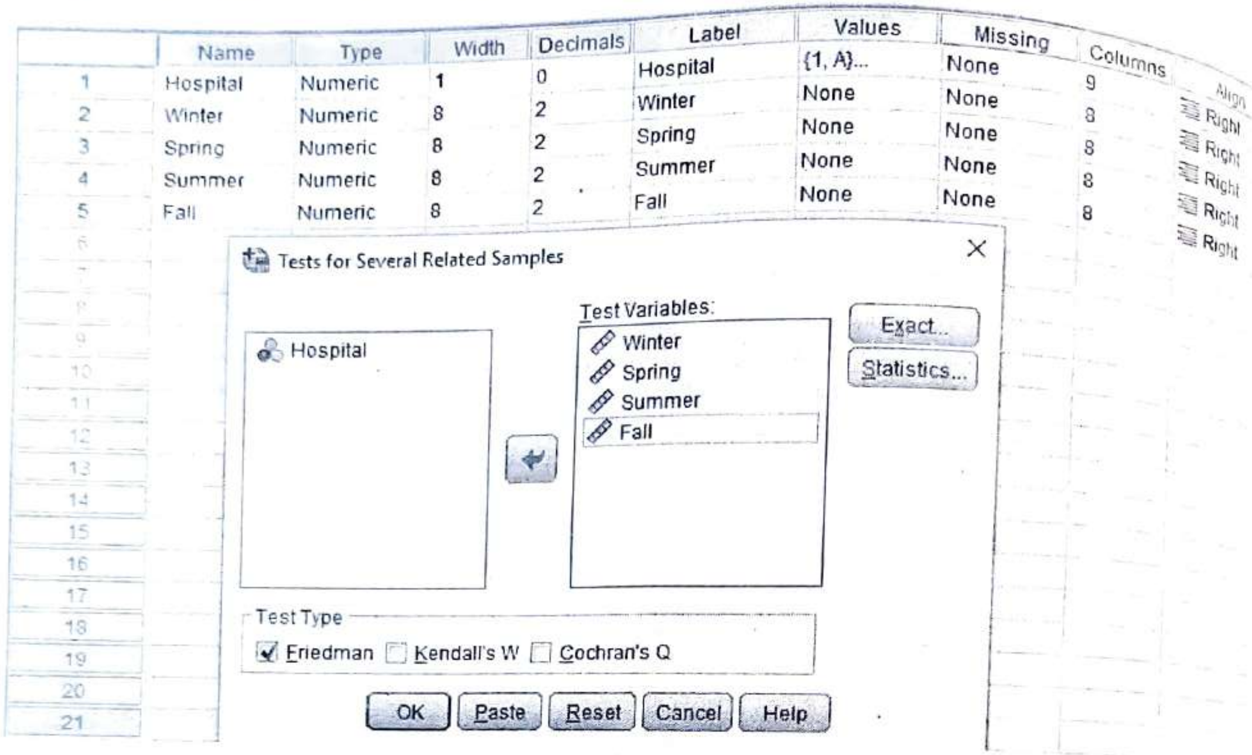
cases	symbols	value	formula
sample size	n	4	
no of samples	k	4	

cases	symbols	value	formula
sample size	n	4	=COUNT(B11:B14)
no of samples	k	4	=COUNT(B11:E11)
Test statistic	$F_r = \frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1)$		
	F_r	0.9	=(12/(C25*C26*(C26+1)))*F16-3*C25*(C26+1)
level of significance	α	0.05	
Critical value	p=P(Fr>0.9)	0.9	from table
decision			
There is no reason to reject Null hypothesis H0			

Answer: As from the Friedman test the p-value of the test is 0.9 we conclude that the null hypothesis is correct. This indicates there is no seasonal effect found in the birth rate in the data.

Using SPSS

Analyze\Nonparametric test\Legacy Dialogs\k Related Samples



Outputs

Friedman Test

Ranks

	Mean Rank
Winter Winter	2.25
Spring Spring	3.00
Summer Summer	2.25
Fall Fall	2.50

Test Statistics[a]

N	4
Chi-Square	.900
df	3
Asymp. Sig.	.825

a. Friedman Test

Friedman test for repeated ranks

Three different advertising media T.V., Radio and Newspaper are being compared to study their effectiveness in promoting sales of WaiWai noodles. Each advertising media is exposed for specified period of time and sales (000 package) from 10 stores located at different areas are recorded.

Advertising Media	Stores									
	A	B	C	D	E	F	G	H	I	J
T.V.	20	21	15	12	14	17	21	16	20	18
Radio	7	9	11	12	10	10	14	12	8	7
News Paper	8	6	11	12	9	6	8	10	8	6

Are three advertising media equally effective, use Friedman two-way ANOVA test.

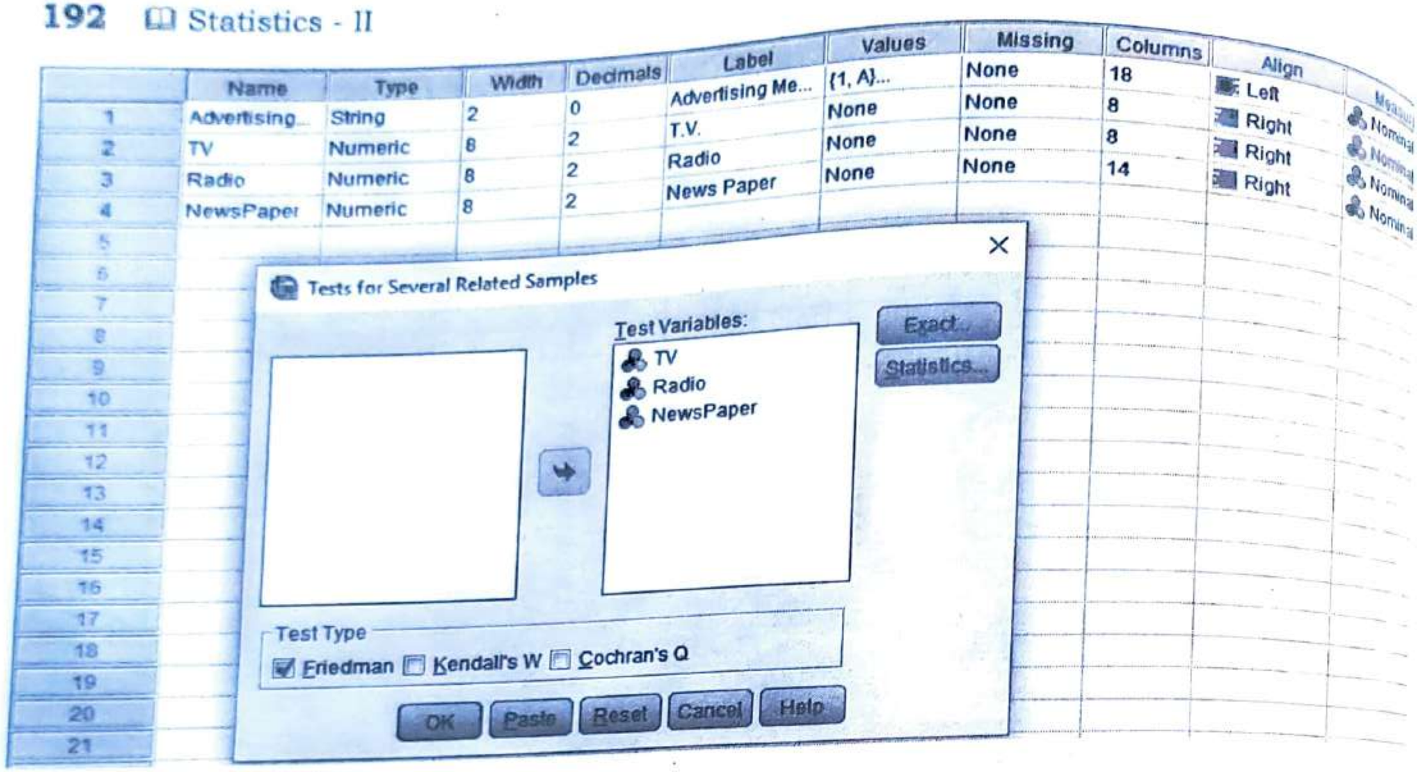
Using EXCEL

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Stores											
2	Advertising Media	A	B	C	D	E	F	G	H	I	J		
3	T.V.	20	21	15	12	14	17	21	16	20	18		
4	Radio	7	9	11	12	10	10	14	12	8	7		
5	News Paper	8	6	11	12	9	6	8	10	8	6		
6													
7													
8	ranks												
9	Advertising Media	A	B	C	D	E	F	G	H	I	J	Ri	Ri^2
10	T.V.	3	3	3	2	3	3	3	3	3	3	29	841
11	Radio	1	2	1.5	2	2	2	2	2	1.5	2	18	324
12	News Paper	2	1	1.5	2	1	1	1	1	1.5	1	13	169
13													1334
14													
15													
16	Test Hypothesis												
17	H0: Three advertising media are equally effective.												
18	H1: Three advertising media are not equally effective.												
19													

	A	B	C	D	G
20	cases	symbols	value	formula	
21	sample size	n	10	=COUNT(B11:B14)	
22	no of samples	k	3	=COUNT(B11:E11)	
23	1.5	t1	2		repeats of C
24	2	t2	3		repeats of D
25	1.5	t3	2		repeats of I
26–28		$F_r = \dfrac{\dfrac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1)}{1-\sum\dfrac{(t_i^3-t_i)}{n(k^3-k)}}$			
29	Test statistic				
30		F_r	15.76471	=((12/(C21*C22*(C22+1)))*M13-3*C21*(C22+1))/(1-(((C23^3-C23)+(C24^3-C24)+(C25^3-C25))/(C21*(C22^3-C22))))	
31					
32	level of significance	α	0.05		
33	degrees of freedom	df	2	=C22-1	
34	Critical value	$\chi^2_{0.05(2)}$	5.991465	=CHISQ.INV.RT(C32,C33)	
35	p value	p	0.000377	=CHISQ.DIST.RT(C30,C33)	
36	decision				
37	significant approach				
38	Reject H0			=IF(C34>C30,"There is no reason to reject null hypothesis H0", "Reject H0")	
39					
40	p value approach				
41	It is significant			=IF(C35>C32,"It is insignificant","It is significant")	

Using SPSS

Analyze\Nonparametric test\Legacy Dialogs\k Related Samples



OUTPUTS

Friedman Test

Ranks

	Mean Rank
TV T.V.	2.90
Radio Radio	1.80
NewsPaper News Paper	1.30

Test Statistics[a]

N	10
Chi-Square	15.765
df	2
Asymp. Sig.	.000

a. Friedman Test

□□□

4

Chapter

MULTIPLE CORRELATION AND REGRESSION

CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Multiple and partial correlation
- Introduction of multiple linear regression, Hypothesis testing of multiple regression, Test of significance of regression, Test of individual regression coefficient
- Model adequacy tests
- Problems and illustrative examples related using software.

Partial Correlation

It is the relationship between two variables keeping all the other remaining variables involved constant. The correlation between two variables keeping one other variable constant is called first order correlation. The correlation between two variables keeping other two variables constant is called second order correlation and so on.

We are interested to study the relationship of production of wheat with seeds, fertilizer, irrigation etc. If we study the relationship between production of wheat with seeds keeping fertilizer and irrigation condition constant is the case of partial correlation. Similarly the study of relationship between production of wheat with fertilizer keeping seeds and irrigation constant, the study of relationship between production of wheat with irrigation keeping seeds and fertilizer constant are the case of partial correlation.

Let us consider three variables X_1, X_2 and X_3 then the partial correlation coefficient between X_1 and X_2 keeping X_3 constant is denoted by $r_{12\cdot3}$ and is given by $r_{12\cdot3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\,.\sqrt{1 - r_{23}^2}}$

Similarly, the partial correlation coefficient between X_1 and X_3 keeping X_2 constant is denoted by $r_{13.2}$ and is given by $r_{13\cdot2} = \dfrac{r_{13} - r_{12}\, r_{32}}{\sqrt{1 - r_{12}^2}\,.\sqrt{1 - r_{32}^2}}$

Also, the partial correlation coefficient between X_2 and X_3 keeping X_1 constant is denoted by $r_{23.1}$ and is given by $r_{23.1} = \dfrac{r_{23} - r_{21}\, r_{31}}{\sqrt{1 - r_{21}^2}\,.\sqrt{1 - r_{31}^2}}$.

Remarks:

1. (i) $r_{12} = r_{21}$ (ii) $r_{13} = r_{31}$ (iii) $r_{23} = r_{32}$
2. (i) $r_{12.3} = r_{21.3}$ (ii) $r_{13.2} = r_{31.2}$ (iii) $r_{23.1} = r_{32.1}$
3. (i) $-1 \leq r_{12\cdot3} \leq 1$ (ii) $-1 \leq r_{13.2} \leq 1$ (iii) $-1 \leq r_{23.1} \leq 1$
4. r_{12}, r_{13}, r_{23} are zero order correlation coefficients.
5. $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ are first order correlation coefficients.
6. $r_{12.34}$, $r_{23.14}$, $r_{13.24}$, $r_{14.23}$, $r_{24.13}$, $r_{34.12}$ are second order correlation coefficients.

Coefficient of Partial Determination

It is the square of partial correlation coefficient. It is used to measure variation in one variable as explained by other variable keeping next variable constant.

If $r_{12.3} = 0.8$ then coefficient of partial determination is $r_{12.3}^2 = (0.8)^2 = 0.64 = 64\%$. It means 64% of the total variation in X_1 has been explained by variable X_2 when the next variable X_3 is held constant.

Example 1: If $r_{12} = 0.8$, $r_{13} = -0.4$ and $r_{23} = -0.58$ find $r_{12.3}$.

Solution:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\,.\sqrt{1 - r_{23}^2}} = \frac{0.8 - (0.4) \times (-0.58)}{\sqrt{1 - (0.4)^2}\sqrt{1 - (0.58)^2}}$$

$$= \frac{0.568}{\sqrt{0.84}\sqrt{0.6636}} = \frac{0.568}{\sqrt{0.5574}} = 0.76$$

Example 2: If $r_{12} = 0.4$, $r_{23} = 0.5$ and $r_{13} = 0.6$. Find (i) $r_{23.1}$ (ii) $r_{23.1}^2$ and interpret.

Solution:

$$r_{23.1} = \frac{r_{23} - r_{21}\,r_{31}}{\sqrt{1 - r_{21}^2}\,.\,\sqrt{1 - r_{31}^2}}$$

$$= \frac{0.5 - 0.4 \times 0.6}{\sqrt{1 - (0.4)^2}\sqrt{1 - (0.6)^2}} = \frac{0.26}{\sqrt{0.84}\sqrt{0.64}} = \frac{0.26}{\sqrt{0.5376}} = 0.35$$

$$r_{23.1}^2 = (0.35)^2 = 0.1225 = 12.25\%$$

It means 12.25% variation in variable X_2 is explained by variable X_3 keeping variable X_1 constant.

Example 3: Are the following data consistent; $r_{12} = -0.8$, $r_{13} = 0.3$ and $r_{23} = 0.4$.

Solution:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1 - r_{13}^2}\,.\,\sqrt{1 - r_{23}^2}}$$

$$= \frac{-0.8 - (0.3) \times (0.4)}{\sqrt{1 - (0.3)^2}\sqrt{1 - (0.4)^2}} = \frac{-0.92}{\sqrt{0.91}\sqrt{0.84}} = \frac{-0.92}{\sqrt{0.7644}} = \frac{-0.92}{0.874} = 1.052$$

Since $r_{12.3}$ should lie between -1 and +1, here $r_{12.3} = 1.051 > 1$. Hence the given data are inconsistent.

Multiple Correlation

The relationship among three or more variables simultaneously (at the same time) is called multiple correlation. In this case relationship of a variable with two or more variables is studied at a time.

We are interested to study the relationship of production of paddy with seeds, fertilizer and irrigation etc. If we study the relationship of production of paddy with seeds, fertilizer and irrigation jointly is called multiple correlation.

Let us consider three variables X_1, X_2 and X_3 the multiple correlation coefficient of X_1 with X_2 and X_3 is denoted by $R_{1.23}$ and is given by $R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}}$

Similarly, the multiple correlation coefficient of X_2 with X_1 and X_3 is denoted by $R_{2.13}$ and is given by $R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}\,r_{23}\,r_{13}}{1 - r_{13}^2}}$

Also multiple correlation coefficient of X_3 with X_1 and X_2 is denoted by $R_{3.12}$ and is given by

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}\,r_{32}\,r_{12}}{1 - r_{12}^2}}$$

Properties of Multiple Correlation Coefficient

1. Multiple correlation coefficient lies between 0 and 1
   (i) $0 \le R_{1.23} \le 1$ (ii) $0 \le R_{2.13} \le 1$ (iii) $0 \le R_{3.12} \le 1$
2. Multiple correlation coefficient is not less than zero order correlation coefficient (simple correlation coefficient)
   (i) $R_{1.23} \ge r_{12}, r_{13}, r_{23}$ (ii) $R_{2.13} \ge r_{21}, r_{23}, r_{13}$ (iii) $R_{3.12} \ge r_{31}, r_{32}, r_{12}$

3. (i) If $R_{1.23} = 0$ then $r_{12} = 0$ and $r_{13} = 0$ (ii) If $R_{2.13} = 0$ then $r_{21} = 0$ and $r_{23} = 0$.
   iii) If $R_{3.12} = 0$ then $r_{31} = 0$ and $r_{32} = 0$.
4. i) $R_{1.23} = R_{1.32}$ (ii) $R_{2.13} = R_{2.31}$ (iii) $R_{3.12} = R_{3.21}$

Coefficient of Multiple Determination

It is the square of multiple correlation coefficient. It is used to measure in variation of one variable as explained by two remaining variables.

If $R_{1.23} = 0.7$ then coefficient of multiple determination is $R_{1.23}^2 = 0.49 = 49\%$. It means 49% variation in variable X_1 is explained by two other variables X_2 and X_3 and remaining 51% is due to the effect of other factors.

Example 4: If $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$ find $R_{1.23}$.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.77)^2 + (0.72)^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - (0.52)^2}}$$

$$= \sqrt{0.7334} = 0.8564$$

Example 5: If $r_{12} = 0.7$, $r_{23} = r_{31} = 0.5$ find (i) $R_{1.23}$ (ii) $R_{1.23}^2$ and interpret

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.7)^2 + (0.5)^2 - 2 \times 0.7 \times 0.5 \times 0.5}{1 - (0.5)^2}} = \sqrt{0.57} = 0.721$$

Now $R_{1.23}^2 = (0.721)^2 = 0.52 = 52\%$.

It means 52% variation in X_1 has been explained by X_2 and X_3.

Example 6: Show that the values $r_{12} = 0.6$, $r_{13} = -0.4$ and $r_{23} = 0.7$ are inconsistent.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (-0.4)^2 - 2 \times 0.6 \times (-0.4) \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (-0.4)^2 - 2 \times 0.6 \times (-0.4) \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.856}{0.51}}$$

$$= 1.29$$

Here $R_{1.23} = 1.29 > 1$

Since $R_{1.23}$ should lie between 0 and 1. Hence inconsistent in the given values.

Example 7: A sample of 10 values of three variables X_1, X_2 and X_3 were obtained as, $\Sigma X_1 = 10$, $\Sigma X_2 = 20$, $\Sigma X_3 = 30$, $\Sigma X_1X_2 = 10$, $\Sigma X_1X_3 = 15$, $\Sigma X_2X_3 = 64$, $\Sigma X_1^2 = 20$, $\Sigma X_2^2 = 68$, $\Sigma X_3^2 = 170$. (i) Find the partial correlation coefficient between X_1 and X_3 eliminating the effect of X_2. (ii) Find the multiple correlation coefficient of X_1 with X_2 and X_3.

Solution:

Here,

$$r_{12} = \frac{n\Sigma X_1X_2 - \Sigma X_1 \Sigma X_2}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_2^2 - (\Sigma X_2)^2}}$$

$$= \frac{10 \times 10 - 10 \times 20}{\sqrt{10 \times 20 - (10)^2}\sqrt{10 \times 68 - (20)^2}}$$

$$= \frac{-100}{\sqrt{100}\sqrt{280}}$$

$$= -0.59$$

$$r_{13} = \frac{n\Sigma X_1X_3 - \Sigma X_1 \Sigma X_3}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_3^2 - (\Sigma X_3)^2}}$$

$$= \frac{10 \times 15 - 10 \times 30}{\sqrt{10 \times 20 - (10)^2}\sqrt{10 \times 170 - (30)^2}}$$

$$= \frac{-150}{\sqrt{100}\sqrt{800}} = -0.53$$

$$r_{23} = \frac{n\Sigma X_2X_3 - \Sigma X_2 \Sigma X_3}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2}\sqrt{nX_3^2 - (\Sigma X_3)^2}}$$

$$= \frac{10 \times 64 - 20 \times 30}{\sqrt{10 \times 68 - (20)^2}\sqrt{10 \times 170 - (30)^2}}$$

$$= \frac{40}{\sqrt{280}\sqrt{800}} = 0.085$$

i) Partial correlation coefficient between X_1 and X_3 eliminating the effect of X_2 is

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

$$= \frac{(-0.53) - (-0.598) \times 0.085}{\sqrt{1 - (-0.598)^2}\sqrt{1 - (0.085)^2}}$$

$$= 0.727$$

ii) Multiple correlation coefficient of X_1 with X_2 and X_3 is

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{32}}}$$

$$= \sqrt{\frac{(-0.598)^2 + (-0.53)^2 - 2 \times (-0.598) \times (-0.53) \times 0.085}{1 - (0.085)^2}}$$

$$= 0.767$$

Example 8: The height and weight of 10 individuals of different ages are given below:

Age (x_1)	11	10	6	10	8	9	10	7	11	8
Height(X_2)	60	67	53	56	64	57	71	58	67	57
Weight(X_3)	57	55	49	52	57	48	59	50	62	51

Find $r_{12.3}$, $r_{13.2}$, $R_{1.23}$.

Solution:

Age(X_1)	Ht(X_2)	Wt(X_3)	$u_1 = X_1 - 10$	$u_2 = X_2 - 60$	$u_3 = X_3 - 50$	u_1^2	u_2^2	u_3^2	u_1u_2	u_1u_3	u_2u_3
11	60	57	1	0	7	1	0	49	0	7	0
10	67	55	0	7	5	0	49	25	0	0	35
6	53	49	-4	-7	-1	16	49	1	28	4	7
10	56	52	0	-4	2	0	16	4	0	0	-8
8	64	57	-2	4	7	4	16	49	-8	-14	28
9	57	48	-1	-3	-2	1	9	4	3	2	6
10	71	59	0	11	9	0	121	81	0	0	99
7	58	50	-3	-2	0	9	4	0	6	0	0
11	67	62	1	7	12	1	49	144	7	12	84
8	57	51	-2	-3	1	4	9	1	6	-2	-3
			$\Sigma u_1 = -10$	$\Sigma u_2 = 10$	$\Sigma u_3 = 40$	$\Sigma u_1^2 = 36$	$\Sigma u_2^2 = 322$	$\Sigma u_3^2 = 358$	$\Sigma u_1u_2 = 42$	$\Sigma u_1u_3 = 9$	$\Sigma u_2u_3 = 248$

Here

$$r_{12} = \frac{n\Sigma u_1u_2 - \Sigma u_1\Sigma u_2}{\sqrt{n\Sigma u_1^2 - (\Sigma u_1)^2}\sqrt{n\Sigma u_2^2 - (\Sigma u_2)^2}}$$

$$= \frac{10 \times 42 - (-10) \times 10}{\sqrt{10 \times 36 - (-10)^2}\sqrt{10 \times 322 - (10)^2}}$$

$$= 0.577$$

$$r_{13} = \frac{n\Sigma u_1u_3 - \Sigma u_1\Sigma u_3}{\sqrt{n\Sigma u_1^2 - (\Sigma u_1)^2}\sqrt{n\Sigma u_3^2 - (\Sigma u_3)^2}}$$

$$= \frac{10 \times 9 - (-10) \times 40}{\sqrt{10 \times 36 - (-10)^2}\sqrt{10 \times 358 - (40)^2}}$$

$$= 0.683$$

$$r_{23} = \frac{n\Sigma u_2u_3 - \Sigma u_2\Sigma u_3}{\sqrt{n\Sigma u_2^2 - (\Sigma u_2)^2}\sqrt{n\Sigma u_3^2 - (\Sigma u_3)^2}}$$

$$= \frac{10 \times 248 - 10 \times 40}{\sqrt{10 \times 322 - (-10)^2}\sqrt{10 \times 358 - (40)^2}}$$

$$= 0.836$$

Now,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.577 - 0.683 \times 0.836}{\sqrt{1 - (0.683)^2}\sqrt{1 - (0.836)^2}}$$

$$= 0.014$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

$$= \frac{0.683 - 0.577 \times 0.836}{\sqrt{1 - (0.577)^2}\sqrt{1 - (0.836)^2}}$$

$$= 0.447$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.577)^2 + (0.683)^2 - 2 \times 0.577 \times 0.683 \times 0.836}{1 - (0.836)^2}}$$

$$= \sqrt{\frac{0.14}{0.3001}} = \sqrt{0.4665} = 0.683$$

Multiple Linear Regression

It is a linear function of one dependent variable with two or more independent variables. With the help of two or more independent variables the value of dependent variable is predicted. For example, if we wish to test the hypothesis that whether or not the 'pass grade' of students depends on many causes such as previous test mark, study hours, IQ, ...then we can test a regression of cause (pass grade) with effect variables. This test will give us which causes are really significant in generating effect variable and among the significant cause variables their relative value responsible to generate the effect variable. If we assume more than one causes (called X or independent variable) responsible for one effect (also called Y or dependent variable), it is known as multiple regression. If we assume that the relation between Y and X's is linear it is called multiple linear regression. However, there can be nonlinear relationship between Y and X's. For example, population growth (Y) is generally considers to have exponential relation with time and other cause variables.

Regression is used for two purpose. To get predicted value of Y for hypothetic X values. This is called prediction method and is more used for time dependent variables. For example, the future value of national income under similar conditions as existing. The other use of regression is to understand the role of cause variables on the generation of effect. It is called exploratory analysis and is more used for special data for example, the district data.

Let us consider three variables Y, X_1 and X_2 in which Y is dependent variable, X_1 and X_2 are independent variables, then the mathematical form of the linear relationship of Y with X_1 and X_2 is expressed as

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

Where,

Y = Dependent variable

X_1 and X_2 = Independent variable or explanatory variable or regressors

b_0 = Intercept and is called average value of Y when X_1 and X_2 are zero.

b_1 = Regression coefficient of Y on X_1 keeping X_2 constant. It measures the amount of change in Y per unit change in X_1 holding the X_2 constant.

b_2 = Regression coefficient of Y on X_2 keeping X_1 constant. It measures the amount of change in Y per unit change in X_2 holding the X_1 constant.

ε = Random error.

Random error (ε) is not created from mistake. It is a technical term that denoted the excess of value from real by model estimation. Error is also called Residual.

So, error = true value - estimated value from regression. Mathematically, $\varepsilon = Y - \hat{Y}$, where Y is the true value and $\hat{Y}$ is the estimate from regression. If we have 20 observations we will have 20 error values. By analyzing error or residual we can understand how the regression model fit to the given data, if assumptions such as linear is really usable, and other problems of the cause and effect variables. Such analysis is called Residual Analysis and is very useful diagnostic for regression.

Assumptions of Linear Regression

Theory of regression assumes that certain assumptions should hold for a reliable and acceptable regression analysis. If one or more assumptions are not satisfied or violated the regression will have specific problem. The major assumptions are as described below.

Let us consider multiple regression model

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

There are certain assumptions about the model. The assumptions are based on relation between error ε and explanatory variables x_i's.

i. Regression model is linear in parameters.

ii. ε is random real variable

iii. The random errors ε have zero mean, i.e. $E(\varepsilon) = 0$

iv. The random errors ε has constant variance ie. $E(\varepsilon) = \sigma^2$ (Noheteroscedaticity).

v. The random variable $\varepsilon$is normally distributed. i.e. $\varepsilon \sim N(0, \sigma^2)$

vi. The random errors ε are independent i.e. $E(\varepsilon_i\varepsilon_j) = 0 : i \neq j$. (No autocorrelation).

vii. X are uncorrelated to the error term ε, ie. $E(X\varepsilon) = 0$ (uniformity of X over samples

viii. The explanatory variables x_i's are measured without error.

ix. The number of observations must be greater than the number of explanatory variables.

x. The explanatory variables X_i's are not perfectly linearly correlated(No multicollinearity)

Estimation of Coefficients in Multiple Linear Regression

The linear relationship of dependent variable Y with explanatory variablesX_1 and X_2 is given by

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

Here b_0, b_1 and b_2 are called parameters of the three variable multiple regression equation.

Error $(e_i) = Y - b_0 - b_1X_1 - b_2X_2$ then $\Sigma e_i^2 = \Sigma(Y - b_0 - b_1X_1 - b_2X_2)^2$

By using the principle of least square by minimizing error sum of square, normal equations to estimate b_0, b_1 and b_2 are

$$\Sigma Y = nb_0 + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{......(i)}$$

$$\Sigma YX_1 = \Sigma b_0X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad \text{......(ii)}$$

$$\Sigma YX_2 = \Sigma b_0X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad \text{......(iii)}$$

Solving i, ii and iii we get, b_0 b_1 and b_2 then substitute values to get multiple regression equation.

$\hat{y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2$, where $\hat{b}_0$, $\hat{b}_1$ and $\hat{b}_2$ are estimated value of b_0, b_1 and b_2 respectively.

Regression equation of X_1 on X_2 and X_3:

Let X_1 be the dependent variable, X_2 and X_3 be the independent variables then the regression equation of X_1 on X_2 and X_3 be

$$X_1 = a + b_2X_2 + b_3X_3$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_2 and b_3 are

$$\Sigma X_1 = na + b_2\Sigma X_2 + b_3\Sigma X_3 \quad \text{.......(i)}$$

$$\Sigma X_1X_2 = a\Sigma X_2 + b_2\Sigma X_2^2 + b_3\Sigma X_2X_3 \quad \text{.......(ii)}$$

$$\Sigma X_1X_3 = a\Sigma X_3 + b_2\Sigma X_2X_3 + b_3\Sigma X_3^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_2 and b_3 and substitute values to get multiple regression equation.

Regression equation of X_2 on X_1 and X_3:

Let X_2 be the dependent variable, X_1 and X_3 be the independent variables then the regression equation of X_2 on X_1 and X_3 be

$$X_2 = a + b_1X_1 + b_3X_3$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_2 and b_3 are

$$\Sigma X_2 = na + b_1\Sigma X_1 + b_3\Sigma X_3 \quad \text{......(i)}$$

$$\Sigma X_1X_2 = a\Sigma X_1 + b_1\Sigma X_1^2 + b_3\Sigma X_1X_3 \quad \text{.......(ii)}$$

$$\Sigma X_2X_3 = a\Sigma X_3 + b_1\Sigma X_1X_3 + b_3\Sigma X_3^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_1 and b_3 and substitute values to get multiple regression equation

Regression equation of X_3 on X_1 and X_2:

Let X_3 be the dependent variable, X_1 and X_2 be the independent variables then the regression equation of X_3 on X_1 and X_2 be

$$X_3 = a + b_1X_1 + b_2X_2$$

By using the principle of least square by minimizing error sum of square, normal equations to estimate a, b_1 and b_2 are

$$\Sigma X_3 = na + b_1\Sigma X_1 + b_2\Sigma X_2 \quad \text{.......(i)}$$

$$\Sigma X_1X_3 = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \quad \text{.......(ii)}$$

$$\Sigma X_2X_3 = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 \quad \text{.......(iii)}$$

Solving i, ii and iii get a, b_1 and b_2 and substitute values to get multiple regression equation.

Example 9: Consider the following results obtained from a sample of 6;

$\Sigma x_1 = 487$, $\Sigma x_2 = 40$, $\Sigma y = 192$, $\Sigma x_1x_2 = 3346$, $\Sigma yx_1 = 15995$, $\Sigma yx_2 = 1390$, $\Sigma x_1^2 = 39901$, $\Sigma x_2^2 = 296$. Find the regression equation of y on x_1 and x_2. Estimate y when $x_1 = 83$ and $x_2 = 7$.

Solution:

Regression equation of y on x_1 and x_2 is $y = b_0 + b_1x_1 + b_2x_2$ ..... (i)

To estimate b_0, b_1 and b_2

$$\Sigma y = nb_0 + b_1\Sigma x_1 + b_2\Sigma x_2$$

or $$192 = 6b_0 + 487b_1 + 40b_2 \quad \text{....(ii)}$$

$$\Sigma yx_1 = b_0\Sigma X_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2$$

or $$15995 = 487b_0 + 39901b_1 + 3346b_2 \quad \text{....(iii)}$$

$$\Sigma yx_2 = b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma x_2^2$$

or $$1390 = 40b_0 + 3346b_1 + 296b_2 \quad \text{....(iv)}$$

Using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
6	487	40	192
487	39901	3346	15995
40	3346	296	1390

$$D = \begin{vmatrix} 6 & 487 & 40 \\ 487 & 39901 & 3346 \\ 40 & 3346 & 296 \end{vmatrix}$$

$$= 6(39901 \times 296 - 3346 \times 3346) - 487(487 \times 296 - 40 \times 3346) + 40(487 \times 3346 - 40 \times 39901)$$

$$= 6416$$

$$D_1 = \begin{vmatrix} 192 & 487 & 40 \\ 15995 & 39901 & 3346 \\ 1390 & 3346 & 296 \end{vmatrix}$$

$$= -352100$$

$$D_2 = \begin{vmatrix} 6 & 192 & 40 \\ 487 & 15995 & 3346 \\ 40 & 1390 & 296 \end{vmatrix}$$

$$= 6776$$

$$D_3 = \begin{vmatrix} 6 & 487 & 192 \\ 487 & 39901 & 15995 \\ 40 & 3346 & 1390 \end{vmatrix}$$

$$= 1114$$

Now,

$$b_0 = \frac{D_1}{D} = \frac{-352100}{6416} = -54.878$$

$$b_1 = \frac{D_2}{D} = \frac{6776}{6416} = 1.056$$

$$b_2 = \frac{D_3}{D} = \frac{1114}{6416} = 0.173$$

Substitute value in equation I we get

$$y = -54.878 + 1.056x_1 + 0.173x_2$$

When $x_1 = 83$ and $x_2 = 7$

$$y = -54.878 + 1.056 \times 83 + 0.173 \times 7 = 33.981$$

Example 10: The following information has been gathered from a random sample of apartment renters in a city. We are trying to predict rent (in dollars per month) based on the size of apartment (number of rooms) and the distance from downtown (in miles)

Rent (Dollar)	360	1000	450	525	350	300
Number of rooms	2	6	3	4	2	1
Distance from downtown	1	1	2	3	10	4

(i) Obtain the multiple regression models that best relate these variables (ii) Interpret the obtained regression coefficients. (iii) If some one is looking for a two bed apartment 2 miles from down town, what rent should he expect to pay?

Solution:

Here, Rent depends upon the number of rooms and distance from downtown.

Let rent = y, number of rooms = x_1 and distance from down town = x_2 then we have to find the regression equation of y on x_1 and x_2.

Rent (y) Dollar	No of rooms (x_1)	Distance (x_2)	x_1^2	x_2^2	yx_1	yx_2	x_1x_2
360	2	1	4	1	720	360	2
1000	6	1	36	1	6000	1000	6
450	3	2	9	4	1350	900	6
525	4	3	16	9	2100	1575	12
350	2	10	4	100	700	3500	20
300	1	4	1	16	300	1200	4
$\Sigma y=2985$	$\Sigma x_1=18$	$\Sigma x_2=21$	$\Sigma x_1^2=70$	$\Sigma x_2^2=131$	$\Sigma yx_1=11170$	$\Sigma yx_2=8535$	$\Sigma x_1x_2=50$

To fit $y = b_0 + b_1x_1 + b_2x_2$

or $\Sigma y = nb_0 + b_1\Sigma x_1 + b_2\Sigma x_2$

$$2985 = 6b_0 + 18b_1 + 21b_2 \quad \text{....(i)}$$

or $\Sigma yx_1 = b_0\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2$

$$11170 = 18b_0 + 70b_1 + 50b_2 \quad \text{....(ii)}$$

or $\Sigma yx_2 = b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma X_2^2$

$$8535 = 21b_0 + 50b_1 + 131\, b_2 \quad \text{....(iii)}$$

Using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
6	18	21	2985
18	70	50	11170
21	50	131	8535

Now,

$$D = \begin{vmatrix} 6 & 18 & 21 \\ 18 & 70 & 50 \\ 21 & 50 & 131 \end{vmatrix}$$

$$= 6(9170-2500) - 18(2358-1050) + 21(900-1470) = 4506$$

$$D_1 = \begin{vmatrix} 2985 & 18 & 21 \\ 11170 & 70 & 50 \\ 8535 & 50 & 131 \end{vmatrix}$$

$$= 2985(9170 - 2500) - 18(1463270 - 426750) + 21(558500 - 597450) = 434640$$

$$D_2 = \begin{vmatrix} 6 & 2895 & 21 \\ 18 & 11170 & 50 \\ 21 & 8535 & 131 \end{vmatrix}$$

$$= 6(1463270 - 426750) - 2985(2358 - 1050) + 21(153630 - 234570) = 615000$$

$$D_3 = \begin{vmatrix} 6 & 18 & 2985 \\ 18 & 70 & 11170 \\ 21 & 50 & 8535 \end{vmatrix}$$

$$= 6(597450 - 558500) - 18(153630 - 234570) + 2985(900 - 1470) = -10830$$

$$b_0 = \frac{D_1}{D} = \frac{434640}{4506} = 96.458,$$

$$b_1 = \frac{D_2}{D} = \frac{615000}{4506} = 136.484,$$

$$b_2 = \frac{D_3}{D} = \frac{-10830}{4506} = -2.403$$

Substituting values in regression equation

(i) $y = 96.458 + 136.484x_1 - 2.403x_2$

(ii) $b_1 = 136.484$ means on average rent is increased by 136.484 when room is increased by 1 holding the effect of distance from down town constant.

$b_2 = -2.403$ means average rent is decreased by 2.403 when the distance from downtown is increased by 1 holding the effect of number of room constant.

(iii) When $x_1 = 2$ and $x_2 = 2$,

$y = 96.458 + 136.484x_1 - 2.403x_2$

$= 96.458 + 136.484 \times 2 - 2.403 \times 2 = 364.62$

Expected rent for two bed room apartment 2 miles from downtown is 364.62 dollar.

Measures of Variation

In regression model value of dependent variable are estimated on the basis of independent variables. In regression analysis total variation is divided into explained variation (sum of square due to regression) and unexplained variation (sum of square due to error). Hence according to Fisher total sum of square is decomposed into sum of square due to regression and sum of square due to error (residual).

Total sum of square (TSS) = Sum of square due to regression (SSR) + Sum of square due to error (SSE)

For regression model $Y = b_0 + b_1X_1 + b_2X_2$, where Y is dependent variable, X_1 and X_2 are independent (explanatory) variables

$TSS = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$

$SSE = \Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma YX_1 - b_2\Sigma YX_2$

$SSR = TSS - SSE$

For regression model $x_1 = a + b_2x_2 + b_3x_3$, where x_1 is dependent variable and x_2, x_3 are independent variables

$TSS = \Sigma(x_1 - \bar{x}_2)^2 = \Sigma x_1^2 - n\bar{x}_1^2$

$SSE = \Sigma(x_1 - \hat{x}_1)^2 = \Sigma x_1^2 - a\Sigma x_1 - b_2\Sigma x_1x_2 - b_3\Sigma x_1x_3$

$SSR = TSS - SSE$

For regression model $x_2 = a + b_1x_1 + b_3x_3$, where x_2 is dependent variable and x_1, x_3 are independent variables

$TSS = \Sigma(x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - n\bar{x}_2^2$

$SSE = \Sigma(x_2 - \hat{x}_2)^2 = \Sigma x_2^2 - a\Sigma x_2 - b_1\Sigma x_1x_2 - b_3\Sigma x_2x_3$

$SSR = TSS - SSE$

For regression model $x_3 = a + b_1x_1 + b_2x_2$, where x_3 is dependent variable and x_1, x_2 are independent variables

$TSS = \Sigma(x_3 - \bar{x}_3)^2 = \Sigma x_3^2 - n\bar{x}_3^2$

$SSE = \Sigma(x_3 - \hat{x}_3)^2 = \Sigma x_3^2 - a\Sigma x_3 - b_1\Sigma x_1x_3 - b_2\Sigma x_2x_3$

$SSR = TSS - SS\,E$

ANOVA table of regression analysis

Source of variation(S.V.)	Degree of freedom (df)	Sum of square (SS)	Mean square(MS) (Variance)
Regression	k(no of independent variable)	SSR	MSR = SSR/k
Error	n-k-1	SSE	MSE = SSE/n-k-1
Total	n-1	TSS	

Standard Error of the Estimate

Standard error is the Square root of the variance computed from sample data. The standard error of the estimate measures the average variation or scatterness of the observed data point around regression line. Standard error of the estimate is used to measure the reliability of the regression equation. Regression line having less standard error of estimate is more reliable than regression line having more standard error of estimate.

It is given by $S_e = \sqrt{\dfrac{SSE}{n - k - 1}}$

SSE = sum of square due to error

k = number of independent variable in regression model

n = number of observations.

When $S_e = 0$, there is no variation of observed data around regression line. In such case regression line is perfect for estimating the dependent variable.

Coefficient of Determination

It measures the proportion of variation in dependent variable that is explained by the set of independent variables .It is the measure based upon measure of variation and is used to determine the fitness of the data to the model. The regression line is reliable if the sum of square due to regression is much greater than sum of square due to error. It is the ratio of sum of square due to regression to the total sum of square. It is denoted by R^2 and is given by, $R^2 = \dfrac{SSR}{TSS}$

It is also obtained by simply squaring the correlation coefficient i.e., $R^2 = r^2$. Higher the value of R^2 the more reliable is the fitted equation .It lies between 0 and 1.

R^2 can never decrease when another independent variable is added to a regression. R^2 will usually increase with increase in number of independent variables.

It is suggested that the adjusted R^2 should be used in place of R^2 in multiple regression model. Adjusted R^2 is simply a R^2 adjusted by its degree of freedom and reflects both the number of independent variables and sample size used in the model. Adjusted R^2 is considered as an important measure for the comparising of two or more regression models that predict same dependent variable with different number of independent variables.

$R^2_{adjusted}\,(\bar{R}^2) = 1 - \dfrac{(n - 1)}{(n - k - 1)}[1 - R^2]$; where n = no of pair of observations, k = no of independent variables.

Example 11: A health research team collects data on ten communities. Measurement are obtained on the following variable.

y = Health care facility utilization

x_1 = Median family income

x_2 = Proportion of worker with health insurance

x_3 = Doctor population ratio.

Source of variation	Sum of square	df
Regression	?	3
Error	88.66	?
Total	476.9	9

(i) Complete the table

(ii) Compute R^2 and interpret

(iii) Compute adjusted R^2

(iv) Compute standard error of estimate.

Solution:

Here

$SSE = 88.66$, $TSS = 476.9$, $k = 3$, $n-1 = 9$

df for error = n-k-1 = 9-3 =6

$SSR = TSS - SSE$

$= 476.9 - 88.66 = 388.24$

$$R^2 = \frac{SSR}{TSS}$$

$$= \frac{388.24}{476.9} = 0.814 = 81.4\%$$

It means 81.5% of the total variation in health care facility utilization can be explained by the variation in median family income, proportion of worker with health insurance and doctor population ratio.

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{(n-k-1)}[1 - R^2]$$

$$= 1 - \frac{9}{6}(1 - 0.814)$$

$$= 1 - 0.279 = 0.721$$

$$S_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{88.66}{6}} = 3.84.$$

Example 12: Find $S_{e(1.23)}$, $R_{1.23}^2$ on the basis of following information:

$\Sigma x_1 = 272$, $\Sigma x_2 = 441$, $\Sigma x_3 = 147$, $\Sigma x_1x_2 = 12005$, $\Sigma x_1x_3 = 4013$, $\Sigma x_2x_3 = 6485$, $\Sigma x_1^2 = 7428$, $\Sigma x_2^2 = 19461$, $\Sigma x_3^2 = 2173$, $n = 10$.

Solution:

We have to find the regression equation of x_1 on x_2 and x_3

$x_1 = a + b_2x_2 + b_3x_3$

To estimate a, b_2 and b_3

$\Sigma x_1 = na + b_2\Sigma x_2 + b_3\Sigma x_3$

Or $272 = 10a + 441b_2 + 147b_3$ ......(i)

$\Sigma x_1x_2 = a\Sigma x_2 + b_2\Sigma x_2^2 + b_3\Sigma x_2x_3$

Or $12005 = 441a + 19461b_2 + 6485b_3$ ......(ii)

$\Sigma x_1x_3 = a\Sigma x_3 + b_2\Sigma x_2x_3 + b_3\Sigma x_3^2$

$$4013 = 147a + 6485b_2 + 2173b_3 \qquad \ldots\ldots(iii)$$

To find a, b_2 and b_3 using Cramer's rule

Coefficient of a	Coefficient of b_2	Coefficient of b_3	Constant
10	441	147	272
441	19461	6485	12005
147	6485	2173	4013

$$\text{Now, } D = \begin{vmatrix} 10 & 441 & 147 \\ 441 & 19461 & 6485 \\ 147 & 6485 & 2173 \end{vmatrix}$$

$$= 10(42288753 - 42055225) - 441(958293 - 953295) + 147(2859885 - 2860767)$$

$$= 1508$$

$$D_1 = \begin{vmatrix} 272 & 441 & 147 \\ 12005 & 19461 & 6485 \\ 4013 & 6485 & 2173 \end{vmatrix}$$

$$= 272(42288753 - 42055225) - 441(26086865 - 26024305) + 147(77852425 - 78096993)$$

$$= -20840$$

$$D_2 = \begin{vmatrix} 10 & 272 & 147 \\ 441 & 12005 & 6485 \\ 147 & 4013 & 2173 \end{vmatrix}$$

$$= 10(26086865 - 26024305) - 272(958293 - 953295) + 147(1769733 - 1764735)$$

$$= 850$$

$$D_3 = \begin{vmatrix} 10 & 441 & 272 \\ 441 & 19461 & 12005 \\ 147 & 6485 & 4013 \end{vmatrix}$$

$$= 10(78096993 - 77852425) - 441(1769733 - 1764735) + 272(2859885 - 2860767)$$

$$= 1658$$

Now

$$a = \frac{D_1}{D} = \frac{-20840}{1508} = -13.819$$

$$b_2 = \frac{D_2}{D} = \frac{850}{1508} = 0.563$$

$$b_3 = \frac{D_3}{D} = \frac{1658}{1508} = 1.099$$

$$\bar{x}_1 = \frac{\Sigma X_1}{n} = \frac{272}{10} = 27.2$$

$$SSE(X_{1.23}) = \Sigma x_1^2 - a\Sigma x_1 - b_2\Sigma x_1x_2 - b_3\Sigma x_1x_3$$

$$= 7428 - (-13.819) \times 272 - 0.563 \times 12005 - 1.099 \times 4013 = 17.661$$

$$TSS = \Sigma x_1^2 - n\bar{x}_1^2$$

$= 7428 - 10 \times (27.2)^2 = 29.6$

$SSR = TSS - SSE$

$= 29.6 - 17.661 = 11.939$

$$S_{e(1\cdot 23)} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{17.661}{10-2-1}} = 1.58$$

$$R^2_{1.23} = \frac{SSR}{TSS} = \frac{11.939}{29.6} = 0.403$$

Test of Significance for Regression Coefficients

To test the significance of the individual regression coefficients t test is used. It helps to determine whether there is significant linear relationship between dependent variable and independent variable.

Let us consider regression equation

$y = b_0 + b_1x_1 + b_2x_2$, for multiple regression equation of three variables. Where y is dependent variable; x_1, x_2 are independent variables, b_0 constant value, b_1 is regression coefficient of y on x_1 keeping x_2 constant, b_2 is regression coefficient of y on x_2 keeping x_1 constant.

Let β_1 and β_2 be the population regression coefficients of the sample regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

Different steps in the test are

Problem to test

$H_0 : \beta_i = 0$ (There is no linear relationship between dependent variable y and independent variable x_i, $i = 1, 2$).

$H_1: \beta_i \neq 0$.

Test statistic

$t = \frac{b_i}{Sb_i} \sim$ t distribution with n-k-1 degree of freedom, n = no of observation and k = no of independent variables

Where b_i = sample regression coefficient and Sb_i = Standard error of regression coefficient

Level of significance

Let α be the level of significance. Usually we take $\alpha = .05$ unless we are given.

Critical value

Critical or tabulated value of t is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision:

Reject H_0 at α level of significance if $|t| > t_{tabulated}$, accept otherwise.

Confidence interval for regression coefficient

At α% level of significance for n-k-1 degree of freedom the critical value of t is $t_{\alpha/2\,(n-k-1)}$, then (100 - α%) confidence or fudicial limits for regression coefficient β_i is given by $b_i \pm t_{\alpha/2\,(n-k-1)} Sb_i$.

Example 13: To study the effect of age (x_1 in years) and weight (x_2 in lbs) on systolic blood pressure (y mm in Hg), the data were recorded for a sample of 15 adult males. The estimated regression model based on data is described below where figures within parenthesis are standard error of estimate.

$$y = \underset{(24.68)}{27.4} + \underset{(0.248)}{0.221}\, x_1 + \underset{(0.155)}{0.56}\, x_2$$

Test the significance of regression coefficients at 1% level of significance.

Solution:

Here,

Sample size (n) = 15, Number of independent variable (k) = 2, $b_0 = 27.4$, $b_1 = 0.221$, $b_2 = 0.56$, $Sb_0 = 24.68$, $Sb_1 = 0.248$, $Sb_2 = 0.115$, $\alpha = 1\%$.

Let β_1 and β_2 be the population regression coefficients.

For the first regression coefficient

Problem to test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Test statistic

$$t = \frac{b_1}{Sb_1} = \frac{0.221}{0.248} = 0.89$$

Critical value

At $\alpha = 0.01$ level of significance, critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2(n-k-1)} = 3.055$.

Decision

$t = 0.89 < t_{tabulated} = 3.055$, accept H_0 at 5% level of significance.

Conclusion

There is no significant linear relationship between y and x_1.

For the second regression coefficient

Problem to test

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Test statistic

$$t = \frac{b_2}{Sb_2} = \frac{0.56}{0.115} = 4.869.$$

Critical value

At $\alpha = 0.01$ level of significance, critical value for two tailed test is

$t_{tabulated} = t_{\alpha/2(n-k-1)} = 3.055$

Decision

$t = 4.869 > t_{tabulated} = 3.055$, reject H_0 at 5% level of significance.

Conclusion

There is a significant linear relationship between y and x_2.

Test of Overall Significance of the Regression Coefficients

To test the significance of over all regression coefficients F test is used. It helps to determine whether there is significant linear relationship between the dependent variable and the set of independent variables.

Let us consider regression equation

$y = b_0 + b_1 x_1 + b_2 x_2$, for multiple regression equation of three variables. Where y is dependent variable; x_1, x_2 are independent variables, b_0 constant value, b_1 is regression coefficient of y on x_1 keeping x_2 constant, b_2 is regression coefficient of y on x_2 keeping x_1 constant.

Let β_1 and β_2 be the population regression coefficients of the sample regression equation $y = b_0 + b_1 x_1 + b_2 x_2$.

Different steps in the test are

Problem to test

$H_0 : \beta_1 = \beta_2 = 0$ (There is no linear relationship between dependent variable y and independent variables)

H_1 : At least one β_i is different from zero (i = 1, 2)

(There is linear relationship between the dependent variable and at least one independent variable)

Test statistic

$F = \frac{MSR}{MSE} \sim$ F distribution with (k, n-k-1) degree of freedom, where k = no of independent variables

MSR = mean sum of square due to regression and

MSE = mean sum of square due to error

ANOVA table for regression analysis

Source of variation (SV)	Degree of freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F	$F_{tabulated}$
Regression	k	SSR	MSR	F_R=MSR/MSE	$F_{\alpha\,(k,n-k-1)}$
Error	n-k-1	SSE	MSE		
Total	n-1	TSS			

$TSS = \Sigma(y - \bar{y})^2$, $SSE = \Sigma (y - \hat{y})^2$, SSR = TSS – SSE.

Level of significance

Let α be the level of significance. Usually we take α = .05 unless we are given.

Critical value

Critical or tabulated value of F is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision

Reject H_0 at α level of significance if $F > F_{tabulated}$, accept otherwise.

Relationship between F and R^2

We know,

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{(n-k-1)}{k} \times \frac{SSR}{SSE}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{SSE}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{(TSS-SSR)}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{\frac{SSR}{TSS}}{\frac{TSS}{TSS} - \frac{SSR}{TSS}}$$

$$= \frac{(n-k-1)}{k} \times \frac{R^2}{1-R^2}$$

Example 14: The following ANOVA summary table was obtained from a multiple regression model with two independent variables.

Source of variation	Degree of freedom	Sum of Square
Regression	2	30
Error	10	120
Total	12	150

Test the overall fit of the model at 0.05 level of significance.

Solution:

Here, $n - k - 1 = 12$, $k = 2$ $SSR = 30$, $SSE = 120$, $TSS = 150$, $\alpha = 0.05$

$n = 12 + k + 1 = 12 + 2 + 1 = 15$

$MSR = \frac{SSR}{k} = \frac{30}{2} = 15$, $MSE = \frac{SSE}{n-k-1} = \frac{120}{10} = 12.$

Problem to test

$H_0 : \beta_1 = \beta_2 = 0$

H_1 : At least one β_i is different from 0, $i = 1, 2$

Test statistic

$$F = \frac{MSR}{MSE} = \frac{15}{12} = 1.25$$

Critical value

At $\alpha = 0.05$ level of significance for one tailed test the critical value is $F_{\alpha(k,\ n-k-1)} = 3.89$

Decision: $F = 1.25 < F_{tabulated} = 3.89$, accept H_0 at 0.05 level of significance.

Conclusion: There is no significant relationship between dependent variable and two independent variables.

Example 15: To study the effect of age (x_1 in years) and weight (x_2 in lbs) on systolic blood pressure (y mm in Hg), the data were recorded for a sample of 15 adult males. The estimated regression model based on data is described below:

$$y = 27.4 + 0.221x_1 + 0.56x_2$$

Further computation shows that $\Sigma(y - \overline{y})^2 = 1835.7$ and $\Sigma(y - \hat{y})^2 = 1101.3$.

Carry out the overall goodness of fit test of the model at 5% level of significance.

Solution:

Here, Sample size (n) = 15, Number of independent variables (k) = 2

$b_0 = 27.4$, $b_1 = 0.221$, $b_2 = 0.56$, Level of significance (α) = 5%

$$TSS = \Sigma(y - \overline{y})^2 = 1835.7$$

$$SSE = \Sigma(y - \hat{y})^2 = 1101.3$$

$$SSR = TSS - SSE = 1835.7 - 1101.3 = 734.4,$$

$$MSR = \frac{SSR}{k} = \frac{734.4}{2} = 367.2$$

$$MSE = \frac{SSE}{n-k-1} = \frac{1101.3}{12} = 91.775$$

Problem to test

$H_0 : \beta_1 = \beta_2 = 0$

H_1 : At least one β_i is different from zero, $i = 1, 2$

Test statistic

$$F = \frac{MSR}{MSE} = \frac{367.2}{91.775} = 4.001$$

Critical value

At $\alpha = 0.05$ level of significance, critical value is $F_{\alpha(k,n-k-1)} = 3.89$.

Decision

$F = 4.001 > F_{tabulated} = 3.89$, reject H_0 at 5% level of significance.

Conclusion

There is linear relationship of dependent variable y with both the independent variables x_1 and x_2.

EXERCISE

1. What do you mean by partial correlation? Write down the relationship between partial and simple correlation coefficients.
2. What do you mean by multiple correlation? Write down the relationship between multiple correlation coefficient and simple correlation coefficients.
3. Write down the properties of multiple correlation coefficient.
4. Differentiate between partial and multiple correlation coefficient.
5. What is multiple regression? Write down the method of obtaining multiple regression line.
6. What are underlying assumptions of linear regression model?
7. What do you mean by standard error of estimate? Write down role of it in regression analysis.
8. What do you mean by coefficient of determination? How is it different from correlation coefficient?
9. If $r_{12} = 0.5$, $r_{23} = 0.1$ and $r = 0.4$ compute $r_{12.3}$ and $r_{13.2}$. **Ans:** 0.5, 0.46
10. For a trivariate distribution $r_{12} = 0.4$, $r_{23} = 0.5$ and $r_{13} = 0.6$. Find (i) $R_{1.23}$ (ii) $r_{23.1}$ (iii) $R_{1.23}^2$ (iv) $r_{23.1}^2$ and comment. **Ans:** 0.611, 0.35, 0.37, 0.125
11. Are the following data consistent; $r_{23} = 0.8$, $r_{31} = -0.5$, $r_{12} = 0.6$. **Ans:** inconsistent
12. From the data related to the yield of dry bark (x_1), height (x_2) and girth (x_3) for 18 cinchona plants the following correlation coefficient were obtained $r_{12} = 0.77$, $r_{13} = 0.72$, $r_{23} = 0.52$. Find the partial correlation coefficients. **Ans:** 0.63, 0.85, - 0.077
13. Suppose a computer has found for a given set of values x_1, x_2 and x_3: $r_{12} = 0.91$, $r_{13} = 0.33$, $r_{32} = 0.81$. Examine whether the computations may be said to be free from error? **Ans:** No
14. The following are zero order correlation coefficients $r_{12} = 0.8$, $r_{13} = 0.44$, $r_{23} = 0.54$. Calculate the partial correlation coefficient between first and third variables keeping the effect of second variable constant. **Ans:** 0.0158
15. Consider the following results obtained from a sample of 10 and x_1,x_2 and x_3 are measured in arbitrary unit $\Sigma x_1 = 10$, $\Sigma x_2 = 20$, $\Sigma x_3 = 30$, $\Sigma x_1^2 = 20$, $\Sigma x_2^2 = 68$, $\Sigma x_3^2 = 170$, $\Sigma x_1x_2 = 10$, $\Sigma x_1x_3 = 15$, $\Sigma x_2x_3 = 64$. Compute $r_{12.3}$ and $R_{1.23}$. **Ans:** -0.65, 0.76
16. From the information given below calculate $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$. **Ans:** 0.86, 0.62, -0.177

x_1	6	8	9	11	12	14
x_2	14	16	17	18	20	23
x_3	21	22	27	29	31	32

17. Given the following information from a multiple regression analysis;

$n = 20$, $b_1 = 4$, $b_2 = 3$, $Sb_1 = 1.2$, $Sb_2 = 0.8$. At 0.05 level of significance, determine whether each of explanatory (dependent) variable makes a significant contribution to the regression model.

Ans: $t = 3.33$, Sig. $t = 3.75$, Sig.

18. In order to establish the functional relationship between annual salaries(y), years of educated high school (x_1) and years of experience with the firm (x_2), data on these three variables were collected from a random sample of 10 persons working in a large firm. Analysis of data produces the following results. Total sum of squares $\Sigma(y - \bar{y})^2 = 397.6$.

Sum of squares due to error $\Sigma(y - \hat{y})^2 = 23.5$. Test the over all significance of regression coefficients at 5% level of significance.

Ans: $F = 55.83$, Sig.

19. Suppose you are given following information;

Multiple regression model $y = 5 + 18x_1 + 20x_2$, sample size $n = 28$

Total sum of squares (TSS) = 250

Sum of square due to error (SSE) = 100

Standard error of regression coefficient of x_1 (Sb_1) = 3.2

Standard error of regression coefficient of x_2 (Sb_2) = 5.5

Test the significance of regression coefficient of x_2 at 1% level of significance

Also test the over all significance of regression coefficients at 5% level of significance.

Ans: $t = 3.63$, Significant, $F = 18.75$, Significant

20. From following information of variables X_1, X_2 and X_3

$\Sigma X_1 = 13, \Sigma X_2 = 11, \Sigma X_3 = 51, \Sigma X_1^2 = 63, \Sigma X_2^2 = 95, \Sigma XX_3 = 77, \Sigma X_2X_3 = 136, \Sigma X_1X_2 = -240, n = 10, \Sigma X_3 = 450.$

(i) Find the regression equation of X_3 on X_1 and X_2 and interpret the regression coefficients.

(ii) Predict X_3 when $X_1 = 1$ and $X_2 = 4$.

(iii) Compute TSS, SSR and SSE

(iv) Compute standard error of estimate

(v) Compute the coefficient of multiple determination and interpret.

Ans: $X_3 = 1.008 + 1.676X_1 + 1.738X_2$, 9.636, 189.9, 156.72, 33.17, 2.17, 0.82

21. From the following information of three variables Y, X_1 and X_2

$\Sigma(y - \bar{y})^2 = 3450$, $\Sigma(y - \hat{Y})^2 = 365.7$, $\Sigma x_1x_2 = 5779$, $\Sigma yx_2 = 6796$, $\Sigma yx_1 = 40830$, $\Sigma x_1 = 643$, $\Sigma x_2 = 106$, $n = 12$

$\Sigma y^2 = 48139$, $\Sigma x_1^2 = 3483$, $\Sigma x_2^2 = 976$, $\Sigma y = 753$, $\Sigma x_1 = 643$, $\Sigma x_2 = 106$, $n = 12$

(i) Find the least square regression of y on x_1 and x_2.

(ii) Find the standard error of estimate.

(iii) find the coefficient of multiple determination.

Ans: $y = 30.69 - 0.0038x_1 + 3.652x_2$, 6.37, 0.89

22. The table shows the corresponding values of the three variables X_1, X_2 and X_3

X_1:	5	7	8	6	10	9
X_2:	12	20	30	40	33	25
X_3:	51	55	58	60	70	66

Find the regression equation of X_1 on X_2 and X_3. Estimate X_1 when $X_2 = 50$ and $X_3 = 100$. Where X_1 represents pull strength, X_2 represents wire length and X_3 represents die height.

Ans: $X_1 = -7.862 - 0.048X_2 + 0.277X_3$, 19.78

23. From the following set of data (i) find the multiple regression equation (ii) Interpret the regression coefficients (iii) Predict y when $X_1 = -10$ and $X_2 = 4$.

Y:	6	10	9	14	7	5
X_1:	1	3	2	-2	3	6
X_2:	3	-1	4	7	2	-4

Ans: $Y = 12.425 - 1.487X_1 - 0.383X_2$, 25.76

24. A developer of food for pig would like to determine what relationship exists among the age of a pig when it starts receiving a newly developed food supplement, the initial weight of the pig and the amount of weight it gains in a week period with the food supplement. The following information is the result of study of eight piglets.

Piglet number	Initial weight (pounds) x_1	Initial age (weeks) x_2	Weight gain y
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

(i) Calculate the least square equation that best describes these three variables.

(ii) Calculate the standard error of estimate.

(iii) How much might we expect a pig to gain weight in a week with the food supplement if it were 9 weeks old and weighted 48 pounds?

Ans: $Y = -3.66 + 0.105X_1 + 0.732x_2$, 1.25, 8

Using Software

Regression Analysis

A developer of food for pig wish to determine what relationship exists among 'age of a pig' when it starts receiving a newly developed food supplement, the initial weight of the pig and the amount of weight it gains in a week period with the food supplement. The following information is the result of study of eight piglets.

Piglet number	Initial weight(pounds)x_1	Initial age (weeks) x_2	Weight gain y
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

I. Determine the least square equation that best describes these three variables.

II. Calculate the standard error.

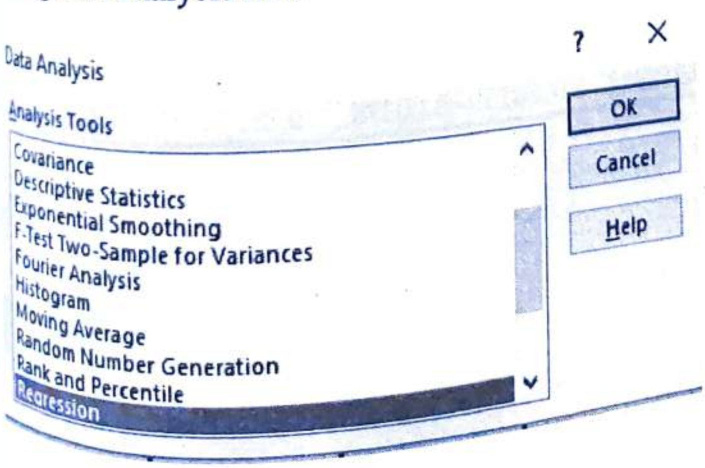
III. How much gain in weight of a pig in a week can we expect with the food supplement if it were 9 weeks old and weighed 48 pounds?

IV. Test the significance of regression coefficients and overall fit of the regression equation

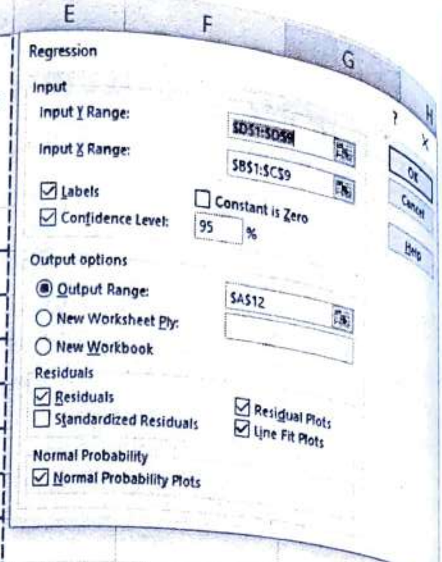
V. Conduct the residual analysis

VI. Determine partial correlations, multiple correlation and coefficient of multiple determination. Interpret.

Using **data analysis tool**



	A	B	C	D
1	Piglet number	Initial weight(pounds)x_1	Initial age (weeks) x_2	Weight gain y
2	1	39	8	7
3	2	52	6	6
4	3	49	7	8
5	4	46	12	10
6	5	61	9	9
7	6	35	6	5
8	7	25	7	3
9	8	55	4	4



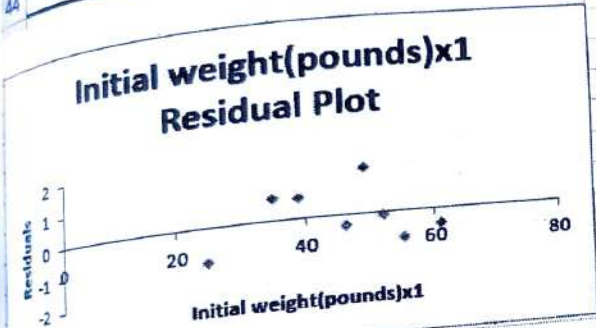
	A	B	C	D	E	F	G
12	SUMMARY OUTPUT						
13							
14	*Regression Statistics*						
15	Multiple R	0.93870818					
16	R Square	0.88117304					
17	Adjusted R Square	0.83364226					
18	Standard Error	0.99907279					
19	Observations	8					
20							
21	ANOVA						
22		*df*	*SS*	*MS*	*F*	*Significance F*	
23	Regression	2	37.00927	18.50463	18.539	0.004867292	
24	Residual	5	4.990732	0.998146			
25	Total	7	42				
26							
27		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
28	Intercept	-4.1917094	1.888119	-2.22004	0.077124	-9.045274309	0.661855502
29	Initial weight(pounds)x1	0.10483433	0.032291	3.246502	0.022784	0.021826458	0.187842193
30	Initial age (weeks) x2	0.80650253	0.158237	5.096815	0.00378	0.399742475	1.213262588
31							

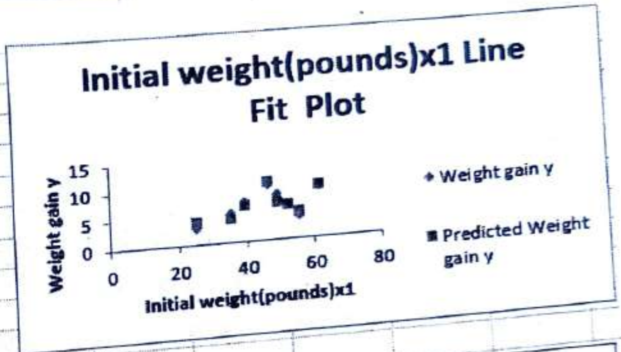
RESIDUAL OUTPUT

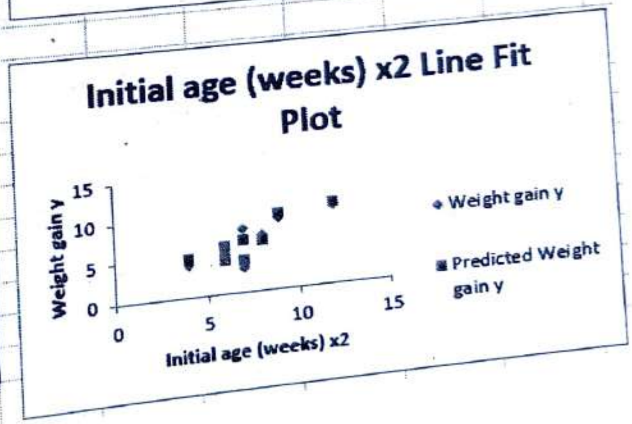
Observation	Predicted Weight gain y	Residuals
1	6.34884955	0.65115
2	6.09869072	-0.09869
3	6.59069027	1.40931
4	10.3087	-0.3087
5	9.46170724	-0.46171
6	4.31650718	0.683493
7	4.07466646	-1.07467
8	4.80018863	-0.80019

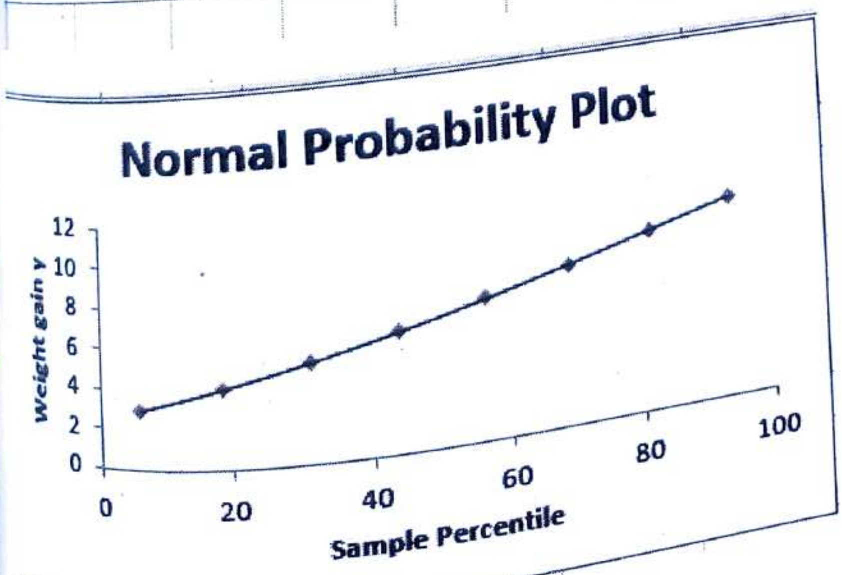
PROBABILITY OUTPUT

Percentile	Weight gain y
6.25	3
18.75	4
31.25	5
43.75	6
56.25	7
68.75	8
81.25	9
93.75	10









Here:

1 The regression equation of weight gain on Initial weight(pounds) and Initial age (weeks) is:

y = (-4.1917) + (0.1048)x1 + (0.8065)x2

2 Standard error = 0.9991

3 Weight gain is 35.4639 units

4 For testing null hypothesis B0 = 0, since p value = 0.077. It is insignificant

For testing null hypothesis B1 = 0, since p value = 0.023. It is significant

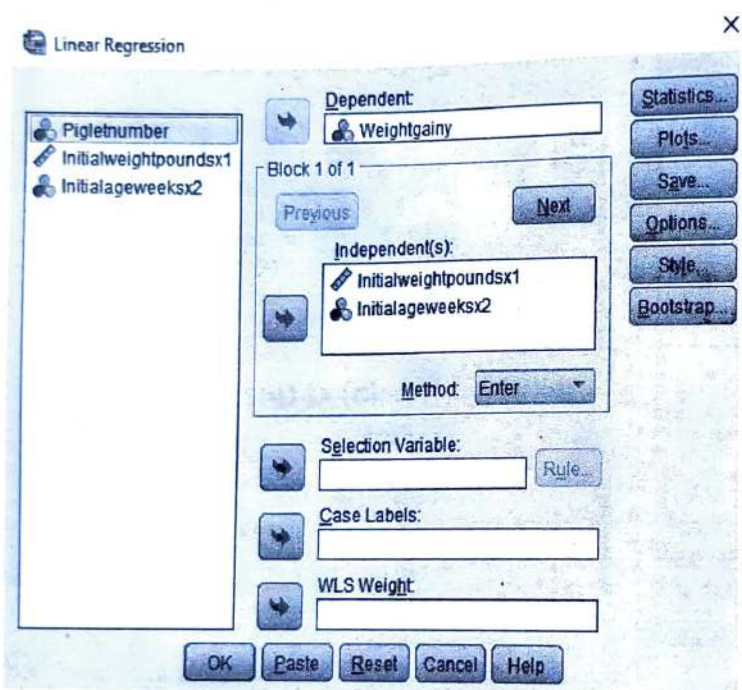
For testing null hypothesis B2 = 0, since p value = 0.004. It is significant

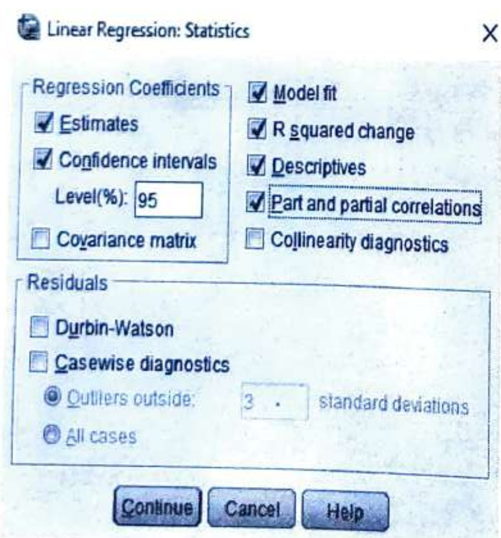
For testing null hypothesis: overall fit of the regression coefficients =0, since here the p value = 0.0048 for F test, that indicates overall fit is significant

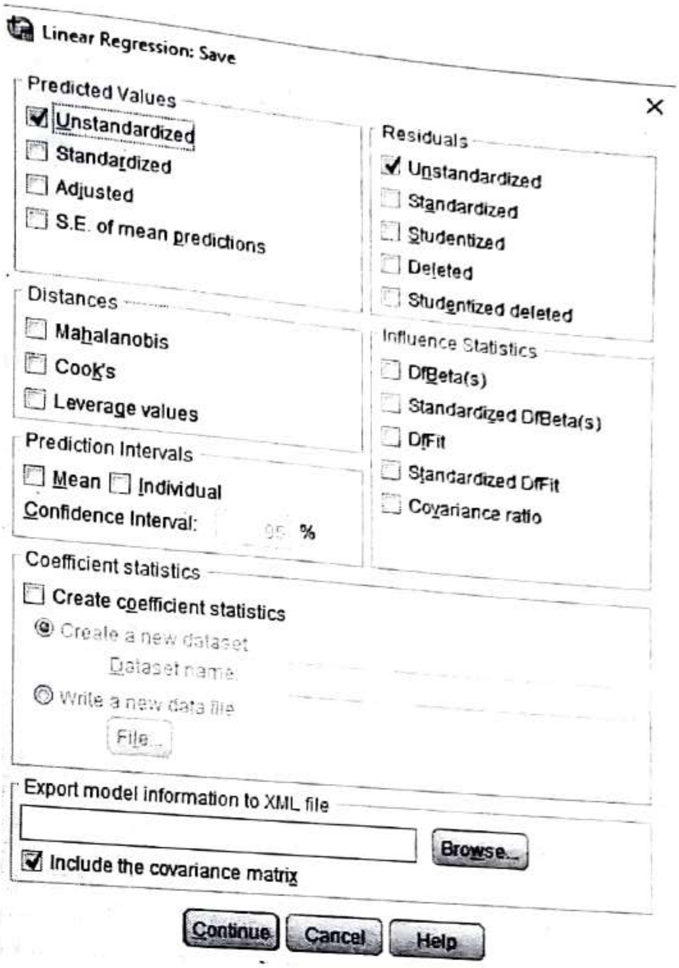
5 Here

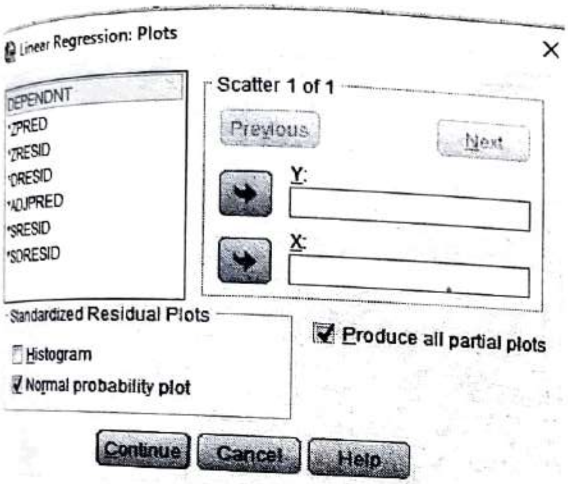
Adj R2= 0.8336. That indicates this regression equation can represent 83.36% of the true observations.

How to use SPSS for Regression









Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Weight gain y	6.50	2.449	8
Initial weight(pounds)x1	45.25	11.696	8
Initial age (weeks) x2	7.38	2.387	8

Correlations

		Weight gain y	Initial weight (pounds)x1	Initial age (weeks) x2
Pearson Correlation	Weight gain y	1.000	.514	.794
	Initial weight(pounds)x1	.514	1.000	.017
	Initial age (weeks) x2	.794	.017	1.000
Sig. (1-tailed)	Weight gain y	.	.096	.009
	Initial weight(pounds)x1	.096	.	.484
	Initial age (weeks) x2	.009	.484	.
N	Weight gain y	8	8	8
	Initial weight(pounds)x1	8	8	8
	Initial age (weeks) x2	8	8	8

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Initial age (weeks) x2, Initial weight (pounds)x1[b]	.	Enter

a. Dependent Variable: Weight gain y

b. All requested variables entered.

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.939[a]	.881	.834	.999	.881	18.539	2	5	.005

a. Predictors: (Constant), Initial age (weeks) x2, Initial weight(pounds)x1

b. Dependent Variable: Weight gain y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37.009	2	18.505	18.539	.005[b]
	Residual	4.991	5	.998		
	Total	42.000	7			

a. Dependent Variable: Weight gain y

b. Predictors: (Constant), Initial age (weeks) x2, Initial weight(pounds)x1

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations		
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-4.192	1.888		-2.220	.077	-9.045	.662			
	Initial weight(pounds)x1	.105	.032	.501	3.247	.023	.022	.188	.514	.824	.500
	Initial age (weeks) x2	.807	.158	.786	5.097	.004	.400	1.213	.794	.916	.786

a. Dependent Variable: Weight gain y

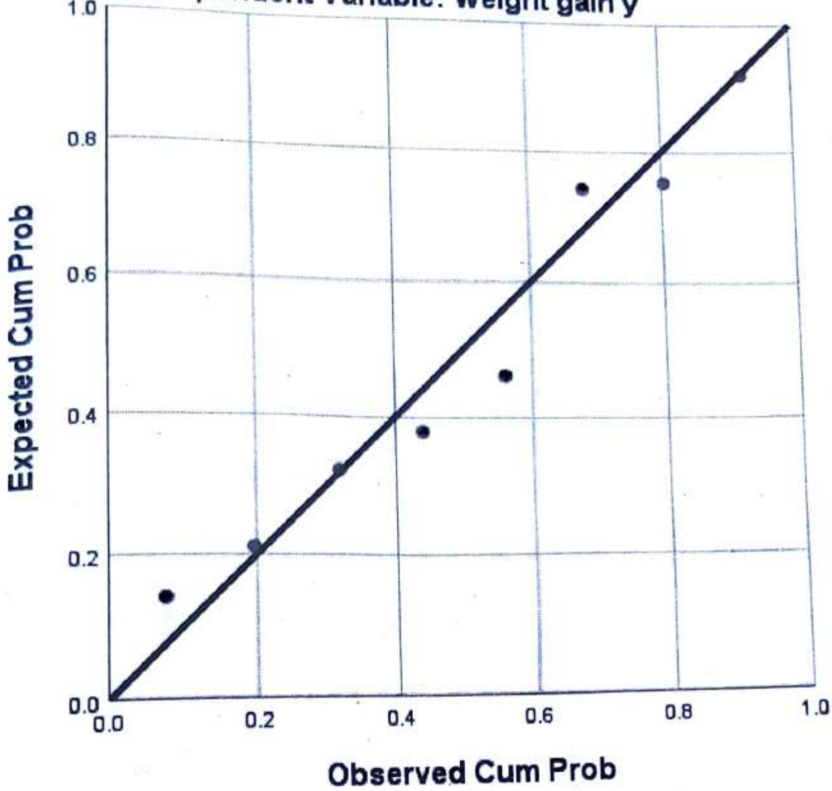
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.07	10.31	6.50	2.299	8
Residual	-1.075	1.409	.000	.844	8
Std. Predicted Value	-1.055	1.656	.000	1.000	8
Std. Residual	-1.076	1.411	.000	.845	8

a. Dependent Variable: Weight gain y

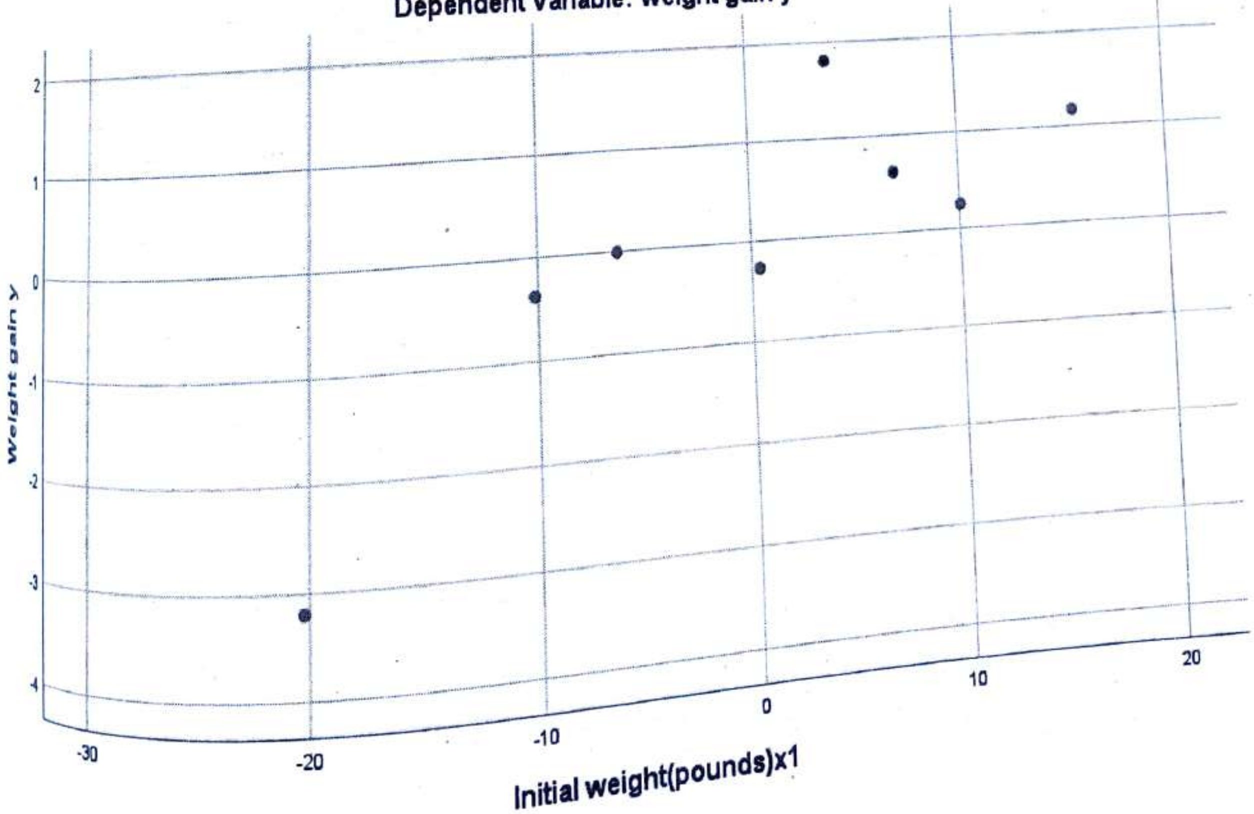
Normal P-P Plot of Regression Standardized Residual

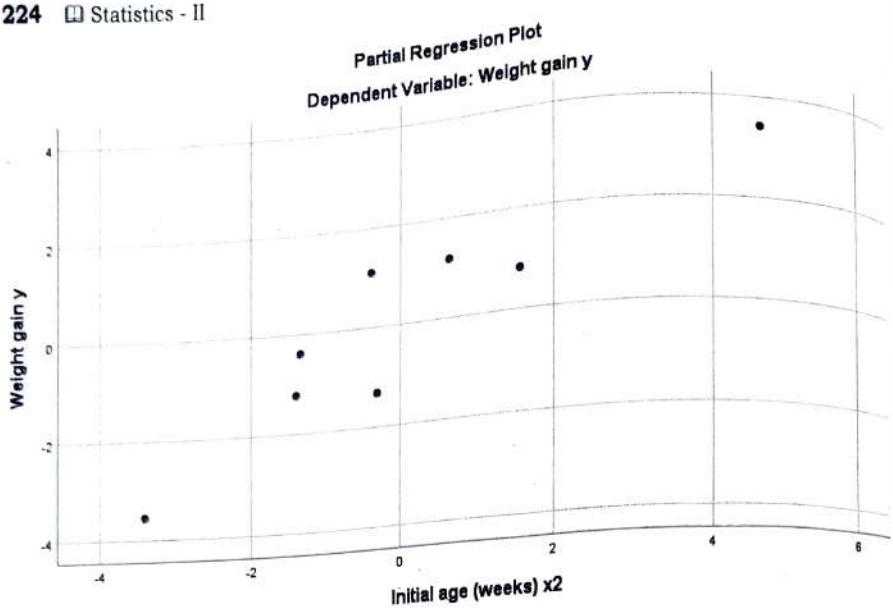
Dependent Variable: Weight gain y



Partial Regression Plot

Dependent Variable: Weight gain y





How to use STATA for Regression

(variable names replaced by y=Weightgainy, x1= Initialweightpoundsx1, x2= Initialageweeksx2)
STATA commands shown in the output display

```
. reg y x1 x2

      Source |       SS           df       MS      Number of obs   =         8
-------------+----------------------------------   F(2, 5)         =     18.54
       Model |  37.0092678         2  18.5046339   Prob > F        =    0.0049
    Residual |  4.99073219         5  .998146438   R-squared       =    0.8812
-------------+----------------------------------   Adj R-squared   =    0.8336
       Total |          42         7           6   Root MSE        =    .99907

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1048343   .0322915     3.25   0.023     .0218265    .1878422
          x2 |   .8065025   .1582366     5.10   0.004     .3997425    1.213263
       _cons |  -4.191709   1.888119    -2.22   0.077    -9.045274    .6618555
------------------------------------------------------------------------------

. display _b[_cons] + _b[x1]*9+ _b[x2]*48
35.463921
```

1. The regression equation of weight gain on Initial weight(pounds) and Initial age (weeks) is:

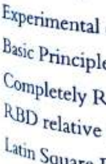
y = (-4.1917) + (0.1048)x1 + (0.8065)x2

2. Standard error (Root MSE)= 0.9991
3. Weight gain is 35.4639 units (the display command)
4. Look at the regression output, P >|t| and for F test, Prob > F
5. Adj R2= 0.8336.

□□□

5
Chapter

DESIGN OF EXPERIMENT



CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Experimental design
- Basic Principles of experimental designs
- Completely Randomized Design (CRD), Randomized Block Design (RBD), ANOVA table, Efficiency of RBD relative to CRD, Estimations of missing value (one observation only), Advantages and disadvantages
- Latin Square Design (LSD), Statistical analysis of m × m LSD for one observation per experimental unit, ANOVA table, Estimation of missing value in LSD (One observation only)
- Efficiency of LSD relative to RBD, Advantage and disadvantages
- Problems and illustrative examples related using software.

Experimental Design

In any experiment the prime objective is either to maximize the precision or to minimize the error. Hence the validity of the experiment depends upon how adequately it has been designed. It helps to determine the relationship between causes and effects of the various experimental treatments. The experiments are carried out under controlled conditions so that input and output can be measured directly.

Objectives of the Experimental Design

i) To estimate the effects of various treatments and to compare the differences of effects are significant or not.

ii) To estimate the interaction effects of various treatments and to compare them

iii) To establish the mathematical relationship between various treatments and their effects.

iv) To estimate error

v) To control error

vi) Proper interpretation of the results

Steps in Design of Experiments

The major steps in the experimental design are

i) The problem which has to be solved by experiment should be stated in clear and significant way.

ii) The hypothesis in accordance with the experiment should be formulated in a clear way.

iii) The experimental technique and designing to solve the problem should be divided in a systematic way.

iv) The result that come from the designing should be examined carefully and reference back to the reasons for the enquiry of the required information. This is done to ensure that experiment provides an adequate extent.

v) The possible outcomes that come from the experiment should be under the consideration of statistical procedure. This is done to ensure the validity of the conditions necessary for statistical procedure.

vi) The conclusion is drawn after the experiment with the measure of reliability of the estimate of any treatment that is evaluated.

vii) The whole investigation is evaluated specially with the other investigation on the same or similar problems.

Terminology in Experimental Design

Experiment

It is means of getting an answer to a question that the experimenter has in mind. In planning experiment, we clearly state our objectives and formulate the hypothesis we want to test. Experiment can be divided into two categories i) Absolute and ii) Comparative Absolute experiment consists of determining the absolute value of some characteristic such as finding correlation coefficient between two variable, average intelligence of a group of people etc. While the comparative experiment consists of comparing different types of fertilizers, different types of cultivation methods, different varieties of crops etc.

Treatment

These are the inputs whose outcomes are to be estimated and compared. These are the different procedures under comparison in the experiment. In agricultural experiment different types of fertilizers, different types of cultivation process, different varieties of crops are treatments.

Experimental Unit

The smallest division of the experimental material in which the treatments are applied and the effects of treatments are measured. In agricultural experiment plot are the experimental unit. In feeding experiment of cows, a cow is experimental unit. Among the patients admitted in a hospital, a patient is an experimental unit.

Yields (Effects)

The outcomes of the experiment due to the application of treatments in experimental units are called yields. In agricultural experiment production of crop on using different fertilizers are yields.

Blocks

The experimental field is divided into relatively homogeneous subgroups or strata which is homogeneous or uniform among themselves than the field as a whole are called blocks.

Experimental Error

A fundamental phenomenon in replicated experiments is the variation in the measurements made on different experimental units even when they get the same treatment. A part of this variation is systematic and can be explained whereas the remainder is to be taken of the random type. This unexplained random part of the variation is termed as experimental error. This is a technical term and does not mean a mistake but includes all types of extraneous variation due to (i) inherient variability in the experimental units, (ii)small errors associated with the measurements made and (iii)lack of representativeness of the sample to the population under study.

It provides a basis for the confidence to be placed in the inference about the population. So it is to be estimated and controlled. It can be estimated by the replication and can be controlled by use of local control.

Precision

It is the amount of information or sensitivity of an experiment. It is given by reciprocal of the variance of the mean. If an experiment is replicated r times the precision of experiment is

$\frac{1}{Var(\overline{x})} = \frac{r}{\sigma^2}$, σ^2 is variance.

The presision increases as the replication increases or the variance decreases.

Efficiency of Design

Consider any two design D_1 and D_2 with replications r_1 and r_2 and variance σ_1^2 and σ_2^2 respectively, then the ratio of the precisions of design D_1 and D_2 is called the relative efficiency of design D_1 with respect to D_2. It is given by $E = \frac{r_1}{s_1^2} / \frac{r_2}{s^2_2}$.

If E = 1, both the design D_1 and D_2 are equally efficient.

If E > 1, design D_1 is more efficient than design D_2.

If E < 1, design D_1 is less efficient than design D_2.

Basic principles of experimental design

Designing of an experiment is deciding how the observations should be taken to answer a particular question in a valid, efficient and economic way. The design and analysis go together in the sense that if an experiment is properly designed then there will exist an appropriate way of analyzing the data. The application of the technique of analysis of variance is appropriate only when the data conform to the basic set up of the analysis of variance. The analysis of the data will be meaningless if the assumption in the analysis of variance are not fulfilled. Hence the layout and analysis of data are coordinated in the design of experiments.

According to R.A. Fisher a good experimental design must posses the following three principles namely;

(i) Replication (ii) Randomization (iii) Local control

Replication

It is the repetition of treatments under investigation. A treatment is repeated a large number of times in order to obtain more reliable result than is possible from single observation.

It works in two ways:

i) Along with randomization, it provides an estimate of the treatment effect.

ii) Along with local control, it provides minimization of error.

The most effective way to increase precision is to increase the number of replication. The precision can be increased by increasing the plot size but it has found that increasing replication of small plot is more efficient than using larger plot. The replication in a particular case depends on the variability of the material, cost of taking observations etc. It broadens the scope of experiment by including different types of experimental units.

Randomization

It is a process of allocating treatments to various plots in a random manner. It ensures that each treatment will have an equal chance of being assigned to an experimental unit. It greatly reduces the bias of applying a particular treatment to a particular unit. It is essential for valid estimate of experimental error and to minimize bias in the results. It is precaution against disturbances (errors)that may or may not occur and may or may not be serious if occur.

It aims:

Equalization of factor not under control

i) Eliminates bias in any form

ii) Provides a basis to estimate the treatment effect.

iii)

Local control

If the experimental material is heterogeneous and different treatments are allocated to various experimental units (plots) in random manner then experimental error will be increased. It is desirable to reduce the experimental error without increasing replication or without interfereing the required randomness.

The experimental error can be minimized by making the relatively heterogeneous experimental material into relatively homogeneous blocks is called local control. For this purpose the whole experimental area(material) is divided into a number of blocks perpendicular to the direction of fertility gradient the objective of local control is to give equal advantage of soil fertility to all the treatments. The experimental material is divided into a number of block rowwise or columnwise or both such that variation between blocks is maximum and variation within each block is minimum. Divide each block into as many plots as the number of treatments and allocate the treatments to the plots of each block separately in a random manner.

Completely randomized design (CRD)

It is simplest of all the design which is based upon only two principles of design namely replication and randomization. In this design treatments are assigned completely at random manner so that each and every experimental unit has equal chance of receiving any treatment. it is appropriate for the homogeneous experimental material.

Layout

Let us consider t treatments in which i^{th} treatment is replicated r_i times, i= 1,2,3,4.........t so that there are n experimental units such that $n = \sum_{i=1}^{t} r_i$. The homogeneous experimental material is divided into n plots(experimental units) and i^{th} treatment repeats in r_i units. The treatments are replicated along row wise as well as column wise. We study the variation between treatments. Observations are classified according to one way. Hence this is the case of one way ANOVA.

Consider a particular case in which $r_1 = r_2 = r_3 = \ldots\ldots = r_t = r$ then n = rt. In this case each treatment is repeated equal number of times r. In general equal number of replication should be made for each treatment except in the case when some treatments are of greater interest than the others. For example, consider t = 3 (A, B, C) and r = 4 then the treatments are allocated as shown below;

C	A	B	A
C	B	A	A
B	B	C	C

Mathematical Model

$$y_{ij} = \mu + \tau_i + e_{ij}\ ;\ i = 1,2,3,\ldots\ldots,t;\ j = 1,2,3,\ \ldots\ldots,\ r$$

Where,

y_{ij} = j^{th} unit receiving i^{th} treatment

μ= general mean effect

τ_i = effect due to i^{th} treatment

e_{ij} = error due to chance

Assumptions

i) All the observations are independent.
ii) All the observations are drawn from population having constant variance.
iii) All the treatments should be homogeneous as far as possible.
iv) Various treatments and environmental effects should be additive in nature.
v) All e_{ij} are i.i.d. $N(0, \sigma_e^2)$

Problem to test

H_{0T}: $\mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{t.}$ (There is no significant difference between the treatments)
H_{1T}: At least one $\mu_{i.}$ is different. $i = 1, 2, 3\ldots\ldots t$ (There is at least one significant difference between treatments)

Statistical Analysis

In the model $y_{ij} = \mu + \tau_i + e_{ij}$, parameters μ and τ_i are determined by using the principle of least square by minimizing error(residual) sum of square.

Total sum of square (TSS) = Sum of square due to treatment (SST) + Sum of square due to error (SSE)

Degree of freedom (d.f.) for various sum of square

Degree of freedom for total sum of square = rt −1 = n − 1

Degree of freedom for sum of square due to treatment = t − 1

Degree of freedom for sum of square due to error = n − t

Mean Sum of Square (MSS)

The sum of square divided by the corresponding degree of freedom gives the respective mean sum of square or variance.

Mean sum of square due to treatment (MST) = SST/t−1

Mean sum of square due to error (MSE) = SSE/ t(r − 1)

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{cal}	F_{tab}
Treatment	t − 1	SST	MST = SST/(t − 1)	F_T = MST/MSE	$F_{\alpha\,[(t-1),\,t(r-1)]}$
Error	t(r − 1)	SSE	MSE = SSE / t(r − 1)		
Total	r(t − 1)	TSS			

Decision

Reject H_{0T} at α% level of significance if $F_T > F_{\alpha\{(t-1),\, t(r-1)\}}$, accept otherwise.

Relation to calculate TSS, SST and SSE

$$TSS = \sum_{i=1}^{t} \sum_{j=1}^{t} (\bar{y}_{ij} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}^2 - C.F.$$

$$SST = r \sum_{i=1}^{t} (y_{i.} - \bar{y}_{..})^2$$

$$= \frac{\sum_{i=1}^{t} T_{i.}^2}{r} - C.F. \text{ where } C.F. = \frac{G^2}{N}$$

$$SSE = TSS - SST$$

Advantages of CRD

i) It is easy to layout,

ii) It allows maximum number of d.f. for MSS due to error which minimize error sum of square.

iii) It is simple to statistical analysis due to one way classification.

iv) If some observations are missing the analysis still remains simple.

Disadvantages of CRD

i) Principle of local control is not used.

ii) It is suitable for small treatments, replication and homogeneous material only.

Uses of CRD

(i) It is used in green house, laboratory etc.

Example 1: The yield of treatments in different plots are as shown below. Carry out analysis.

t_4 1401	t_3 2536	t_3 2459	t_1 2537	t_3 2827	t_1 2069
t_2 2211	t_1 1797	t_4 1170	t_4 1516	t_4 2104	t_3 2385
t_2 3366	t_1 2104	t_2 2591	t_3 2460	t_4 1077	t_2 2544

Solution:

Problem to test

$H_{0T}: \mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$

H_{1T}: At least one of the $\mu_{i.}$ is different, $i = 1, 2, 3, 4$ $(1 = t_1, 2 = t_2, 3 = t_3, 4 = t_4)$

						$T_{i.}$
t_1	2537	2069	1797	2104		6049
t_2	2211	3366	2591	2544		10712
t_3	2536	2459	2827	2385	2460	12667
t_4	1401	1170	1516	2104	1077	6427
						$\Sigma T_{i.} = 35855$

Now, $\sum_{i,j} y_{ij}^2 = 2537^2 + 2069^2 + 1797^2 + 2104^2 + \ldots\ldots + 1077^2 = 88374337$

$G = \Sigma T_{i.} = 35855$

$N = n_1 + n_2 + n_3 + n_4 = 4 + 4 + 5 + 5 = 18$

$$C.F. = \frac{G^2}{N} = \frac{(35855)^2}{18} = 71421168.06$$

$$TSS = \sum_{i,j} y_{ij}^2 - C.F.$$

$= 88374337 - 71421168.06$

$= 16953168.94$

$$SST = \sum_i \frac{T_{i.}^2}{n_i} - C.F.$$

$$= \left\{\frac{(6049)^2}{4} + \frac{(10712)^2}{4} + \frac{(12667)^2}{5} + \frac{(6427)^2}{5}\right\} - 71421168.06 = 5768691.79$$

$SSE = TSS - SST$

$= 16953168.94 - 5768691.79 = 11184477.15$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Treatment	3	5768691.79	1922897.26	22.72	$F_{0.05(3,14)}=5.66$
Error	14	11184477.15	84605.51		
Total	17	16953168.94			

Decision

$F_T = 22.72 > F_{0.05(3,14)} = 5.66$, reject H_{0T} at 5% level of significance.

Conclusion

Treatment difference is highly significant.

Example 2: There are different types of text entry techniques A, B and C. The efficiency is measured by its error rate. The error rate found while entering in four different mobile are given below.

A	8	C	5	C	4	B	5
B	7	A	4	B	1	A	6
C	2	B	5	C	3	A	7

Test the significant difference among the three text entry techniques.

Solution:

Problem to test

H_0: The difference in efficiency among the text entry technique are not significant.

H_1: The difference in efficiency among the text entry technique are significant.

Now,

A	8	4	6	7
B	7	5	5	1
C	2	5	4	3

Changing origin to 4 ($u_{ij} = x_{ij} - 4$)

Text entry technique					$T_{i.}$
A	4	0	2	3	9
B	3	1	1	–3	2
C	–2	1	0	–1	–2
					$\Sigma T_{i.} = 9$

Here,

$G = \Sigma T_{i.} = 9, N = 12$

$$C.F. = \frac{G^2}{N} = \frac{(9)^2}{12} = 6.75$$

$$\sum_{i,j} u_{ij}^2 = 16 + 0 + 4 + 9 + 9 + 1 + 1 + 9 + 4 + 1 + 0 + 1 = 55$$

$$TSS = \sum_{i,j} u_{ij}^2 - C.F. = 55 - 6.75 = 48.25$$

$$SST = \frac{\Sigma_i T_{i.}^2}{r} - C.F.$$

$$= \frac{1}{4}\{81+4+4\} - 6.75 = 22.25 - 6.75 = 15.5$$

$$SSE = TSS - SST$$

$$= 48.25 - 15.5 = 32.75$$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Treatment	2	15.5	7.75	2.73	$F_{0.05(2,9)} = 4.26$
Error	9	32.75	3.63		
Total	11	48.25			

Decision

$F_T = 2.73 < F_{0.05(2,9)} = 4.26$, accept H_0 at 5% level of significance.

Conclusion

The efficiency of three techniques are not significant.

Randomised Block Design (RBD)

When the experimental material is not homogeneous the RBD is better than CRD. The RBD is the design where the treatments are allocated in random manner but randomization is restricted that each treatment must occur once in each row or once in each column. Hence this design is row wise or column wise. It is based upon the all principles of design namely replication, randomization and local control.

Lay out

Let us consider t treatments with r replication each so that there are N = r t experimental unit.

Let us divide the non homogeneous experimental material into a number of relatively homogeneous blocks perpendicular to the direction of fertility gradient. Divide each block into as many plot as the number of treatment and allocate these t treatments randomly to the plot of each block separately.

The treatments are replicated along row wise or column wise. In this case we study the variation between treatments and the variation between blocks. Observation are classified according to two ways. Hence it is the case of two way ANOVA.

Let us consider t = 3(A, B, C) and r = 4. The treatments are allocated in the blocks as shown below.

Block I	Block II	Block III	Block IV
A	C	A	C
B	A	C	B
C	B	B	A

Mathematical model

$Y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$

Where,

y_{ij} = j^{th} block receiving i^{th} treatment.

$i = 1, 2, \ldots\ldots, t,\ j = 1, 2, \ldots\ldots, r$

μ = constant effect

τ_i = Effect due to i^{th} treatment

β_j = Effect due to j^{th} block

e_{ij} = Error due to chance

Assumptions

(i) All the observations are independent.

(ii) All the observations should be drawn from normal population having constant variance.

(iii) All the treatments should be homogeneous as far as possible.

(iv) Various treatments and environmental effect are additive in nature.

(v) e_{ij} are i.i.d. $N(0, \sigma_e^2)$

Statistical analysis

In the model $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$, parameters μ, τ_i and β_j are determined by using the principle of least square by minimizing error(residual) sum of square.

Total sum of square (TSS) = Sum of square due to treatment (SST) + Sum of square due to block (SSB) + Sum of square due to error (SSE)

Degree of freedom (d.f.) for various sum of square:

Degree of freedom for total sum of square = rt –1 = N – 1

Degree of freedom for sum of square due to treatment = t – 1

Degree of freedom for sum of square due to block = r – 1

Degree of freedom for sum of square due to error = t(r – 1)

Mean Sum of Square (MSS):

The sum of square divided by the corresponding degree of freedom gives the respective mean sum of square or variance.

$$\text{Mean sum of square due to treatment (MST)} = \frac{SST}{t-1}$$

$$\text{Mean sum of square due to block (SSB)} = \frac{SSB}{r-1}$$

$$\text{Mean sum of square due to error (MSE)} = \frac{SSE}{(t-1)(r-1)}$$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{cal}	F_{tab}
Treatment	t – 1	SST	MST = SST/(t–1)	F_T= MST/MSE	$F_{\alpha[(t-1),\,(t-1)(r-1)]}$
Block	r – 1	SSB	MSB = SSB/(r–1)	F_B = MSB/MSE	$F_{\alpha[(r-1),\,(t-1)(r-1)]}$
Error	(r–1)(t–1)	SSE	MSE = SSE/(t – 1)(r –1)		
Total	rt–1	TSS			

Decision:

Reject H_{0T} at α% level of significance if $F_T > F_{\alpha\{(t-1),\, t(r-1)\}}$, accept otherwise.

Reject H_{0C} at α% level of significance if $F_C > F_{\alpha\{(r-1),\, t(r-1)\}}$, accept otherwise.

Relation to calculate TSS, SST, SSB and SSE:

$$TSS = \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}^2 - \frac{G^2}{N}$$

$$= \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}^2 - C.F.$$

$$SST = r \sum_{i=1}^{t} (y_{i.} - \bar{y}_{..})^2$$

$$= \frac{\sum_{i=1}^{t} T_{i.}^2}{r} - C.F.$$

$$SSB = t \sum_{j=1}^{r} (\bar{y}_{.j} - \bar{y}_{..})^2$$

$$= \frac{\Sigma T_{.j}^2}{r} - C.F.$$

$$SSE = TSS - SST - SSB$$

Advantages of RBD

(i) RBD provides the better result than CRD.

(ii) There is no restriction on the number of treatment or replication. But at least two replication is necessary.

(iii) The statistical analysis remains simple if some observations are missing.

Disadvantage of RBD

It is suitable for only small number of treatments.

Example 3: Carry out ANOVA of following design.

X	7	Y	8	Z	12	X	8
Y	10	X	9	X	6	Z	6
Z	8	Z	11	Y	9	Y	5

Solution:

Here, treatments are replicated along rowwise as only. Hence this is the case of RBD.

Problem to test

H_{0T}: There is no significant difference between treatments.

H_{1R}: There is significant difference between treatments.

H_{0B}: There is no significant difference between block

H_{1B}: There is significant difference between block.

Here,

Treatments	Block				$T_{i.}$
	I	II	III	IV	
X	7	9	6	8	30
Y	10	8	9	5	32
Z	8	11	12	6	37
$T_{.j}$	25	28	27	19	99

$N = 3 \times 4 = 12$

$$G = \sum_i T_{i.} = \sum_j T_{j.} = 99$$

$$C.F. = \frac{G^2}{N} = \frac{(99)^2}{12} = 816.75$$

$$\sum_i \sum_j y_{ij}^2 = 7^2 + 9^2 + 6^2 + 8^2 + 10^2 + 8^2 + 9^2 + 5^2 + 8^2 + 11^2 + 12^2 + 6^2 = 865$$

$$TSS = \sum_i \sum_j y_{ij}^2 - C.F. = 865 - 816.75 = 48.25$$

$$SST = \frac{1}{r}\sum_i T_{i.}^2 - C.F.$$

$$= \frac{1}{4}[30^2 + 32^2 + 37^2] - 816.75 = 823.25 - 816.75 = 6.5$$

$$SSB = \frac{1}{t}\sum_i T_{.j}^2 - C.F.$$

$$= \frac{1}{3}[25^2 + 28^2 + 27^2 + 19^2] - 816.75 = 833 - 816.75 = 16.25$$

$SSE = TSS - SST - SSB$

$= 48.25 - 6.5 - 16.25$

$= 25.5$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Treatments	2	6.5	3.25	$F_T = 0.764$	$F_{0.05(2,6)} = 5.14$
Blocks	3	16.25	5.416	$F_R = 1.274$	$F_{0.05(3,6)} = 4.76$
Error	6	25.5	4.25		
Total	11	48.25			

Decision

$F_T = 0.764 < F_{0.05(2,6)} = 5.14$, accept H_{0T} at 5% level of significance.

$F_B = 1.274 < F_{0.05(3,6)} = 4.76$, accept H_{0B} at 5% level of significance.

Conclusion

There is no significant difference between treatments and there is no significant difference between blocks.

Example 4: Let A, B, C and D are four different page replacement algorithms. The following table gives the running times of programs under each replacement algorithm in 5 different blocks and in each block 5 different programs were used.

Block I		Block II		Block III		Block IV		Block V	
A	32	B	33	D	30	A	35	C	36
B	34	C	34	C	35	C	32	D	29
C	31	A	34	B	36	B	37	A	37
D	29	D	26	A	33	D	28	B	35

Analyse the above result to test whether there is significant difference between yields of four varieties.

Problem to test

H_{0T}: There is no significant difference between algorithms.

H_{1T} : There is significant difference between algorithms.

Here,

Treatment	Block				
	I	II	III	IV	V
A	32	34	33	35	37
B	34	33	36	37	35
C	31	34	35	32	36
D	29	26	30	28	29

Changing origin to 32 ($u_{ij} = y_{ij} - 32$)

Treatment	Block					$T_{i\cdot}$
	I	II	III	IV	V	
A	0	2	1	3	5	11
B	2	1	4	5	3	15
C	-1	2	3	0	4	8
D	-3	-6	-2	-4	-3	-18
$T_{\cdot j}$	-2	-1	6	4	9	16

Now,

$N = t \times r = 4 \times 5 = 20$

$G = \sum_i T_{i.} = \sum_j T_{.j} = 16$

$C.F. = \frac{G^2}{N} = \frac{(16)^2}{20} = 12.8$

$\sum_i \sum_j T\, u_{ij}^2 = (0 + 4 + 1 + 9 + 25 + 4 + 1 + 16 + 25 + 9 + 1 + 4 + 9 + 0 + 16 + 9 + 36 + 4 + 16 + 9) = 198$

$TSS = \sum_i \sum_j u_{ij}^2 - C.F.$

$= 198 - 12.8 = 185.2$

$SST = \frac{1}{r} \sum_i T_{i.}^2 - C.F.$

$= \frac{1}{5} [(11)^2 + (15)^2 + (8)^2 + (-18)^2] - 12.8$

$= 146.8 - 12.8 = 134$

$SSB = \frac{1}{t} \sum_j T_{.j}^2 - C.F.$

$= \frac{1}{4} [(-2)^2 + (-1)^2 + (6)^2 + (4)^2] + (9)^2] - 12.8$

$= 34.5 - 12.8 = 21.7$

$SSE = TSS - SST - SSB$

$= 185.2 - 134 - 21.7$

$= 29.5$

ANOVA table,

Source of variation	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Between treatment	3	134	46.66	18.982	$F_{0.05(3,12)} = 3.49$
Between block	4	21.7	5.425	2.207	$F_{0.05(4,12)} = 3.26$
Error	12	29.5	2.458		
Total	19	185.2			

Decision

$F_T = 18.982 > F_{0.05(3,12)} = 3.49$, reject H_{0T} at 5% level of significance.

Conclusion

There is significant difference between running times of four algorithms.

Example 5: Calculate the minimum number of replication so that an observed difference of 10% of the mean will be taken as significant at 5% level, the coefficient of variation (c.v.) of the plot values being 12%.

Solution:

Let r be the replication and μ be the mean, then

$$\text{C.V.} = \frac{\sigma}{\mu} \times 100\%$$

or $12\% = \frac{\sigma}{\mu} \times 100\%$

or $\sigma = 0.12\mu$ ....(*)

Also, $\bar{x}_1 - \bar{x}_2 = 10\%$ of μ

or $\bar{x}_1 - \bar{x}_2 = 0.1\mu$ ....(**)

Now,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{2}{r}}} = \frac{0.1\,\mu}{0.12\,\mu\sqrt{\frac{2}{r}}} = \frac{5}{6}\sqrt{\frac{r}{2}}$$

For t to be significant at 5% level of significance, we have $|t| > 1.96$

or $\frac{5}{6}\sqrt{\frac{r}{2}} > 1.96$

or $r > (1.96)^2 \times (1.2)^2 \times 2$

or $r > 11.1$

Hence required no of replication is 12.

Efficiency of RBD relative to CRD

The mechanism of the precision of RBD as compared to CRD is called efficiency of RBD relative to CRD.

Let us consider design having t treatments with r replication each.

If we perform RBD then,

The mathematical model of RBD is $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$; $i = 1, 2, 3.......t$, $j = 1, 2, 3..... r$

If we perform CRD then,

The mathematical model of CRD is $y_{ij} = \mu + \tau_i + e_{ij}$; $i = 1, 2, 3......t$, $j = 1, 2, 3 ... r$

Now the efficiency of RBD relative to CRD is given by

$$\text{Precision of RBD/ Precision of CRD} = \frac{1}{MSE} \Big/ \frac{1}{MSE'} = \frac{MSE'}{MSE} = \frac{\sigma_e'^2}{\sigma_e^2} = \frac{r(t-1)\,MSE + (r-1)\,MSB}{(rt-1)\,MSE}$$

If $\frac{\sigma_e'^2}{\sigma_e^2} < 1$ then RBD is less efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} > 1$ then RBD is more efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} = 1$ then RBD and CRD are equally effective.

Example 6: From the following ANOVA table of RBD, determine it's efficiency with respect to CRD.

S.V.	d.f.	S.S.	M.S.S.
Between treatments	5	750	150
Between blocks	3	180	60
Error	15	200	13.33
Total	23	1130	

Solution:

Here, $t = 6$, $r = 4$, $MSE = 13.33$, $MST = 150$, $MSB = 60$

Now, efficiency of RBD with respect to CRD is

$$\frac{s_e'^2}{s_e^2} = \frac{r(t-1)\,MSE + (r-1)\,MSB}{(rt-1)\,MSE} = \frac{4\times5\times13.33 + 3\times60}{23\times13.33} = 1.456$$

RBD is 45.6% efficient than CRD.

Missing plot for RBD

Let us consider an RBD involving t treatment with r replication each. Let one of the observation say x receiving i^{th} treatment in the j^{th} block is missing.

Let, G' = total of all known $rt-1$ values.

T' = total of all known values of i^{th} treatment.

B' = total of all kown values of j^{th} block.

	B_1 B_2 $B_3 B_j$ B_r	$T_{i\cdot}$
T_1	y_{11} y_{12} y_{13} ....y_{1j} .... y_{1r}	$T_{1\cdot}$
T_2	y_{21} y_{22} y_{23} ....y_{2j} .... y_{2r}	$T_{2\cdot}$
T_3	y_{31} y_{32} y_{33} .... y_{3j} .... y_{3r}	$T_{3\cdot}$
	: : : : :	
	: : : : :	
T_i	y_{i1} y_{i2} y_{i3} .... x ....y_{ir}	$T_{i\cdot} = T'+x$
	: : : : :	
	: : : : :	
	: : : : :	
T_t	y_{t1} y_{t2} y_{t3} ....y_{tj} ....y_{tr}	$T_{t\cdot}$
$T_{\cdot j}$	$T_{\cdot 1}$ $T_{\cdot 2}$ $T_{\cdot 3}$ $T_{\cdot j}=B'+x$ $T_{\cdot r}$	$G'+x$

$$x = \frac{T't + B'r - G'}{(t-1)(r-1)}$$

Substitute the value of x in place of missing value and carry out analysis as usual except that one degree of freedom is substracted from total and consequently from error. Because of the change in level of degree of freedom we obtain an upward bias in SST. Hence to get better result substract an adjustment factor from SST.

$$\text{Adjustment factor (k)} = \frac{(B' + T't - G')^2}{t(t-1)(r-1)^2}$$

Adjusted SST (SST_A) = SST – k.

Example 7: The table given below represents the yield of 3 varieties in 4 replicate experiment for which one observation is missing. Estimate the missing value and analyse the data.

A	18.1	B	?	A	15.2	C	13.2
C	16.0	A	12.1	B	17.5	A	16.6
B	16.3	C	13.4	C	16.3	B	18.1

Solution:

Treatment	Block				$T_{i.}$
	I	II	III	IV	
A	18.1	12.1	15.2	16.6	62
B	16.3	? (x)	17.5	18.1	51.9+x
C	16.0	13.4	16.3	13.2	58.9
$T_{.j}$	50.4	25.5+x	49.0	47.9	172.8+x

Here T' = 51.9, B' = 25.5, G' = 172.8, t = 3, r = 4

$$\text{Now, missing value } x = \frac{T't + B'r - G'}{(t-1)(r-1)} = \frac{51.9 \times 3 + 25.5 \times 4 - 172.8}{2 \times 3} = 14.15$$

Problem to test

H_{0T} : There is no significant difference between treatments.

H_{1T} : There is significant difference between treatments.

H_{0B} : There is no significant difference between blocks.

H_{1B} : There is significant difference between blocks.

$G = G' + x = 172.8 + 14.15 = 186.95$, $N = 4 \times 3 = 12$

$$C.F. = \frac{G^2}{N} = \frac{(186.95)^2}{12} = 2912.525$$

$$\sum_i \sum_j y_{ij}^2 = (18.1)^2+(12.1)^2+(15.2)^2+(16.6)^2+(16.3)^2+(14.15)^2+(17.5)^2+(18.1)^2+16^2+(13.4)^2$$

$$+ (16.3)^2+(13.2)^2 = 2955.88$$

$$TSS = \sum_i \sum_j y_{ij}^2 - C.F.$$

$$= 2955.88 - 2912.525 = 43.354$$

$$SST = \frac{1}{r}\sum_i T_{i.}^2 - C.F.$$

$$= \frac{1}{4}\{62^2+66.05^2+58.9^2\} - 2912.525 = 2918.953 - 2912.525 = 6.428$$

$$SSB = \frac{1}{t}\sum_j T_{.j}^2 - C.F.$$

$$= \frac{1}{3}\{50.4^2+ 39.65^2+ 49^2+47.9^2\} - 2912.525 = 2935.897 - 2912.525$$

$$= 23.372$$

$$\text{Adjustment factor (k)} = \frac{(B' + tT' - G')^2}{t(t-1)(r-1)^2} = \frac{(25.5 + 3 \times 51.9 - 172.8)^2}{3 \times 2 \times 9} = 1.306$$

$$SST_A = SST - k$$

$$= 6.428 - 1.306 = 5.121$$

$$SSE = TSS - SST_A - SSB$$

$$= 43.354 - 5.121 - 23.372$$

$$= 14.861$$

S.V.	d.f.	S.S.	M.S.S	F_{Cal}	F_{Tab}
Treatments	2	5.121	2.56	0.861	$F_{0.05(2,5)} = 5.79$
Blocks	3	23.372	7.79	2.62	$F_{0.05(3,5)} = 5.41$
Error	5	14.861	2.972		
Total	10	43.354			

Decision

$F_T = 0.861 < F_{0.05(2,5)} = 5.79$, accept H_{0T} at 5% level of significance.

$F_B = 2.62 < F_{0.05(3,5)} = 5.41$, accept H_{0B} at 5% level of significance.

Conclusion

There is no significant difference between treatments.

There is no significant difference between blocks.

Latin Square Design (LSD)

When the experimental material is not homogeneous the LSD is better than RBD. In RBD local control is used according to one way grouping i.e. according to blocks but in LSD local control is used according to two way grouping i.e. rows and columns. Hence it is used when two sources of errors are to be controlled simultaneously. In this design number of treatments are equal to the number of replication and the treatments are allocated in such a way that each of the treatment occurs once and only once in each row and column. In this design Latin alphabet are used to denote the treatments, and shape is square due to equal number of treatments and replication so called Latin square design. It is based upon the all principles of design namely replication, randomization and local control.

Lay out

Let us consider m treatments with m replication each so that there are $N = m^2$experimental unit. Let us divide the experimental material into m^2 experimental units arranged in square so that each row as well as column contains m units. In this design none of treatments are replicated along row wise or column wise. In this case we study the variation between treatments, the variation between rows and variation between columns. It has only m^2 experimental unit but studies variation of three factors i.e. rows, columns and treatments. Hence it is the case of incomplete three way ANOVA. For complete three way ANOVA we need m^3 experimental unit.

Let us consider t = 4(A, B, C, D) then 4 × 4 LSD is as shown below.

A	D	B	C
B	C	D	A
C	B	A	D
D	A	C	B

Mathematical model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk}\ ;\ i = 1, 2, 3 \ldots\ldots m;\ j = 1, 2, 3\ldots\ldots m;\ k = 1, 2, 3 \ldots\ldots m$$

Where,

y_{ijk} = i^{th} row and j^{th} column receiving k^{th} treatment.

μ = constant effect

α_i = effect due to i^{th} row

β_j = effect due to j^{th} column

τ_k = effect due to k^{th} treatment

e_{ijk} = error due to chance

Assumptions

(i) All the observations are independent.

(ii) All the observations should be drawn from normal population having constant variance.

(iii) All the treatments should be homogeneous as far as possible.

(iv) Various treatments and environmental effect are additive in nature.

(v) e_{ijk} are i.i.d. $N(0, \sigma_e^2)$

Statistical analysis:

In the model $y_{ij} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk}$, parameters μ, α_i, β_j and τ_k are determined by using the principle of least square by minimizing error(residual) sum of square.

Total sum of square (TSS) = Sum of square due to row(SSR) + Sum of square due to column (SSC) + Sum of square due to treatment (SST) + Sum of square due to error (SSE)

Degree of freedom (d f) for various sum of square:

Degree of freedom for total sum of square = $m^2 - 1 = N - 1$

Degree of freedom for sum of square due to row = $m-1$

Degree of freedom for sum of square due to column = $m-1$

Degree of freedom for sum of square due to treatment = $m-1$

Degree of freedom for sum of square due to error = $(m-1)(m-2)$

Mean Sum of Square (MSS)

The sum of square divided by the corresponding degree of freedom gives the respective mean sum of square or variance.

Mean sum of square due to row (MSR) = $\dfrac{SSR}{m-1}$

Mean sum of square due to column (MSC) = $\dfrac{SSC}{m-1}$

Mean sum of square due to treatment (MST) = $\dfrac{SST}{m-1}$

Mean sum of square due to error (MSE) = $\dfrac{SSE}{(m-1)(m-1)}$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{cal}	F_{tab}
Row	$m-1$	SSR	$MSR = \dfrac{SSR}{m-1}$	$F_R = MSR/MSE$	$F_{\alpha\{(m-1),(m-1)(m-2)\}}$
Column	$m-1$	SSC	$MSC = \dfrac{SSC}{m-1}$	$F_C = MSC/MSE$	$F_{\alpha\{(m-1),(m-1)(m-2)\}}$
Treatment	$m-1$	SST	$MST = \dfrac{SST}{m-1}$	$F_T = MST/MSE$	$F_{\alpha\{(m-1),(m-1)(m-2)\}}$
Error	$(m-1)(m-2)$	SSE	$MSE = \dfrac{SSC}{(m-1)(m-2)}$		
Total	m^2-1	TSS			

Decision:

Reject H_{0R} at α% level of significance if $F_R > F_{\alpha\{(m-1),(m-1)(m-2)\}}$, accept otherwise.

Reject H_{0C} at α% level of significance if $F_C > F_{\alpha\{(m-1),(m-1)(m-2)\}}$, accept otherwise.

Reject H_{0T} at α% level of significance if $F_T > F_{\alpha\{(m-1),(m-1)(m-2)\}}$, accept otherwise.

Relation to calculate TSS, SSR, SSC, SST and SSE:

$$TSS = \sum_{(i,j,k)} (y_{ijk} - \Sigma\bar{y}_{\ldots})^2 = \sum_{(i,j,k)} y_{ijk}^2 - \frac{G^2}{N} = \sum_{(i,j,k)} y_{ijk}^2 - C.F.$$

$$SSR = m\sum_{i} (\bar{y}_{i..} - \bar{y}_{\ldots})^2 = \frac{\Sigma_i T_{i..}^2}{m} - CF$$

$$SSC = m\sum_{i} (\bar{y}_{.j.} - \bar{y}_{\ldots})^2 = \Sigma\frac{T_{.j.}^2}{m} - CF$$

$$SST = m\sum_{i} (\bar{y}_{..k} - \bar{y}_{\ldots})^2 = \Sigma\frac{T_{..k}^2}{m} - CF$$

$$SSE = TSS - SSR - SSC - SST$$

Advantages of LSD

i) Due to the use of two way grouping of controls more variation than CRD and RBD.

ii) It is incomplete three way layout. It's advantage over complete three way layout is that instead of m^3 experimental units only m^2 units are needed.

iii) The statistical analysis remains simple if some observations are missing.

Disadvantages of LSD

i) The assumption of factors are independent is not always true.

ii) It is suitable for treatments 5 to 10.

iii) It is not easy in the field layout.

Example 8: The following is the 5×5 Latin square design for data taken from a manurial experiment with sugarcane. The five treatments were A = no manure; B = an inorganic manure;

C, D and E = three levels of farm yard manure.

Plan and yield of sugarcane (in a suitable unit) per plot.

Row	Column				
	I	II	III	IV	V
I	A 52.5	E 46.3	D 44.1	C 48.1	B 40.9
II	D 44.2	B 42.9	A 51.3	E 49.3	C 32.6
III	B 49.1	A 47.3	C 38.1	D 41.0	E 47.2
IV	C 43.2	D 42.5	E 67.2	B 55.1	A 45.3
V	E 47.0	C 43.2	B 46.7	A 46.0	D 43.2

Analyze the above data to find if there are any treatment effects.

Solution:

Problem to test

H_{0T} : There is no significant difference between treatments.

H_{1T}: There is significant difference between treatments.

Row	Column					
	I	II	III	IV	V	$T_{i..}$
I	A 52.5	E 46.3	D 44.1	C 48.1	B 40.9	231.9
II	D 44.2	B 42.9	A 51.3	E 49.3	C 32.6	220.3
III	B 49.1	A 47.3	C 38.1	D 41.0	E 47.2	222.7
IV	C 43.2	D 42.5	E 67.2	B 55.1	A 45.3	253.3
V	E 47.0	C 43.2	B 46.7	A 46.0	D 43.2	226.1
$T_{.j.}$	236.0	222.2	247.4	239.5	209.2	1154.3

$T_{..A} = 52.5 + 51.3 + 47.3 + 45.3 + 46.0 = 242.4$

$T_{..B} = 40.9 + 42.9 + 49.1 + 55.1 + 46.7 = 234.7$

$T_{..C} = 48.1 + 32.6 + 38.1 + 43.2 + 43.2 = 205.2$

$T_{..D} = 44.1 + 44.2 + 41.0 + 42.5 + 43.2 = 215.0$

$T_{..E} = 46.3 + 49.3 + 47.2 + 67.2 + 47.0 = 257.0$

$G = 1154.3,\ N = 25$

$$C.F. = \frac{G^2}{N} = \frac{(1154.3)^2}{25} = 53296.33$$

$$\sum_{(i,j,k)} y_{ijk}^2 = (52.5)^2 + (46.3)^2 + (44.1)^2 + (48.1)^2 + (40.9)^2 + \dots\dots\dots + (43.2)^2 = 54273.51$$

$$TSS = \sum_{(i,j,k)} y_{ijk}^2 - C.F. = 54273.51 - 53296.33 = 977.18$$

$$SSR = \frac{\Sigma_i T_{i..}^2}{m} - C.F. = \frac{1}{5}\{(231.9)^2+(220.3)^2+(222.7)^2+(253.3)^2+(226.1)^2\} - 53296.33$$

$$= 53437.41 - 53296.33 = 141.08$$

$$SSC = \frac{\Sigma_j T_{.j.}^2}{m} - C.F. = \frac{1}{5}\{(236)^2 + (222.2)^2 + (247.4)^2 + (239.5)^2 + (209.2)^2\} - 53296.33$$

$$= 53480.09 - 53296.33$$

$$= 183.76$$

$$SST = \Sigma \frac{T_{..k}^2}{m} - C.F. = \frac{1}{5}\{(242.4)^2 + (234.7)^2 + (205.2)^2 + (215)^2 + (257)^2\} - 53296.33$$

$$= 53644.57 - 53296.33$$

$$= 348.24$$

$$SSE = TSS - SSR - SSC - SST$$

$$= 977.18 - 141.08 - 183.76 - 348.24 = 304.1$$

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Row	4	141.08	35.27		
Column	4	183.76	45.94		
Treatment	4	348.24	87.06	$F_T = 3.436$	$F_{0.05(4,12)} = 3.26$
Error	12	304.1	25.34		
Total	24	977.18			

Decision

$F_T = 3.436 > F_{0.05(4,12)} = 3.26$, reject H_{0T} at 5% level of significance.

Conclusion

There are treatment effects.

Example 9: The layout and yield of four treatments in a 4×4 experiment is shown in the following table. Analyse the data.

D 20	B 17	A 20	C 19
B 21	A 18	C 18	D 17
A 18	C 21	D 17	B 17
C 20	D 19	B 17	A 18

Solution:

Problem to test

H_{0R}: There is no significant difference between rows

H_{1R}: There is significant difference between rows.

H_{1C}: There is no significant difference between columns

H_{1C}: There is significant difference between columns.

H_{0T}: There is no significant difference between treatments

H_{1T}: there is significant difference between treatments.

Now, changing origin to 19 ($u_{ijk} = y_{ijk} - 19$)

					$T_{i..}$
	D 1	B -2	A 1	C 0	0
	B 2	A -1	C -1	D -2	-2
	A -1	C 2	D -2	B -2	-3
	C 1	D 0	B -2	A -1	-2
$T_{.j.}$	3	-1	-4	-5	-7

$G = -7, N = 4^2 = 16$

$$C.F. = \frac{G^2}{N} = \frac{(-7)^2}{16} = 3.0625$$

$$\sum_{(i,j,k)} u_{ijk}^2 = 1+4+1+0+4+1+1+4+1+4+4+4+1+0+4+1 = 35$$

$$TSS = \sum_{(i,j,k)} u_{ijk}^2 - C.F. = 35 - 3.0625 = 31.9375$$

$$SSR = \frac{1}{4}\sum_{i} T_{i..}^2 - C.F. = \frac{1}{4}\{(0)^2 + (-2)^2 + (-3)^2 + (-2)^2\} - 3.0625 = 3.75 - 3.0625 = 0.6875$$

$$SSC = \frac{1}{4}\sum_{j} T_{.j.}^2 - C.F. = \frac{1}{4}\{(3)^2+(-1)^2+(-4)^2+(-5)^2\} - 3.0625 = 12.75 - 3.0625 = 9.6875$$

$T_{..A} = 1-1-1-1 = -2$

$T_{..B} = -2+2-2-2 = -4$

$T_{..C} = 0-1+2+1 = 2$

$T_{..D} = 1-2-2+0 = -3$

$$SST = \frac{1}{4}\sum_{k} T_{..k}^2 - C.F. = \frac{1}{4}\{(-2)^2 + (-4)^2 + (2)^2 + (-3)^2\} - 3.0625 = 8.25 - 3.0625 = 5.1875$$

$$SSE = TSS - SSR - SSC - SST = 31.9375 - 0.6875 - 9.6875 - 5.1875 = 16.375$$

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Row	3	0.6875	0.2291	0.0839	$F_{0.05(3,6)} = 4.76$
Column	3	9.6875	3.2291	1.1832	$F_{0.05(3,6)} = 4.76$
Treatment	3	5.1875	1.7291	0.6335	$F_{0.05(3,6)} = 4.76$
Error	6	16.375	2.7291		
Total	15	31.9375			

Decision

$F_R = 0.0839 < F_{0.05(3,6)} = 4.76$, accept H_{0R} at 5% level of significance.

$F_C = 1.1832 < F_{0.05(3,6)} = 4.76$, accept H_{0C} at 5% level of significance.

$F_T = 0.6335 < F_{0.05(3,6)} = 4.76$, accept H_{0T} at 5% level of significance.

Conclusion

There is no significant difference between rows, there is no significant difference between columns and there is no significant difference between treatments.

Efficiency of LSD Relative to CRD

The mechanism of the precision of LSD as compared to CRD is called efficiency of LSD relative to CRD.

If we perform LSD then,

The mathematical model of LSD is $y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk}$; $i = 1, 2, 3 \ldots\ldots m$; $j = 1, 2, 3 \ldots\ldots m$; $k = 1, 2, 3 \ldots\ldots m$

If we perform CRD then,

The mathematical model of CRD is $y_{ij} = \mu + \tau_i + e_{ij}'$; $i = 1, 2, 3 \ldots\ldots m$, $j = 1, 2, 3 \ldots\ldots m$

Now the efficiency of LSD relative to CRD is given by

$$\text{Precision of LSD/ Precision of CRD} = \frac{1}{MSE} \Big/ \frac{1}{MSE'} = \frac{MSE'}{MSE} = \frac{\sigma_e'^2}{\sigma_e^2}$$

$$= \frac{(m-1)\,MSE + MSR + MSC}{(m+1)\,MSE}$$

If $\frac{\sigma_e'^2}{\sigma_e^2} < 1$ then LSD is less efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} > 1$ then LSD is more efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} = 1$ then LSD and CRD are equally effective.

Example 10: From the following ANOVA table of 4×4 LSD determine it's efficiency with respect to CRD.

S.V.	d.f.	S.S.	M.S.S.
Rows	3	2.133	0.711
Columns	3	2.203	0.734
Treatments	3	10.663	3.554
Error	6	7.059	1.177
Total	15	22.058	

Solution:

m = 4, MSR = 0.711, MSC = 0.734, MST = 3.554, MSE = 1.177

Now,

Efficiency of LSD with respect to CRD is

$$\frac{\sigma_e'^2}{\sigma_e^2} = \frac{(m-1)\,MSE + MSR + MSC}{(m+1)\,MSE} = \frac{3 \times 1.177 + 0.711 + 0.734}{5 \times 1.177} = 0.845 = 84.5\%$$

LSD is 15.5% less efficient than CRD.

Efficiency of LSD relative to RBD

The mechanism of the precision of LSD as compared to RBD is called efficiency of LSD relative to RBD.

If we perform LSD then,

The mathematical model of LSD is $y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk}$; $i = 1, 2, 3 \ldots\ldots m$; $j = 1, 2, 3 \ldots\ldots m$; $k = 1, 2, 3 \ldots\ldots m$

If we perform RBD then,

The mathematical model of RBD is $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}'$; $i = 1, 2, 3 \ldots\ldots m$, $j = 1, 2, 3 \ldots\ldots m$

(i) When row is taken as block:

$$\frac{\sigma_e'^2}{\sigma_e^2} = \frac{(m-1)\,MSE + MSC}{mMSE}$$

If $\frac{\sigma_e'^2}{\sigma_e^2} < 1$ then LSD is less efficient than RBD.

If $\frac{\sigma_e'^2}{\sigma_e^2} > 1$ then LSD is more efficient than RBD.

If $\frac{\sigma_e'^2}{\sigma_e^2} = 1$ then LSD and RBD are equally effective.

When column is taken as block:

(ii) $$\frac{\sigma_e'^2}{\sigma_e^2} = \frac{(m-1)\,MSE + MSR}{mMSE}$$

If $\frac{\sigma_e'^2}{\sigma_e^2} < 1$ then LSD is less efficient than RBD.

If $\frac{\sigma_e'^2}{\sigma_e^2} > 1$ then LSD is more efficient than RBD.

If $\frac{\sigma_e'^2}{\sigma_e^2} = 1$ then LSD and RBD are equally effective.

Example 11: From the following ANOVA table of 4×4 LSD determine it's efficiency with respect to RBD.

S.V.	d.f.	S.S.	M.S.S.
Rows	3	2.133	0.711
Columns	3	2.203	0.734
Treatments	3	10.663	3.554
Error	6	7.059	1.177
Total	15	22.058	

Solution:

$m = 4$, $MSR = 0.711$, $MSC = 0.734$, $MST = 3.554$, $MSE = 1.177$

When row is taken as block

$$\frac{\sigma_e'^2}{\sigma_e^2} = \frac{(m-1)\,MSE + MSC}{mMSE} = \frac{3 \times 1.177 + 0.734}{4 \times 1.177} = 0.905 = 90.5\%$$

Hence, LSD is 9.5% less efficient than RBD when row is taken as block.

When column is taken as block

$$\frac{\sigma_e'^2}{\sigma_e^2} = \frac{(m-1)\,MSE + MSR}{mMSE} = \frac{3 \times 1.177 + 0.711}{4 \times 1.177} = 0.901 = 90.1\%$$

Hence LSD is 9.9% less efficient than RBD when column is taken as block.

Missing plot for LSD

Let us consider a $m \times m$ LSD. Let one of the observation say x occurring in i^{th} row, j^{th} column and k^{th} treatment is is missing.

Let, G' = total of all known $m^2 - 1$ values.

R' = total of all known values of i^{th} row

C' = total of all known values of j^{th} column

T' = total of all known values of k^{th} treatment.

	Column						$R_{i..}$
Row	y_{111}	y_{122}	y_{133}	y_{1j4}	y_{1mm}		$R_{1..}$
	y_{212}	y_{223}	y_{234}	y_{2j5}	y_{2m1}		$R_{2..}$
	y_{313}	y_{324}	y_{335}	y_{3j6}	y_{3m2}		$R_{3..}$
	:	:	:	:	:		
	y_{i14}	y_{i25}	y_{i36}	x	y_{im3}		$R_{i..}=R'+x$
	:	:	:	:	:		
	Y_{m1m}	y_{m21}	y_{m32}.....	y_{mj3}.........	y_{mm6}		$R_{m..}$
$C_{.j.}$	$C_{.1.}$	$C_{.2.}$	$C_{.3.}$	$C_{.j.}=C'+x$	$C_{.m.}$		$G'+x$

$$x = \frac{m(R' + C' + T') - 2G'}{(m-1)(m-2)}$$

Substitute the value of x in place of missing value and carry out analysis as usual except that one degree of freedom is substracted from total and consequently from error. Because of the change in level of degree of freedom we obtain an upward bias in SST .Hence to get better result substract an adjustment factor from SST.

$$\text{Adjustment factor (k)} = \frac{\{(m-1)T' + R' + C' - G'\}^2}{\{(m-1)(m-2)\}^2}$$

Adjusted SST (SST_A) = SST – k.

Example 12: Determine the missing value and carryout ANOVA of following design.

D	20.1	B	19.4	C	30.6	A	7.9
C	17.5	A	10.4	D	21.2	B	19.1
A	?	D	18.1	B	24.6	C	25.2
B	25.1	C	30.4	A	10.2	D	28.0

Solution:

								$T_{i..}$
D	20.1	B	19.4	C	30.6	A	7.9	78
C	17.5	A	10.4	D	21.2	B	19.1	68.2
A	(x)	D	18.1	B	24.6	C	25.2	67.9+x
B	25.1	C	30.4	A	10.2	D	28.0	93.7
$T_{.j.}$	62.7+x		78.3		86.6		80.2	307.8+x

Here,

$m = 4, R' = 67.9, C' = 62.7, T' = 10.4 + 10.2 + 7.9 = 28.5, G' = 307.8$

$$\text{Missing value } (x) = \frac{m(R' + C' + T') - 2G'}{(m-1)(m-2)}$$

$$= \frac{4(67.9 + 62.7 + 28.5) - 2 \times 307.8}{3 \times 2}$$

$$= \frac{636.4 - 615.6}{6} = 3.46$$

Problem to test

H_{0R}: There is no significant difference between rows

H_{1R}: There is significant difference between rows.

H_{1C}: There is no significant difference between columns

H_{1C}: There is significant difference between columns.

H_{0T}: There is no significant difference between treatments

H_{1T}: there is significant difference between treatments.

Now,

$G = G' + x = 307.8 + 3.46 = 311.26$

$N = m^2 = 4^2 = 16$

$$C.F. = \frac{G^2}{N} = \frac{(311.26)^2}{16} = 6055.17$$

$$\sum_{i,j,k} y_{ijk}^2 = 20.1^2 + 19.4^2 + 30.6^2 + 7.9^2 + 17.5^2 + 10.4^2 + 21.2^2 + 19.1^2 + 3.46^2 + 18.1^2 + 24.6^2 + 25.2^2 + 25.1^2 + 30.4^2 + 10.2^2 + 28.0^2$$

$$= 7029.79$$

$$TSS = \sum_{i,j,k} y_{ijk}^2 - C.F.$$

$$= 7029.79 - 6055.17 = 974.62$$

$$SSR = \frac{1}{m} \sum_{i} T_{i..}^2 - C.F.$$

$$= \frac{1}{4} \{78^2 + 68.2^2 + 71.36^2 + 93.7^2\} - 6055.17 = 6151.79 - 6055.17$$

$$= 96.62$$

$$SSC = \frac{1}{m} \sum_{j} T_{.j.}^2 - C.F.$$

$$= \frac{1}{4} \{66.16^2 + 78.3^2 + 86.6^2 + 80.2^2\} - 6055.17 = 6109.9 - 6055.17$$

$$= 54.73$$

$$T_{..A} = 3.46 + 10.4 + 10.2 + 7.9$$

$$= 31.96$$

$T_{..B} = 25.1+19.4+24.6+19.1$

$= 88.2$

$T_{..C} = 17.5+30.4+30.6+25.2$

$= 103.7$

$T_{..D} = 20.1+18.1+21.2+28$

$= 87.4$

$$SST = \frac{1}{m}\sum_{i} T_{..k}^2 - C.F.$$

$$= \frac{1}{4}\{31.96^2 + 88.2^2 + 103.7^2 + 87.4^2\} - 6055.17$$

$$= 6798.28 - 6055.17$$

$$= 743.11$$

$$\text{Adjustment factor (k)} = \frac{\{(m-1)T' + R' + C' - G'\}^2}{\{(m-1)(m-2)\}^2}$$

$$= \frac{\{3 \times 28.5 + 67.9 + 62.7 - 307.8\}^2}{\{3 \times 2\}^2} = 233.58$$

$$SST_A = SST - k$$

$$= 743.11 - 233.58 = 509.53$$

$$SSE = TSS - SSR - SSC - SST_A$$

$$= 974.62 - 96.62 - 54.73 - 509.53$$

$$= 313.94$$

ANOVA table

S.V.	d.f.	S.S.	M.S.S.	F_{Cal}	F_{Tab}
Rows	3	96.62	32.2	0.513	$F_{0.05(3,5)} = 5.41$
Columns	3	54.73	18.24	0.29	$F_{0.05(3,5)} = 5.41$
Treatments	3	509.53	169.84	2.706	$F_{0.05(3,5)} = 5.41$
Error	5	313.74	62.748		
Total	14				

Decision

$F_R = 0.513 < F_{0.05(3,5)} = 5.41$, accept H_{0R} at 5% level of significance.

$F_C = 0.29 < F_{0.05(3,5)} = 5.41$, accept H_{0C} at 5% level of significance.

$F_T = 2.706 < F_{0.05(3,5)} = 5.41$, accept H_{0T} at 5% level of significance.

Conclusion

There is no significant difference between rows. There is no significant difference between columns. There is no significant difference between treatments.

EXERCISE

1. Describe basic principles of experimental design.
2. Explain the terms with examples; Experiment, Treatments, Experimental units, Blocks, Experimental error, Precision.
3. What do you mean by CRD? Write down it's advantages and disadvantages.
4. What do you mean by RBD? Write down it's advantages and disadvantages.
5. What do you mean by LSD? Write down it's advantages and disadvantages.
6. Differentiate between CRD and RBD.
7. Differentiate between RBD and LSD.
8. What are the assumptions underlying the analysis of the results of LSD.
9. Clearly state the restrictions that are being imposed on the number of treatments and number of replications in CRD, RBD and LSD.
10. What is meant by relative efficiency? Give expression for efficiency of (i) LSD relative to RBD. (ii) LSD relative to CRD (iii) RBD relative to CRD.
11. Carry out ANOVA of following output of wheat per field obtained as a result of 3 varieties of wheat A, B and C.

A 10	B 5	A 20	C 15
B 6	A 15	C 11	B 10
C 22	B 12	C 18	A 16

Ans: $F_T = 4.793$, insig.

12. Carry out ANOVA for following design.

A 8	C 10	A 6	B 10
C 12	B 8	B 9	A 8
B 10	A 8	C 10	C 9

Also calculate the relative efficiency of the design with respect to CRD.

Ans: $F_T = 7.97$ insig., $F_B = 1.6$, insig., Efficiency = 1.163

13. Set up the analysis of variance for the following results of a design.

A 10	B 15	C 20
B 25	C 10	A 15
C 25	A 20	B 15

Also calculate the efficiency of the design over i) RBD ii) CRD.

Ans: $F_R = 0.25$, $F_C = 0.25$, $F_T = 0.142$ insig. Efficiency = 0.625, 0.75, 0.75

14. The table given below are yields of 3 varieties in a 4 replicate experiment for which one observation is missing. Estimate the missing value and then analyse the data.

P 19	R 29	P 23	Q 33
Q 26	P ?	Q 27	R 26
R 21	Q 28	R 22	P 26

Ans: 25.3, $F_T = 4.69$, $F_B = 4.72$, insig.

15. The table given below represents the yields of 4 varieties in a 4 replicate experiment for which one observation is missing. Estimate the missing value and then carry out the ANOVA

A	12	C	19	B	10	D	8
C	18	B	12	D	6	A	?
B	22	D	10	A	5	C	21
D	12	A	7	C	27	B	17

Ans: 16.67, FR = 0.26, FC = 0.559, $F_T = 3.32$, insig.

16. Complete the following table for the analysis of variance of a design.

S.V.	d.f.	S.S.	M.S.S.	F
Blocks	4	26.8	?	?
Treatment	3	?	?	?
Error	?	?	2.5	
Total	?	85.3		

Ans: 12, 19, 28.5, 30, 6.7, 9.5, 2.68, 3.8

17. Fill in the blanks in the following analysis of variance table of a design.

Source of Variation	d.f.	S.S.	M.S.S.	F
Rows	?	72	?	2
Columns	?	?	36	?
Treatments	?	180	?	?
Error	6	?	12	
Total	?	?		

Ans: 3, 3, 3, 15, 108, 72, 432, 24, 60, 3, 5

18. Complete the following table for analysis of variance of a design.

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Columns	5	?	?	?
Rows	?	4.2	?	?
Treatments	?	?	2.43	?
Error	?	?	0.65	
Total	?	39.65		

The columns as representing schools, the rows as classes, the treatments as methods of teaching and the observations as grades based on 100 points. Test the hypothesis that the treatment effects are equal to zero.

Ans: 5, 5, 20, 35, 10.3, 12.15, 13, 2.06, 0.84, 3.16, 1.29, 3.73, Reject H_{0T}

19. Consider the partially completed ANOVA table below. Complete the ANOVA table and answer the followings. What design was employed? How many treatments were compared? How many observations were analyzed? At 0.05 level of significance, can one conclude that the treatments have different effects? Why?

Source of variation	Sum of Square	Degree of freedom	Mean Square	F
Treatments	231.5	2	?	?
Blocks	?	7	?	?
Error	573.75	?	?	
Total	903.75	23		

Ans: 98.5, 14, 115.75, 14.07, 40.98, 2.82, 0.34, Accept H_{0T}

20. From the following ANOVA table of RBD, determine it's efficiency with respect to CRD.

Source	D.F.	S.S.	M.S.S.
Between Blocks	5	21.55	4.31
Between Treatments	3	15.66	5.22
Error	15	12.3	0.82
Total	23	49.51	

Ans: 1.925

21. From the following ANOVA table of LSD , determine it's efficiency i)with respect to CRD ii) with respect to RBD when columns are taken as blocks iii) with respect to RBD when rows are taken as blocks.

Source of variation	Degree of freedom	Sum of squares	Mean sum of squares
Rows	3	259.5375	86.4375
Columns	3	155.2725	51.7575
Treatments	3	1372.1225	457.3742
Error	6	156.3700	26.0616
Total	15	1943.0775	

Ans: 1.6605, 1.6605, 1.2464

Using Software

CRD

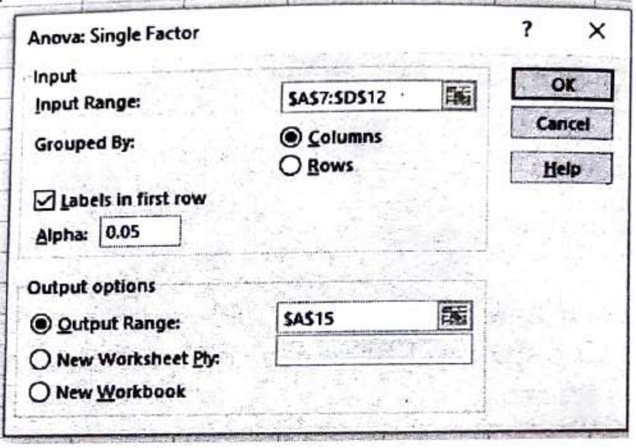
The yield of treatments in different plots are as shown in the following plots. Carry out analysis.

t_4 1401	t_3 2536	t_3 2459	t_1 2537	t_3 2827	t_1 2069
t_2 2211	t_1 1797	t_4 1170	t_4 1516	t_4 2104	t_3 2385
t_2 3366	t_1 2104	t_2 2591	t_3 2460	t_4 1077	t_2 2544

Using excel

Data\Data analysis\Anova: Single Factor

	A	B	C	D	E	F
1	t_4 1401	t_3 2536	t_3 2459	t_1 2537	t_3 2827	t_1 2069
2	t_2 2211	t_1 1797	t_4 1170	t_4 1516	t_4 2104	t_3 2385
3	t_2 3366	t_1 2104	t_2 2591	t_3 2460	t_4 1077	t_2 2544
4						
5						
6						
7	t1	t2	t3	t4		
8	2537	2211	2536	1401		
9	2069	3366	2459	1170		
10	1797	2597	2827	1516		
11	2104	2544	2385	2104		
12	1077		2460			



	A	B	C	D	E	F	G
15	Anova: Single Factor						
16							
17	SUMMARY						
18	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
19	t1	5	9584	1916.8	290618.2		
20	t2	4	10718	2679.5	238647		
21	t3	5	12667	2533.4	29788.3		
22	t4	4	6191	1547.75	158217.6		
23							
24							
25	ANOVA						
26	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
27	Between Groups	3567435	3	1189145	6.734041	0.004844	3.343889
28	Within Groups	2472220	14	176587.1			
29							
30	Total	6039654	17				
31							

Problem to test

$H_{0T}: \mu_{1.} = \mu_{2.} = \mu_{3.} = \mu_{4.}$

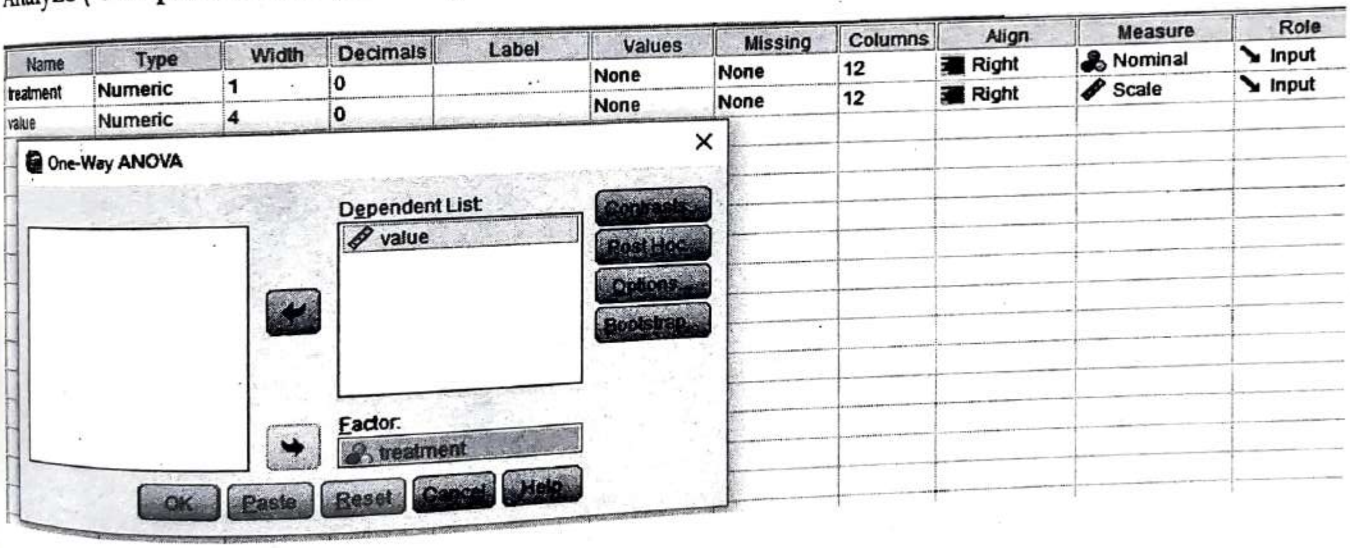
H_{1T}: At least one of the $\mu_{i.}$ is different, i = 1,2,3,4 (1 = t_1, 2 = t_2, 3 = t_3, 4 = t_4)

	A	B	C	D	E	F	G
31	α	0.05					
32							
33	Decision			Formula			
34	the null hypothesis H0 is rejected			=IF(E26<G26,"there is no reason to reject Null hypothesis H0", "the null hypothesis H0 is rejected")			
35							
36	it is significant			=IF(F26<B31,"it is significant", "it is not significant")			
37							

Using SPSS

Create variable for treatment and its values

Analyze\Compare means\One way anova



Select LSD in Post HOC test

ANOVA

value

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3561394.361	3	1187131.454	6.720	.005
Within Groups	2473236.750	14	176659.768		
Total	6034631.111	17			

Multiple Comparisons

Dependent Variable: value
LSD

(I) treatment	(J) treatment	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
1	2	-761.200*	281.952	.017	-1365.93	-156.47
	3	-616.600*	265.827	.036	-1186.74	-46.46
	4	369.050	281.952	.212	-235.68	973.78
2	1	761.200*	281.952	.017	156.47	1365.93
	3	144.600	281.952	.616	-460.13	749.33
	4	1130.250*	297.203	.002	492.81	1767.69
3	1	616.600*	265.827	.036	46.46	1186.74
	2	-144.600	281.952	.616	-749.33	460.13
	4	985.650*	281.952	.004	380.92	1590.38
4	1	-369.050	281.952	.212	-973.78	235.68
	2	-1130.250*	297.203	.002	-1767.69	-492.81
	3	-985.650*	281.952	.004	-1590.38	-380.92

*. The mean difference is significant at the 0.05 level.

The result is significant so we have to perform Post HOC test to determine the cause of difference. Here treatment 1 and 4 have similar effect and they are different from treatments 2 and 3.

Using STATA

oneway value treatment

```
. oneway value treatment

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F

Between groups     3561394.36      3   1187131.45      6.72     0.0049
 Within groups     2473236.75     14   176659.768

    Total          6034631.11     17   354978.301

Bartlett's test for equal variances:  chi2(3) =   4.0595  Prob>chi2 = 0.255
```

For pairwise comparison

oneway value treatment, bonferroni tabulate

```
oneway value treatment, bonferroni tabulate
```

treatment	Summary of value Mean	Std. Dev.	Freq.
1	1916.8	539.09016	5
2	2678	488.86194	4
3	2533.4	172.59287	5
4	1547.75	397.76574	4
Total	2175.2222	595.80055	18

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	3561394.36	3	1187131.45	6.72	0.0049
Within groups	2473236.75	14	176659.768		
Total	6034631.11	17	354978.301		

Bartlett's test for equal variances: chi2(3) = 4.0595 Prob>chi2 = 0.255

Comparison of value by treatment
(Bonferroni)

Row Mean-Col Mean	1	2	3
2	761.2 0.104		
3	616.6 0.216	-144.6 1.000	
4	-369.05 1.000	-1130.25 0.012	-985.65 0.021

Two way ANOVA

The following table gives the result of the experiment on four varieties of a crop in 5 blocks of plot.

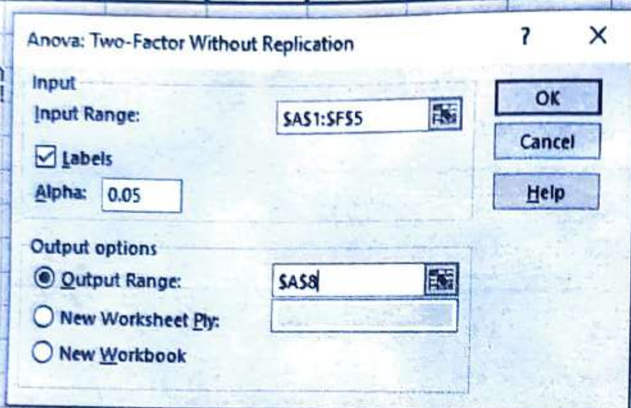
Block I		Block II		Block III		Block IV		Block V	
A	32	B	33	D	30	A	35	C	36
B	34	C	34	C	35	C	32	D	29
C	31	A	34	B	36	B	37	A	37
D	29	D	26	A	33	D	28	B	35

Analyse the above result to test whether there is significant difference between yields of four varieties.

Using EXCEL

Data\Data analysis\Anova: Two Factor without Replication

Treatment	Block 1	Block 2	Block 3	Block 4	Block 5
A	32	34	33	35	37
B	34	33	36	37	35
C	31	34	35	32	36
D	29	26	30	28	29



Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
A	5	171	34.2	3.7
B	5	175	35	2.5
C	5	168	33.6	4.3
D	5	142	28.4	2.3
Block 1	4	126	31.5	4.333333
Block 2	4	127	31.75	14.91667
Block 3	4	134	33.5	7
Block 4	4	132	33	15.33333
Block 5	4	137	34.25	12.91667

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	134	3	44.66667	18.16949	9.3E-05	3.490295
Columns	21.7	4	5.425	2.20678	0.129622	3.259167
Error	29.5	12	2.458333			
Total	185.2	19				

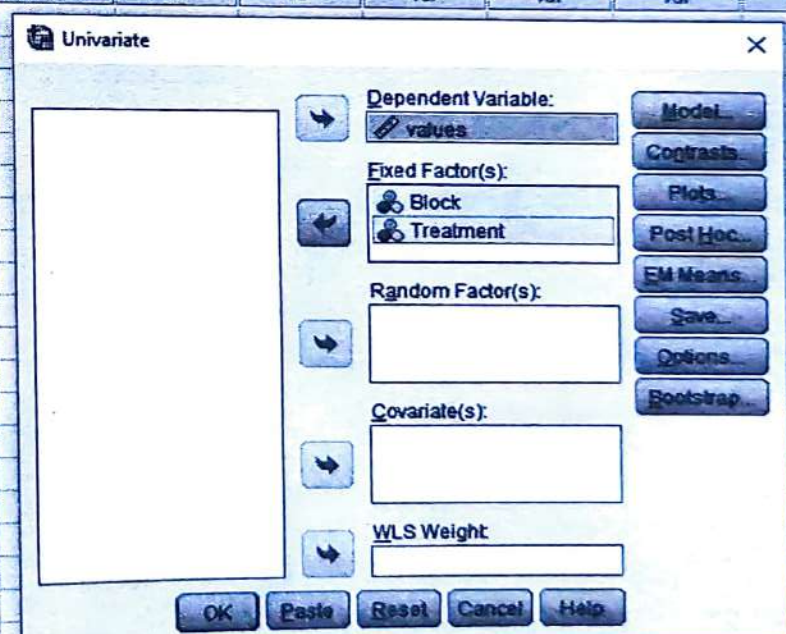
	A	B	C	D	E	F	G	H
29	Total	185.2	19					
30								
31	α	0.05						
32								
33	Decision			Formula				
34	the null hypothesis H0 is rejected			=IF(E25<G25,"there is no reason to reject Null hypothesis H0", "the null hypothesis H0 is rejected")				
35	there is no reason to reject Null hypothesis H0			=IF(E26<G26,"there is no reason to reject Null hypothesis H0", "the null hypothesis H0 is rejected")				
36	it is significant			=IF(F25<B31,"it is significant", "it is not significant")				
37	it is significant			=IF(F26<B31,"it is significant", "it is not significant")				

Using **SPSS**

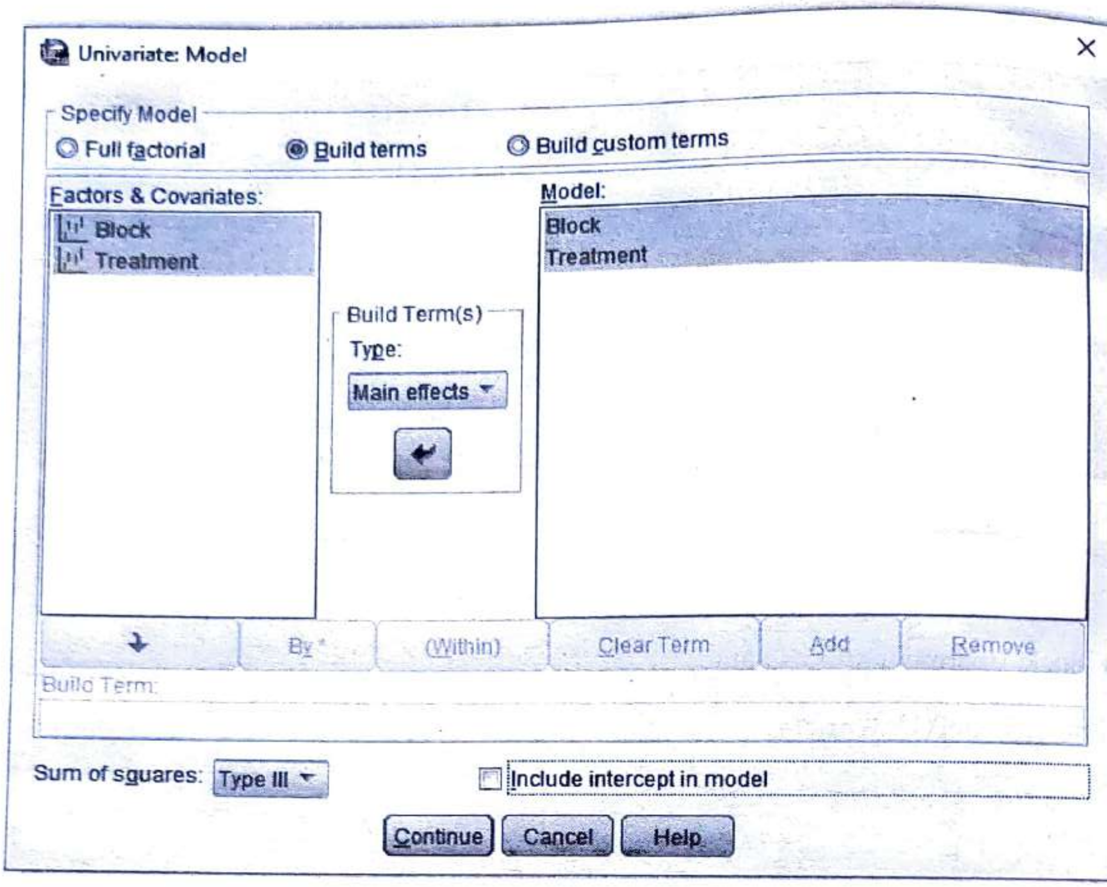
Create variable for block, treatment and its values

Analyze\General linear model\Univariate

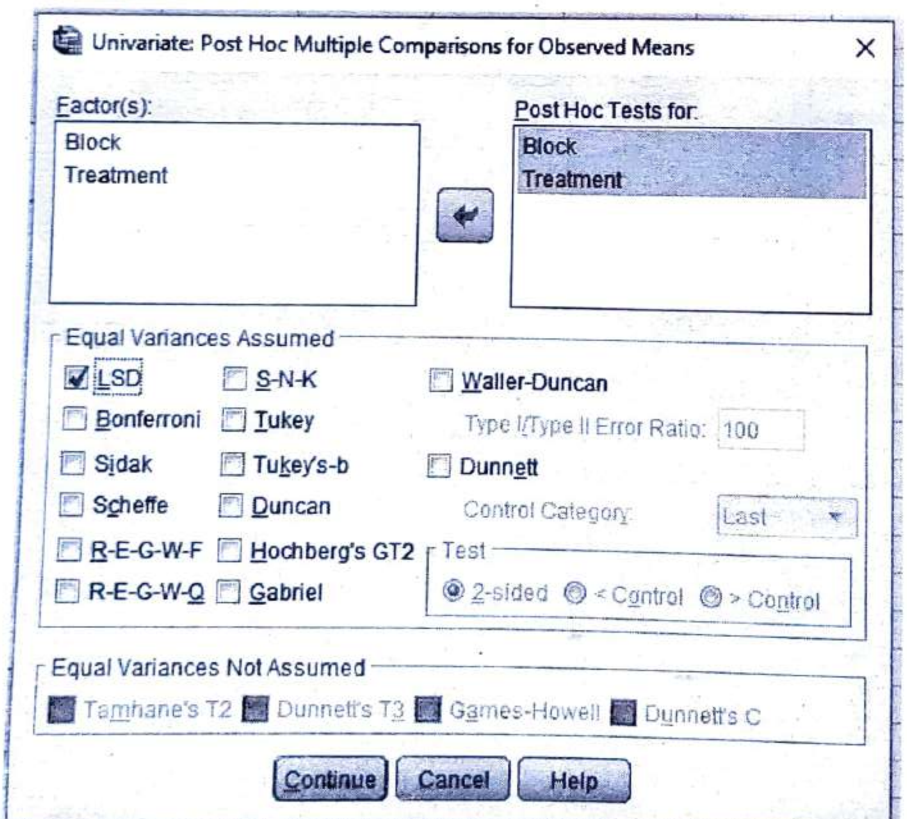
	Block	Treatment	values
1	1	A	32
2	1	B	34
3	1	C	31
4	1	D	29
5	2	A	34
6	2	B	33
7	2	C	34
8	2	D	26
9	3	A	33
10	3	B	36
11	3	C	35
12	3	D	30
13	4	A	35
14	4	B	37
15	4	C	32
16	4	D	28
17	5	A	37
18	5	B	35
19	5	C	36
20	5	D	29



Click model



Click Post HOC



outputs

Between-Subjects Factors

		Value Label	N
Block Block	1		4
	2		4
	3		4
	4		4
	5		4
Treatment Treatment	1	A	5
	2	B	5
	3	C	5
	4	D	5

Tests of Between-Subjects Effects

Dependent Variable: values values

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	21672.500[a]	8	2709.062	1101.992	.000
Block	21.700	4	5.425	2.207	.130
Treatment	134.000	3	44.667	18.169	.000
Error	29.500	12	2.458		
Total	21702.000	20			

a. R Squared = .999 (Adjusted R Squared = .998)

Multiple Comparisons

Dependent Variable: values values

LSD

(I) Block	(J) Block	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
1	2	-.25	1.109	.825	-2.67	2.17
	3	-2.00	1.109	.096	-4.42	.42
	4	-1.50	1.109	.201	-3.92	.92
	5	-2.75*	1.109	.029	-5.17	-.33
2	1	.25	1.109	.825	-2.17	2.67
	3	-1.75	1.109	.140	-4.17	.67
	4	-1.25	1.109	.282	-3.67	1.17
	5	-2.50*	1.109	.044	-4.92	-.08
3	1	2.00	1.109	.096	-.42	4.42
	2	1.75	1.109	.140	-.67	4.17
	4	.50	1.109	.660	-1.92	2.92
	5	-.75	1.109	.512	-3.17	1.67
4	1	1.50	1.109	.201	-.92	3.92
	2	1.25	1.109	.282	-1.17	3.67
	3	-.50	1.109	.660	-2.92	1.92
	5	-1.25	1.109	.282	-3.67	1.17
5	1	2.75*	1.109	.029	.33	5.17
	2	2.50*	1.109	.044	.08	4.92
	3	.75	1.109	.512	-1.67	3.17
	4	1.25	1.109	.282	-1.17	3.67

Based on observed means.

The error term is Mean Square(Error) = 2.458.

*. The mean difference is significant at the 0.05 level.

Using STATA

anova values Block Treatment

anova values Block Treatment

```
Number of obs =         20     R-squared     =  0.8407
Root MSE      =    1.56791     Adj R-squared =  0.7478
```

Source	Partial SS	df	MS	F	Prob>F
Model	155.7	7	22.242857	9.05	0.0006
Block	21.7	4	5.425	2.21	0.1296
Treatment	134	3	44.666667	18.17	0.0001
Residual	29.5	12	2.4583333		
Total	185.2	19	9.7473684		

LSD

The following is the 5 × 5 Latin square design for data taken from a manurial experiment with sugarcane. The five treatments were A=no manure; B= an inorganic manure;

C,D and E =three levels of farm yard manure.

Plan and yield of sugarcane (in a suitable unit) per plot.

Row	Column				
	I	II	III	IV	V
I	A 52.5	E 46.3	D 44.1	C 48.1	B 40.9
II	D 44.2	B 42.9	A 51.3	E 49.3	C 32.6
III	B 49.1	A 47.3	C 38.1	D 41.0	E 47.2
IV	C 43.2	D 42.5	E 67.2	B 55.1	A 45.3
V	E 47.0	C 43.2	B 46.7	A 46.0	D 43.2

Analyse the above data to find if there are any treatment effects.

Using Excel

	A	B	C	D	E	F	G	H	I
11	Table to calculate sums								
12		columns							
13	Row	1	2	3	4	5	Grand Total	formula	
14	1	52.5	46.3	44.1	48.1	40.9	231.9	=SUM(B14:F14)	
15	2	44.2	42.9	51.3	49.3	32.6	220.3	=SUM(B15:F15)	
16	3	49.1	47.3	38.1	41	47.2	222.7	=SUM(B16:F16)	
17	4	43.2	42.5	67.2	55.1	45.3	253.3	=SUM(B17:F17)	
18	5	47	43.2	46.7	46	43.2	226.1	=SUM(B18:F18)	
19	Grand Total	236	222.2	247.4	239.5	209.2	1154.3		
20	formula	=SUM(B14:B18)	=SUM(C14:C18)	=SUM(D14:D18)	=SUM(E14:E18)	=SUM(F14:F18)	=SUM(G14:G18)		
21									
22	treatment sums								
23	Treatments	Sums	formula						
24	1	242.4	=SUMIF(Table1[Treatment],A24,Table1[Values])						
25	2	234.7	=SUMIF(Table1[Treatment],A25,Table1[Values])						
26	3	205.2	=SUMIF(Table1[Treatment],A26,Table1[Values])						
27	4	215	=SUMIF(Table1[Treatment],A27,Table1[Values])						
28	5	257	=SUMIF(Table1[Treatment],A28,Table1[Values])						
29	Grand Total	1154.3	=SUM(B24:B28)						

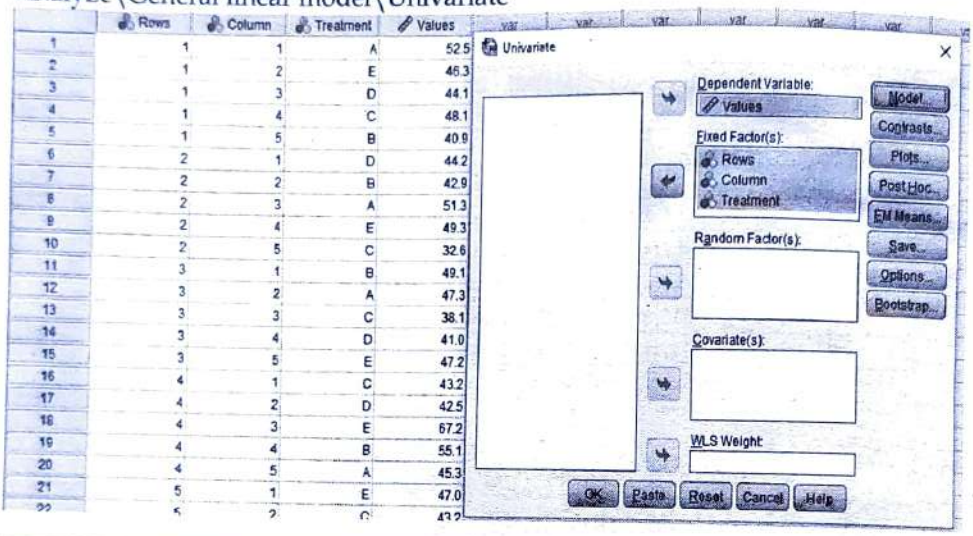
	A	B	C	D	E	F	G
31		symbol	value	formula			
32	no. of rows	m	5	=COUNT(A14:A18)			
33	no. of columns	m	5	=COUNT(B13:F13)			
34	no. of treatments	m	5	=COUNT(A24:A28)			
35	Grand total	G	1154.3	=B29			
36	correction factor	C.F.	53296.3	=C35^2/C32^2			
37	total sum of square	TSS	977.17	=SUMSQ(B14:F18)-C36			
38	sum of square due to row	SSR	141.078	=(SUMSQ(G14:G18)/C32)-C36			
39	sum of square due to column	SSC	183.758	=(SUMSQ(B19:F19)/C33)-C36			
40	sum of square due to treatment	SST	348.238	=(SUMSQ(B24:B28)/C34)-C36			
41	sum of square due to error	SSE	304.095	=C37-C38-C39-C40			
42	level of significance	α	0.05				
43							
44	ANOVA table						
45	source of variation	degrees of freedom	Sum of squares	Mean sum of square	F ratio	F tabulated	sig
46		=C32-1	=C38	=C47/B47	=D47/$D$50	=F.INV.RT($C$42,B47,$B$50)	=F.DIST.RT(E47,B47,$B$50)
47	Row	4	141.078	35.2696	1.392	3.2591667	0.29477755
48	Column	4	183.758	45.9396	1.813	3.2591667	0.19122217
49	Treatment	4	348.238	87.0596	3.435	3.2591667	0.04313196
50	Error	12	304.095	25.341267			
51	Total	24	977.17				

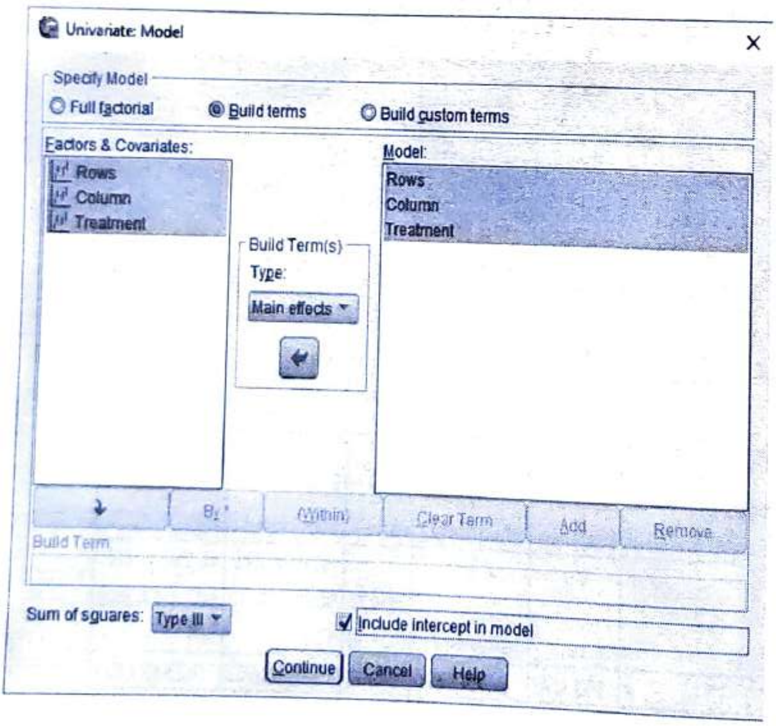
	A	B	C	D	E	F	G	H	I	J
53	Decision									
54	Significant approach									
55		1 There is no reason to reject Null Hypothesis of no significant difference between Rows,Hence it is accepted								
56		2 There is no reason to reject Null Hypothesis of no significant difference between Rows,Hence it is accepted								
57		3 Null hypothesis is rejected								
58										
59	P value approach									
60		1 It is insignificant								
61		2 It is insignificant								
62		3 It is significant								

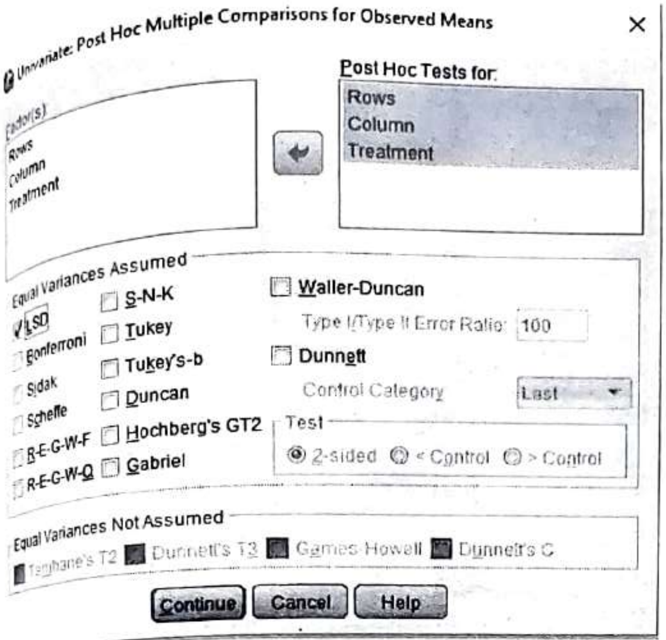
Using SPSS

Create variable for rows, columns, treatment and its values

Analyze\General linear model\Univariate







Outputs

Between-Subjects Factors

		Value Label	N
Rows Rows	1		5
	2		5
	3		5
	4		5
	5		5
Column Column	1		5
	2		5
	3		5
	4		5
	5		5
Treatment Treatment	1	A	5
	2	B	5
	3	C	5
	4	D	5
	5	E	5

Tests of Between-Subjects Effects

Dependent Variable: Values Values

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	673.075[a]	12	56.090	2.213	.092
Intercept	53296.340	1	53296.340	2103.144	.000
Rows	141.078	4	35.270	1.392	.295
Column	183.758	4	45.940	1.813	.191
Treatment	348.238	4	87.060	3.435	.043
Error	304.095	12	25.341		
Total	54273.510	25			
Corrected Total	977.170	24			

a. R Squared = .689 (Adjusted R Squared = .378)

Multiple Comparisons

Dependent Variable: Values Values
LSD

(I) Rows	(J) Rows	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
1	2	2.320	3.1838	.480	-4.617	9.257
	3	1.840	3.1838	.574	-5.097	8.777
	4	-4.280	3.1838	.204	-11.217	2.657
	5	1.160	3.1838	.722	-5.777	8.097
2	1	-2.320	3.1838	.480	-9.257	4.617
	3	-.480	3.1838	.883	-7.417	6.457
	4	-6.600	3.1838	.060	-13.537	.337
	5	-1.160	3.1838	.722	-8.097	5.777
3	1	-1.840	3.1838	.574	-8.777	5.097
	2	.480	3.1838	.883	-6.457	7.417
	4	-6.120	3.1838	.079	-13.057	.817
	5	-.680	3.1838	.834	-7.617	6.257
4	1	4.280	3.1838	.204	-2.657	11.217
	2	6.600	3.1838	.060	-.337	13.537
	3	6.120	3.1838	.079	-.817	13.057
	5	5.440	3.1838	.113	-1.497	12.377
5	1	-1.160	3.1838	.722	-8.097	5.777
	2	1.160	3.1838	.722	-5.777	8.097
	3	.680	3.1838	.834	-6.257	7.617
	4	-5.440	3.1838	.113	-12.377	1.497

Based on observed means.

The error term is Mean Square(Error) = 25.341.

Using STATA

anova Values Rows Column Treatment

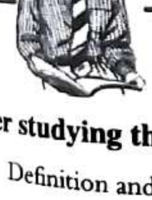
```
anova Values Rows Column Treatment

                  Number of obs =      25     R-squared     =  0.6888
                  Root MSE      = 5.03401     Adj R-squared =  0.3776

       Source |  Partial SS     df       MS           F     Prob>F
   -----------+----------------------------------------------------
        Model |  673.0752       12    56.0896        2.21   0.0916
              |
         Rows |  141.0784        4    35.2696        1.39   0.2948
       Column |  183.7584        4    45.9396        1.81   0.1912
    Treatment |  348.2384        4    87.0596        3.44   0.0431
              |
     Residual |  304.0952       12    25.341267
   -----------+----------------------------------------------------
        Total |  977.1704       24    40.715433
```

5135

6

Chapter

STOCHASTIC PROCESS



CHAPTER OUTLINE

After studying this chapter, students will be able to understand the:

- Definition and Classification.
- Markov Process, Markov chain, Matrix approach, Steady-state distribution.
- Counting Process, Binomial process, Poisson process, simulation of stochastic process.
- Queuing system: Main component of queuing system, Little's law.
- Bernulli single sever queing process, system with limited capacity.
- M/M/1 system: Evaluating the system performance.

Definition and Classification

A family of random variables indexed by a parameter such as time is called stochastic process. It is also called chance or random process. Hence stochastic process is a family of random variables $\{X(t)/t \in T\}$ defined on given probability space, indexed by the parameter t where t varies over an index set T. The values assumed by random variable X(t) are called states and the set of all possible values forms the state space of the process. The state space is denoted by I. If the state space of a stochastic process is discrete then it is called a discrete state process often referred as chain. In this case state space is assumed to be {0, 1, 2, 3 ....}. If the state space of stochastic process is continuous then it is called continuous state process. If the index set is discrete then it is called discrete parameter process and if the index set is continuous then it is called continuous parameter process.

		Index set T	
		Discrete	Continuous
State space I	Discrete	Discrete parameter stochastic chain	Continuous parameter stochastic chain
	Continuous	Discrete parameter continuous state process	Continuous parameter continuous state process

Markov Process

Stochastic process X(t) is called Markov process if for any $t_1 < t_2 < t_3 .... < t_n < t$ and any sets $A_1, A_2, A_3 ..... A_n$.

$P\{X(t) \in A\ X(t_1) \in A_1\ X(t_2) \in A_2 .... X(t_n) \in A_n\} = P[X(t) \in A, X(t_n) \in A_n\}$

P(future/ past, present) = P(future/present)

A stochastic process such that probability distribution for its future development depend only on the present state and not on how process arrived in that state. If the state space I is discrete then the Markov Process is called Markov Chain. If we further assume that parameter space T is also discrete then we get discrete parameter Markov Chain.

Markov Chain

Let $\{X_n\}$ be a sequence of values describing a mutually exclusive and exhaustive system of events. Let X_n values take only discrete the union I of all possible values of X_n is then a countable set called the state space of the process. Each element $i \in I$ is called a state. The index n is of time. The number of events may be finite or infinite the values of $\{X_n\}$ is said to be a Markov chain or Markov dependent if for all $i_0, i_1, i_2, ... i_{n-1}, i_n \in I$ and for all n.

$P(X_n = i_n\ (X_0 = i_0, X_1 = i_1 .... X_{(n-1)} = i_{(n-1)}) = P(X_n = i_n\ X_{n-1} = i_{n-1})$

The conditional distribution of X_n given the values $X_0, X_1, ..., X_{n-1}$ depends only on X_{n-1} not on the preceding values. If the state space is finite then we have finite Markov Chain.

Transition Probability

The probability of moving from one state to another or remain in the same state in a single period of time is called transition probability. The probability of moving from one state to another depends upon the probability of preceding state. Hence transition probability is a conditional probability.

$P\{X(t+1) = j/X(t) = i\} = P_{ij}(t)$ is transition probability from state i to state j in time t

Transition probability matrix

It is matrix obtained by using transition probabilities of various states.

Let the state space be I = {0, 1, 2, 3, ... n} then

$$P = \begin{pmatrix} P_{00} & P_{01} & P_{02} & P_{0n-2} & P_{0n-1} & P_{0n} \\ P_{10} & P_{11} & P_{12} & P_{1n-2} & P_{1n-1} & P_{1n} \\ P_{20} & P_{21} & P_{22} & P_{2n-2} & P_{2n-1} & P_{2n} \\ P_{n0} & P_{n1} & P_{n2} & P_{nn-2} & P_{nn-1} & P_{nn} \end{pmatrix}$$

$= \{P_{ij}\}$ i, j $\in$ I is called transition probability matrix.

Here P_{ij} is the conditional probability of moving from state i to state j . All conditional one step state probabilities can be represented as elements of a square matrix which is called transition probabilities matrix.

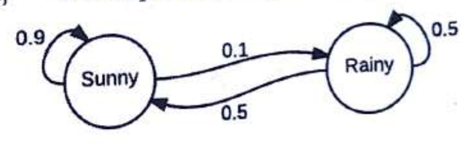
Here, $P_{ij} \geq 0$

$$\sum_i P_{ij} = 1$$

$$\sum_j P_{ij} = 1$$

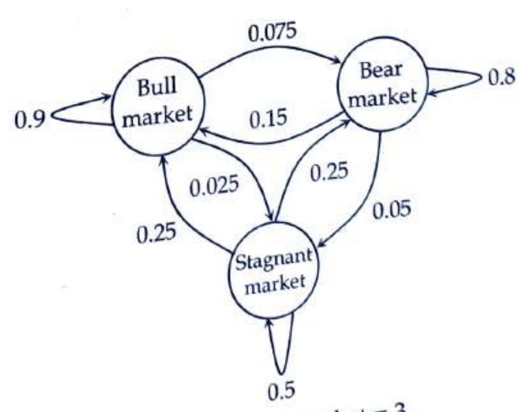
$P_{ij} = P\{X(t+1) = j/X(t) = i\}$ is transition probability

$P^{(n)}_{ij} = P\{X(t+n) = j/X(t) = i\}$ is n step transition probability



Let sunny = 1, Rainy = 2

Transition probability matrix $P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$



Let Bull market = 1, Bear mrket = 2 and stagnant market = 3.

Transition probability matrix

$$P = \begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

n Step Transition Probability

Probability $P_{ij}(t) = P\{X(t+1) = j / X(t) = i\}$
$= P\{X(t+1) = j/X(t) = i, X(t-1) = h, X(t-2) = g \ldots\}$ is called transition probability

Probability $P_{ij}(t)^n = P\{X(t+n) = j / X(t) = i\}$ of moving from state i to state j is n step transition probability

$$P_{ij}(n) = \sum_{k=1}^{m} P_{ik}^{(n-1)} P_{kj}^{(1)}$$

$$P_{ij}(2) = \sum_{k=1}^{m} P_{ik} P_{kj}$$

$$P_{ij}(3) = \sum_{k=1}^{m} \sum_{l=1}^{m} P_{jk} P_{kl} P_{lj}$$

Example 1: Find 2 step and 3 step transition probability matrix from the transition probability matrix

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Solution:

2 step transition probability matrix

$$P^{(2)} = P\,P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 0+0 \\ 0+0 & 1+0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

3 step transition probability matrix

$$P^{(3)} = P^{(2)}\,P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 0+0 \\ 0+0 & 1+0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 2: Find 2 step and 3 step transition probability matrix from the transition probability matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix}$$

Solution:

2 step transition probability matrix

$$P^{(2)} = P\,P = \begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & p \\ q & q+p & 0 \\ 0 & 0 & p \end{bmatrix} = \begin{bmatrix} q & 0 & p \\ 0 & 1 & 0 \\ q & 0 & p \end{bmatrix}$$

3 step transition probability matrix

$$P^{(3)} = P^{(2)}\,P = \begin{bmatrix} q & 0 & p \\ 0 & 1 & 0 \\ q & 0 & p \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & q+p & 0 \\ q & 0 & p \\ 0 & q+p & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ q & 0 & p \\ 0 & 1 & 0 \end{bmatrix}$$

Example 3: If $P = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$

Find $P_{12}(2)$ $P_{22}(2)$.

$$P_{12}(2) = \sum_{k=1}^{m} P_{1k}P_{k2} = P_{11}P_{12} + P_{12}P_{22}$$
$$= 0.5 \times 0.5 + 0.5 \times 0.6$$
$$= 0.25 + 0.30 = 0.55$$

$$P_{22}(2) = \sum_{k=1}^{m} P_{2k}P_{k2} = P_{21}P_{12} + P_{22}P_{22}$$
$$= 0.4 \times 0.5 + 0.6 \times 0.6$$
$$= 0.2 + 0.36 = 0.56$$

Example 4: If $P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$

Find $P_{11}(3)$, $P_{21}(3)$.

$$P_{11}(3) = \sum_{k=1}^{m} \sum_{l=1}^{m} P_{1k}P_{kl}\,P_{l1}$$
$$= P_{11}P_{11}P_{11} + P_{21}P_{12}P_{21} + P_{12}P_{21}P_{11} + P_{12}P_{22}P_{21}$$
$$= 0.7 \times 0.7 \times 0.7 + 0.7 \times 0.3 \times 0.4 + 0.3 \times 0.4 \times 0.7 + 0.3 \times 0.6 \times 0.4$$
$$= 0.583$$

$$P_{21}(3) = \sum_{k=1}^{m} \sum_{l=1}^{m} P_{2k}P_{kl}\,P_{l1}$$
$$= P_{21}P_{11}P_{11} + P_{21}P_{12}P_{21} + P_{22}P_{21}P_{11} + P_{22}P_{22}P_{21}$$
$$= 0.4 \times 0.7 \times 0.7 + 0.4 \times 0.3 \times 0.4 + 0.6 \times 0.4 \times 0.7 + 0.6 \times 0.6 \times 0.4$$
$$= 0.196 + 0.048 + 0.168 + 0.144$$
$$= 0.556$$

Example 5: In some town each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, whereas a rainy day is followed by sunny day with probability 0.4. It rains on Monday Make forecast for Tuesday, Wednesday and Thursday.

Solution:

Let state 1 = sunny and state 2 = rainy

Transition probabilities are

$P_{11} = 0.7$, $P_{12} = 0.3$, $P_{21} = 0.4$, $P_{22} = 0.6$

One step transition probability matrix

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

For Tuesday rainy chance $(P_{22}) = 0.6 = 60\%$

For Tuesday sunny chance $(P_{21}) = 0.4 = 40\%$

For Wednesday

$$P^{(2)} = PP = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.49 + 0.12 & 0.21 + 0.18 \\ 0.28 + 0.24 & 0.12 + 0.36 \end{bmatrix} = \begin{bmatrix} 0.61 & 0.40 \\ 0.52 & 0.48 \end{bmatrix}$$

For Wednesday rainy chance $(P_{22}) = 0.48 = 48\%$

For Wednesday sunny chance $(P_{21}) = 0.52 = 52\%$

For Thursday

$$P^{(3)} = P^{(2)} P = \begin{bmatrix} 0.61 & 0.4 \\ 0.52 & 0.48 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.427 + 0.16 & 0.183 + 0.24 \\ 0.364 + 0.192 & 0.156 + 0.288 \end{bmatrix} = \begin{bmatrix} 0.632 & 0.423 \\ 0.556 & 0.444 \end{bmatrix}$$

For Thursday rainy chance $(P_{22}) = 0.444 = 44.4\%$

For Thursday sunny chance $(P_{21}) = 0.556 = 55.6\%$

Steady State Distribution

A collection of limiting probabilities $\pi_x = \lim_{h \to 0} P_h(x)$ is called steady state distribution of a markov chain $X(t)$.

When limit exists, it can be used as a forecast of the distribution of X after many transitions.

When steady state distribution exists $\pi P = \pi$ and $\sum_x \pi_x = 1$.

Example 6: Obtain steady state distribution of a Markov Chain having transition probability matrix $\begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix}$

Solution:

Here $P = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix}$

Let $\pi = (\pi_1 \quad \pi_2)$

Now, $\pi P = \pi$

or $[\pi_1 \quad \pi_2] \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} = [\pi_1 \quad \pi_2]$

or $[0.2\pi_1 + 0.5\pi_2 \quad 0.8\pi_1 + 0.5\pi_2] = [\pi_1 \quad \pi_2]$

Hence $0.2\pi_1 + 0.5\pi_2 = \pi_1$ .....(i)

$0.8\pi_1 + 0.5\pi_2 = \pi_2$ ......(ii)

From (i) $0.5\pi_2 = \pi_1 - 0.2\pi_1$

or $0.5\pi_2 = 0.8\pi_1$

or $\pi_2 = 8/5\, \pi_1$

$= 1.6\, \pi_1$

Since, $\pi_1 + \pi_2 = 1$

or $\pi_1 + 1.6\pi_1 = 1$

or $2.6\, \pi_1 = 1$

or $\pi_1 = \frac{1}{2.6} = \frac{5}{13}$

$\pi_2 = \frac{1.6}{2.6} = \frac{8}{13}$

Hence in long-run probability of state 1 is 5/13 and state 2 is 8/13.

Counting Process

Counting process includes counts of arrived jobs, completed task, transmitted messages, detected errors, scored goals etc.

A stochastic process X is called counting if $X(t)$ is the number of items counted by the time t. On passes of time additional items can not be counted. Counts are non negative integers. $X(t) \in \{0, 1, 2, 3, \ldots\ldots\}$. Hence counting are discrete state.

Hence $\{X(t), t \geq 0\}$ is called a counting process if

(i) $X(0) = 0$

(ii) For $0 < t_1 < t_2 < \ldots\ldots < t_n < t$

$$X(0) < X(t_1) < X(t_2) < X(t_n) < \ldots\ldots X(t_n) < X(t)$$

(iii) $X(t_i) - X(t_{(i-1)})$ denotes the number of events in (t_{i-1}, t_i)

(iv) $X(t_i) - X(t_{(i-1)})$ are independently distributed.

Binomial Process

It is discrete time discrete space counting stochastic process. Binomial process $X(n)$ is the number of successes in the first n independent Bernoulli trials, where $n = 0, 1, 2, 3\ldots\ldots\ldots\ldots$

Let λ = arrival rate

Δ = frame size

P = probability of success (arrival) during one frame (trial)

$X(t/\Delta)$ = Number of arrivals by the time t

T = inter arrival time

The inter arrival period consists of a Geometric number of frames Y, each frame taking Δ seconds. Hence the inter arrival time can be computed as $T = Y\Delta$. It is rescaled Geometric random variable taking possible values $\Delta, 2\Delta, 3\Delta \ldots\ldots\ldots$

$$\lambda = p/\Delta$$

$$n = t/\Delta$$

$$X(n) = \text{Binomial } (n,p)$$

$$Y = \text{Geometric } (p)$$

$$T = Y\Delta$$

$$E(T) = E(Y\Delta) = \Delta\, E(Y) = \Delta/p = 1/\lambda$$

$$V(T) = V(Y\Delta) = \Delta^2\, V(Y) = (1-p)\left(\frac{\Delta}{p}\right)^2 = \frac{1-p}{\lambda^2}$$

Example 7: Suppose that a number of defects coming from an assembly line can be modeled as a Binomial counting process with frames of one-half-minute length and probability $p = 0.02$ of a defect during each frame.

i) Find the probability of going more than 20 minutes without a defect.

ii) Determine the arrival rate in units of defects per hour.

iii) If the process is stopped for inspection each time a defect is found, on average how long will the process run until it is stopped?

Solution:

Let X_n = Number of defects in n frames

Here, $p = 0.02$

$\Delta = 0.5$ minutes

T = time between two successive defects

Solution:

i) For $t = 20$ minutes $n = \frac{t}{\Delta} = \frac{20}{0.5} = 40$

$P\{X(40) = 0\} = C(40,0)\ 0.02^{\circ}(1 - 0.02)^{40} = 0.446$

ii) Since 1 hour has 120 frames, λ = no. of defects per hour = np = 120(0.02) = 2.4 defects per hour.

iii) $E(T) = \frac{\Delta}{p} = \frac{0.5}{0.02} = 250.5$

Example 8: Customers come to a self-service gas station at the rate of 20 per hour. Their arrivals are modeled by a Binomial counting process.

i) How many frames per hour should we choose, and what should be the length of each frame if the probability of an arrival during each frame is to be 0.05?

ii) With this frames, find the expected value and standard deviation of the time between arrivals at the gas station.

Solution:

Arrival rate $\lambda = 20\ hr^{-1}$.

i) $p = 0.05$.

Δ = duration of 1 frame = $p/\lambda = 0.05/(20\ hrs^{-1}) = (1/400)$ hrs = 9 sec.

Also, n = number of frames in 1 hr = $1\ hr/\Delta = 400$ frames.

ii) Let T = inter-arrival time. $E(T) = \Delta/p = 9/0.05 = 180$ sec $= 180/60 = 3$ min.

$SD(T) = (\Delta/p)\sqrt{1-p} = 180\sqrt{1-0.05} = 175.44$ sec $= 175.44/60 = 2.92$ min.

Example 9: Jobs are sent to a mainframe computer at a rate of 4 jobs per minute. Arrivals are modeled by a Binomial counting process.

i. Choose a frame size that makes the probability of a new jobs received during each frame equal to 0.1

ii. Using the chosen frame compute the probability of more than 4 jobs received during one minute

iii. What is probability of more than 20 jobs during 5 minutes

iv. What is average inter arrival time and variance?

v. What is probability that next job does not arrive during next 30 seconds?

Solution:

Here, $\lambda = 4$ per minute, $p = 0.1$

(i) $\Delta = p/\lambda = 0.1/4 = 0.025$ min

For $t = 1$, $n = t/\Delta = 1/.025 = 40$ frames

$n = 40$, $p = 0.1$

(ii) $P(X(n) > 4) = 1 - P(X(n) \leq 4) = 1 - \left[\sum_{x=0}^{4} {}^{40}C_x (0.1)^x (0.9)^{40-x}\right]$

$= 1 - [(0.9)^{40} + 40 \times 0.1 \times (0.9)^{39} + 780 \times (0.1)^2 (0.9)^{38} + 9880 \times (0.1)^3 (0.9)^{37} + 91350 \times (0.1)^4 (0.9)^{36}] = 0.37$

(iii) $P(X(n) > 20) = P(X(n) > 20.5)$

Using continuity correction

$= P\left\{\frac{x(n) - np}{\sqrt{npq}} > \frac{20.5 - 200 \times 0.1}{\sqrt{200 \times 0.1 \times 0.9}}\right\}$

$= P(Z > 0.12) = 0.5 - P(0 < z < 0.12) = 0.5 - 0.0478 = 0.4522$

(iv) $E(T) = 1/\lambda = 1/4 = 0.25$ min = 15 sec

$V(T) = \frac{1 - p}{\lambda^2} = \frac{0.9}{4^2} = 0.056$

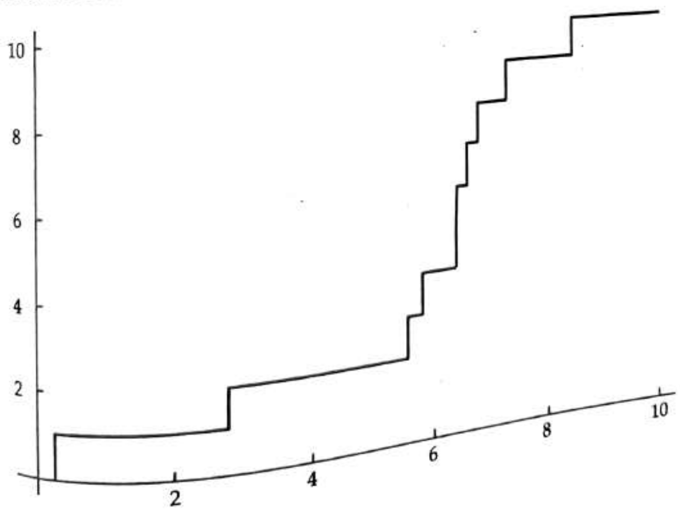
(v) $T = \Delta Y = 0.025 Y$

$P(T > 30 \text{ sec}) = P(T > 0.5 \text{ min}) = P[Y(0.025) > 0.5]$

$= P(Y > 20) = (1 - P)^n = (1 - 0.1)^{20} = 0.1215$

Poisson Process

It is limiting case of Binomial process as binomial counting process Δ tends to 0. Poisson process is a continuous time counting stochastic process obtained from Binomial counting process when its frame size Δ decreases to 0 while the arrival rate λ remains constant.



Let $X(t)$ = No. of arrivals occurring until time t

T = inter arrival time

T_k = time of k^{th} arrival.

$X(t)$ = Poisson(λt)

T = Exponential (λ)

$T_k = \text{Gamma}(k, \lambda)$

$E\,X(t) = np = \frac{t}{\Delta} p = \lambda t$

$V\,X(t) = \lambda t \qquad F_T(t) = 1 - e^{-\lambda_t}$

Prob. of k^{th} arrival before time t

$P(T_k \leq t) = P[X(t) \geq k]$
$P(T_k > t) = P[X(t) < k]$

Example 10: The number of hits to a certain web site follows Poisson process with 5 hits per minute.0

i) What is time required to get 5000 hits?

ii) What is probability that hitting occurs within 12 hours?

Solution:

Number of hits $k = 5000$, $\lambda = 5\ \text{min}^{-1}$

Expected time $= \frac{k}{\lambda} = \frac{5000}{5} = 1000$ minutes.

Standard deviation $(\sigma) = \frac{\sqrt{k}}{\lambda} = 14.14$

$P(T_k < 12\text{ hrs.}) = P(T_k < 720) = P\left(\frac{T_k - \mu}{\sigma} < \frac{720 - \mu}{\sigma}\right)$

$= P\left(z < \frac{720 - 1000}{14.14}\right) = P(Z < -19.44) = 0$

Example 11: Customers arrive at a shop at the rate of 2 per minute. Find (i) expected number of customers in a 5 minute period (ii) the variance of the number of customers in the same period (iii) the probability that there will be at least one customer.

Solution:

Here, $\lambda = 2$
$t = 5$

(i) $E(X) = \lambda t = 5 \times 2 = 10$

(ii) $V(X) = \lambda t = 5 \times 2 = 10$

(iii) $P\{X(5) \geq 1\} = 1 - P\{X(5) < 1\}$
$= 1 - P\{X(5) = 0\}$
$= 1 - e^{-10} = 0.999$

Example 12: Shipments of paper arrive at a printing shop according to a Poisson process at a rate of 0.5 shipments per day.

i) Find the probability that the printing shop receives more than two shipments in a day.

ii) If there are more than 4 days between shipments, all the paper will be used up and the presses will be idle. What is the probability that this will happen?

Solution:

Arrival rate $\lambda = 0.5$ per day.

$X(t)$ = number of arrivals (shipments) in t days, it is Poisson (0.5t)

T = inter-arrival time measured in days, it is Exponential (0.5).

i) $P[X(1) > 2] = 1 - P[X(1) \leq 2] = 1 - [P\{X(1) = 0\} + P\{X(1)=1\} + P\{X(1) = 2\}]$

$$= 1 - \left\{\frac{e^{-0.5}\,0.5^0}{0!} + \frac{e^{(-0.5)}\,(0.5)^1}{1!} + \frac{e^{(-0.5)}\,(0.5)^2}{2!}\right\}$$

$$= 1 - e^{-0.5}\{1 + 0.5 + 0.125\}$$

$$= 1 - 0.6065 \times 1.625 = 0.014$$

(ii) $P[T > 4] = \int_4^{\infty} 0.5e^{-0.5t}\,dt = 0.5\left[\frac{e^{-0.5t}}{-0.5}\right]_4^{\infty} = e^{-0.5\times4} = 0.135$

Simulation of Stochastic Process

It requires lengthy computation process to determine different characteristics of stochastic process.

It is used in prediction of future behaviour of stochastic process. Monte Carlo method is used for simulation process. Simulation is carried out for discrete time process, market chain continuous time process, possion process etc.

Queuing system

It is facility consisting of one or several servers designed to perform certain tasks or process certain jobs and a queue of jobs waiting to be processed.

Jobs arrive at the queuing system, wait for an available server, get processed by the server and leave.

Examples are

- A medical office serving patients
- An internet service provider whose customers connect to the internet, browse and disconnect
- A printer processing job sent to it from different computers
- A TV channel viewed by many people at various times
- A personal or shared computer executing tasks sent by its users

Features of Queue

Calling population:
It is infinite population with independent arrivals and not influenced by queuing system.

Arrival process:
Arrival rate follows Poisson distribution with parameter λ.

Queuing configuration:
Queue is single waiting line with unlimited space

Queue discipline:
Queue discipline is based upon first come first serve (FCFS).

Service process:
Service rate follows exponential distribution with parameter μ.

Main Components of Queuing System

Arrival

Job arrives to the queuing system at random times. A counting process $A(t)$ tells the number of arrivals that occurred by time t. In stationary queuing system arrivals occur at arrival rate

$$\lambda = \text{Average number of arrivals per unit time} = \frac{EA(t)}{t} \text{ for any } t > 0$$

Queuing and Routing to Servers

Arrived jobs are processed according to the order of their arrivals, on a first come first serve basis. When new job arrives it may find the system in different states. If one server is available at a time it will certainly take a new job. If several servers are available the job may be randomized to one of them or server may be chosen according to some rule.

Service

Once a server becomes available, it immediately starts processing the next assigned job. In practice service time are random because they depend upon amount of work required by each task. The average service time is μ. It varies from one server to other. The service rate is defined as the average number of jobs processed by a continuously working server during one unit of time. $S(t)$ tells the number of customers served by time t

$$\mu = \text{Average number of customers served per unit time} = \frac{ES(t)}{t}, t > 0$$

Departure

When the service is completed, the job leaves the system

Little's law

It is Law given by John Little.

It gives the relationship between the expected number of jobs, the expected response time and the arrival rate. It is valid for any stationary queuing system .It is applied to M/M/1 queuing system and its components, the queue and server. Assuming the system is functional, all the jobs go through the entire system and thus each component is subject to the same arrival rate λ. According to Little law

$$\lambda E(R) = E(X)$$
$$\lambda E(S) = E(X_s)$$
$$\lambda E(W) = E(X_w)$$

Bernoulli Single Server Queuing Process

It is discrete time queuing process with one server, unlimited capacity, arrivals occur according to Bernoulli counting process. The probability of a new arrival during each frame is P_A, during each frame each busy server completes its job Bernoulli counting process with probability P_s independent of other server and independent of the process arrivals. There is geometric number of frames (P_A) between successive arrivals, each service takes a Geometric number of frames (P_S), service of any job takes at least one frame. Arrival $X(t)$ increases by 1 and departure decreases by 1

$$P_A = \lambda\Delta$$
$$P_S = \mu\Delta$$

It is homogeneous Markov chain as P_A and P_S never change. The number of jobs in system increases by 1 with each arrival or decreases by 1 in each departure. It guarantees at most one arrival and one departure during each frame.

In this case;

$P_{00} = P(\text{no arrivals}) = 1 - P_A$

$P_{01} = P(\text{new arrivals}) = P_A$

For all $i \geq 1$

$P_{i,i-1} = P(\text{ no arrivals} \cap \text{one departure}) = (1 - P_A)P_S$

$P_{i,i} = P\{(\text{no arrivals} \cap \text{no departure}) \cup (\text{no arrivals} \cap \text{no departure})\} = (1 - P_A)(1 - P_S) + P_AP_S$

$P_{i,i+1} = P(\text{one arrivals} \cap \text{no departure}) = P_A(1 - P_s)$

Now transition probability matrix is

$$P = \begin{bmatrix} 1 - P_A & P_A & 0 \\ (1 - P_A)P_S & (1 - P_A)(1 - P_S) + P_AP_S & P_A(1 - P_S) \\ 0 & (1 - P_A)P_S & (1 - P_A)(1 - P_S) + P_AP_S \\ 0 & 0 & (1 - P_A)P_S \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Example 13: Laptop computers arrive at a repair shop at the rate of four per day. Assume an 8-hour working day. The expected time to complete service on a laptop is 1.25 hours. Model this process as a single-server Bernoulli queuing process with 15-minute frames. a) Find the service rate. b) Find the arrival and service probabilities.

Solution:

Here, $\lambda = 4$ per day $= 4$ per 8 hour $= ½ = 0.5$ per hour

$\Delta = 15$ minutes $= 15/60 = ¼$ hr

Service time of 1 laptop = 1.25 hrs

1/1.25 laptop = 1 hr

Hence $\mu = 0.8$ per hour

$$P_A = \lambda\,\Delta = 0.5 \times \frac{1}{4} = 0.125$$

$$P_S = \mu\,\Delta = 0.8 \times \frac{1}{4} = 0.2$$

Example 14: A barbershop has one barber and two chairs for waiting. The expected time for a barber to cut customer's hair is 15 minutes. Customers arrive at the rate of two per hour provided the barbershop is not full. However, if the barbershop is full (three customers), potential customers go elsewhere. Assume that the barbershop can be modeled as single-server Bernoulli queuing process with limited capacity. Use frame size of 3 minutes. a) Derive the one-step transition probability matrix for this process. b) Find steady-state probabilities and interpret them.

Solution:

Service time for 1 customer = 15 minutes

Service time for 4 customers = 1 hr

Hence $\mu = 4$ per hour

Arrival of customers = 2 per hour

Hence $\lambda = 2$ per hour

Frame size $\Delta = 3$ minutes $= 3/60 = 1/20$ hr $= 0.05$ hr

Capacity $C = 3$

$P_A = \lambda\,\Delta = 2 \times 0.05 = 0.1$

$P_S = \mu\,\Delta = 4 \times 0.05 = 0.2$

$P_{00} = 1 - P_A = 1 - 0.1 = 0.9$

$P_{01} = P_A = 0.1$

For all $i \geq 1$

$P_{i,i-1} = (1 - P_A)\,P_S = 0.9 \times 0.2 = 0.18$

$P_{i,i} = (1 - P_A)(1 - P_S) + P_A P_S = 0.9 \times 0.8 + 0.1 \times 0.2 = 0.72 + 0.02 = 0.74$

$P_{i,\,i+1} = P_A(1 - P_S) = 0.1 \times 0.8 = 0.08$

Now transition probability matrix is

$$P = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.18 & 0.74 & 0.08 & 0 \\ 0 & 0.18 & 0.74 & 0.08 \\ 0 & 0 & 0.18 & 0.82 \end{bmatrix}$$

For steady state distribution

Now, $\pi p = \pi$

or $$[\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3]\begin{bmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.18 & 0.74 & 0.08 & 0 \\ 0 & 0.18 & 0.74 & 0.08 \\ 0 & 0 & 0.18 & 0.82 \end{bmatrix}$$

$$= [\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3]$$

or $$[0.9\pi_0 + 0.18\pi_1 \quad 0.1\pi_0 + 0.74\pi_1 + 0.08\pi_2 \quad 0.08\pi_1 + 0.74\pi_2 + 0.18\pi_3 \quad 0.08\pi_2 + 0.82\pi_3]$$

$$= [\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3]$$

Hence $0.9\pi_0 + 0.18\pi_1 = \pi_0$ ....(i)

$0.1\pi_0 + 0.74\pi_1 + 0.08\pi_2 = \pi_1$ ....(ii)

$0.08\pi_1 + 0.74\pi_2 + 0.18\pi_3 = \pi_2$ ......(iii)

$0.08\pi_2 + 0.82\pi_3 = \pi_3$ ......(iv)

Also $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$ ......(v)

From (i) $0.18\pi_1 = 0.1\pi_0$

From (ii) $0.18\pi_1 + 0.74\pi_1 + 0.08\pi_2 = \pi_1$

or $0.08\pi_2 = 0.02\pi_1$

or $0.16\pi_2 = 0.08\pi_1$

From (iii) $0.1\pi_2 + 0.74\pi_2 + 0.18\pi_3 = \pi_2$

or $0.18\pi_3 = 0.1\pi_2$

Now from (v) $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$

or $1.8\,\pi_1 + \pi_1 + \pi_2 + \pi_3 = 1$

or $2.8\pi_1 + \pi_2 + \pi_3 = 1$

or $2.8 \times 2\pi_2 + \pi_2 + \pi_3 = 1$

$$6.6\pi_2 + \pi_3 = 1$$

or $$6.6 \times 1.8\pi_3 + \pi_3 = 1$$

or $$12.88\pi_3 = 1$$

or $$\pi_3 = \frac{1}{12.88} = 0.077$$

or $$\pi_2 = 1.8\pi_3 = \frac{1.8}{12.88} = 0.139$$

$$\pi_1 = 2\pi_2 = 2 \times \frac{1.8}{12.88} = \frac{3.6}{12.88} = 0.279$$

$$\pi_0 = 1.8\pi_1 = 1.8 \times \frac{3.6}{12.88} = 0.503$$

According to steady state probabilities 50.3% of time there are no customers in barber shop. 27.9% of time there is no waiting line but barber is working 13.9% of time barber is working and one more customer is waiting and 7.7% of time barber shop is completely full and no vacant seats for waiting.

Example 15: Any printer represents a single server system; the job is sent to a printer at the rate of 10 per hour and takes an average of 50 seconds to print a job. Printer is printing a job and there is another job stored in queue. Assuming single server queuing process with 10 seconds frame. Find out transition probability matrix.

Solution:

Here $\lambda = 10$ per hour $= 1/6$ per minute

$\mu = 1$ per 50 sec $= 6/5$ per minute

$\Delta = 10$ sec $= 1/6$ minute

$P_A = \lambda\Delta = 1/36$

$P_S = \mu\Delta = 1/5$

$P_{00} = 1 - P_A = 1 - 1/36 = 35/36 = 0.972$

$P_{01} = P_A = 1/36 = 0.028$

$P_{i,i-1} = (1 - P_A)P_S = 35/36 \times 1/5 = 7/36 = 0.195$

$P_{i,i} = (1 - P_A)(1 - P_S) + P_AP_S = 35/36 \times 4/5 + 1/36 \times 1/5 = 7/9 + 1/180 = 141/180 = 0.783$

$P_{i,i+1} = P_A(1 - P_S) = 1/36 \times 4/5 = 1/45 = 0.022$

Now transition probability matrix is

$$P = \begin{bmatrix} 0.972 & 0.028 & 0 & 0 & 0 \\ 0.195 & 0.783 & 0.022 & 0 & 0 \\ 0 & 0.195 & 0.783 & 0.022 & 0 \\ 0 & 0 & 0.195 & 0.783 & 0.022 \end{bmatrix}$$

Parameters of queuing system

λ = arrival rate

μ = service rate

$\rho = \frac{\lambda}{\mu}$ = utilization or arrival to service ratio

If $\lambda > \mu$ then queue is infinite, service provider is busy, service system is failure.

If $\lambda = \mu$ then customers come and served, there will be no queue, server will be busy

If $\lambda < \mu$ then there will be no queue, there is idle time for server

Random variables in queuing system

$X_S(t)$ = Number jobs receiving service at time t

$X_W(t)$ = Number of jobs waiting in a queue at time t

$X(t)$ = Total number of jobs in the system

S_k = Service time of k^{th} job

W_k = Waiting time of k^{th} job

R_k = Total time a job spends in the system from its arrival until departure

N = Number of customers in the system (waiting and in service)

L_s = Mean (average) number of customers in the system

L_q = Mean (average) number of customers in the queue

L_b = Mean length of non empty queue

W_S = Mean waiting time in the system

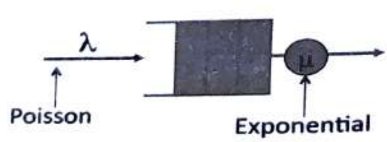
W_q = Mean waiting time in queue

P_w = Probability that an arriving customer has to wait

System with limited capacity

M/M/1 System

It is continuous time Markov process with one server, unlimited capacity, exponential inter arrival times with arrival rate λ, exponential service time with service rate μ, service time and interarrival time are independent. It is limiting case of a Bernoulli queuing process. When frame Δ is small then Δ^2 is negligible.



It is both death process in which only one customer is served at time. Transition are due to arrival or departure of customer. Only nearest neighbour transitions are allowed.

$a_{i,i+1} = \lambda$, $a_{i,i-1} = \mu$

$a_{ij} = 0$ for $|i - j| > 1$.

Transition probabilities for Bernoulli single server queuing process can be obtained as

$$P_{00} = 1 - P_A = 1 - \lambda\Delta$$

$$P_{10} = P_A = \lambda\Delta$$

For $I \geq 1$

$p_{i,i-1} = (1 - P_A) P_S = (1 - \lambda \Delta) \mu \Delta = \mu\Delta - \lambda\mu\Delta^2 = \mu\Delta$

$p_{i,i+1} = P_A (1 - P_S) = \lambda\Delta (1 - M\Delta) = \lambda\Delta - \lambda\mu\Delta^2 = \lambda\Delta$

$p_{i,i} = (1 - P_A)(1 - P_S) + P_A P_S = (1 - \lambda\Delta)(1 - \mu\Delta) + \lambda\Delta.\mu\Delta$

$= 1 - \lambda\Delta - \lambda\Delta + \lambda\mu\Delta^2 + \lambda\mu\Delta^2$

$= 1 - \lambda\Delta - \mu\Delta$

Transition probability matrix is

$$P = \begin{bmatrix} 1-\lambda\Delta & \lambda\Delta & 0 & 0 \\ \mu\Delta & 1-\lambda\Delta-\mu\Delta & \lambda\Delta & 0 \\ 0 & \mu\Delta & 1-\lambda\Delta-\mu\Delta & \lambda\Delta \\ 0 & 0 & \mu\Delta & 1-\lambda\Delta-\mu\Delta \end{bmatrix}$$

Steady state distribution of M/M/1 system

$$\pi_0 = 1 - \rho$$

$$\pi_1 = \rho(1 - \rho)$$

$$\pi_2 = \rho^2 (1 - \rho)$$

and so on.

Evaluating the System Performance

Many important characteristics of the system are as follows;

Utilization rate $= \dfrac{\text{Arrival rate}}{\text{Service rate}} = \dfrac{\lambda}{\mu} = \rho$

Idle rate = 1 – utilization rate $= 1 - \dfrac{\lambda}{\mu} = 1 - \rho$

Probability of no customer in queue

$P_0 = 1 - \dfrac{\lambda}{\mu} = 1 - \rho$

Probability of one customer in queue

$P_1 = \rho P_0 = \rho(1 - \rho)$

Probability of two customers in queue

$P_2 = \rho P_1 = \rho^2(1 - \rho)$

Probability of n customers in queue

$P_n = \rho^n(1 - \rho),\ \rho < 1,\ n = 0, 1, 2, 3, 4, 5 \ldots\ldots\ldots$

Probability of server being busy $= 1 - P_0 = \rho$

Expected (average) number of customers in the system

$L_S = \dfrac{\text{Utilization rate}}{\text{Idle rate}} = \dfrac{\rho}{1 - \rho}$

Expected queue length (Expected number of customers waiting in the queue)

$L_q = L_s$ – Utilization factor

$= L_s - \rho = \dfrac{\rho}{1 - \rho} - \rho = \dfrac{\rho - \rho(1 - \rho)}{1 - \rho} = \dfrac{\rho^2}{1 - \rho}$

Expected (average) waiting time of a customer in the queue

$$W_q = \frac{\text{Average number of customer in queue}}{\text{Arrival rate}} = \frac{L_q}{\lambda}$$

Expected (average) waiting time of a customer in the system

$$W_S = \frac{\text{Average number of customer in system}}{\text{Arrival rate}} = \frac{L_s}{\lambda}$$

Probability of k or more customers in the system

$$P(n \geq k) = \rho^k$$

Variance of queue length

$$V(n) = \frac{\rho}{(1-\rho)^2}$$

Expected number of customers served per busy period

$$L_b = \frac{L_S}{1 - P_0} = \frac{1}{1-\rho}$$

Expected length of non empty queue

$$L'_q = \frac{\text{Expected number of customers in the queue}}{\text{P(more than one customer in queue)}} = \frac{L_q}{P(n > 1)}$$

Example 16: In computer network of a software company arrival of printing message follows Poisson law on an average at every 10 minutes, printer prints messages takes on an average 6 minutes to print following exponential law. Find (i) average arrival rate and average service rate for 1 hour (ii) average arrival rate and average service rate for 15 minutes.

Solution:

i) 1 message arrives in 10 minutes

6 messages arrive in 60 minutes

Hence $\lambda = 6$ per hour (average arrival rate)

1 message serves in 6 minutes

10 messages serve in 60 minutes

Hence $\mu = 10$ per hour (average service rate)

ii) 1 message arrives in 10 minutes

1.5 messages arrive in 15 minutes

Hence $\lambda = 1.5$ per 15 minutes (average arrival rate)

1 message serves in 6 minutes

2.5 messages serve in 15 minutes

Hence $\mu = 2.5$ per 15 minutes (average service rate)

Example 17: A computer repairman finds that the time spent on his jobs has an exponential distribution with mean 20 minutes. If he repairs sets in an order in which they come in and if the arrival of computers follow Poisson distribution with an average rate of 8 per 6 hour. (i) Find repairs idle time each day (ii) Find average number of jobs brought in?

Solution:

Here, service rate of computer is 1 computer per 20 minute

3 computers per 1 hour

Hence $\mu = 3$ per hour

Arrival rate of computer is 8 per 6 hour

4/3 per1 hour

Hence $\lambda = 4/3$ per hour

$$\text{Utility rate } (\rho) = \frac{\lambda}{\mu} = \frac{\frac{4}{3}}{3} = \frac{4}{9}$$

$$\text{Ideal rate} = 1 - \rho = 1 - \frac{4}{9} = \frac{5}{9}$$

$$\text{Ideal time in 6 hours} = 6 \times \text{ideal rate} = 6 \times \frac{5}{9} = \frac{10}{3} \text{ hrs}$$

$$\text{Average length of system } (L_s) = \frac{\rho}{1-\rho} = \frac{4/9}{5/9} = \frac{4}{5}$$

Example 18: Telephone calls arrive at telephone booth following Poisson distribution at an average time of 5 minutes between one and next. The length of phone call is assumed to be exponentially distributed with an average of 4 minutes. (i) What is probability that a person arriving at booth will have to wait? (ii) What is average length of queue that forms from time to time?

Solution:

$\lambda = 1$ call per 5 minute = 1/5 per minute

$\mu = 1$ call per 4 minute = 1/4 per minute

(i) Probability that server is busy $(\rho) = \frac{\lambda}{\mu} = \frac{1/5}{1/4} = \frac{4}{5} = 0.8$

(ii) Average length of queue $(L_q) = \frac{\rho^2}{1-\rho} = \frac{0.64}{1-0.8} = 3.2$

Example 19: In a health clinic, the average rate of arrival of patients is 12 patients per hour. On an average, a doctor can serve patients at the rate of one patient every four minutes. Assume, the arrival of patients follows a Poisson distribution and service to patients follows an exponential distribution. (i) Find the average number of patients in the waiting line and in the clinic (ii) Find the average waiting time in the waiting line or in the queue and (iii) average waiting time in the clinic.

Solution:

Arrival rate of patient, $\lambda = 12$ patients per hour

Service rate of patient, $\mu = 1$ in 4 minutes = 15 patients per hour

Now, $\rho = \frac{\lambda}{\mu} = \frac{12}{15} = 0.8$

Average number of patients in system $L_s = \frac{\rho}{1-\rho} = \frac{0.8}{1-0.8} = 4$ patients

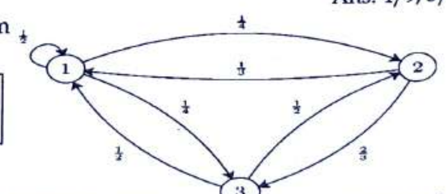
Average number of patients in queue $L_q = \frac{\rho^2}{1-\rho} = \frac{0.64}{1-0.8} = 3.2$ patients

Average waiting time in queue $W_q = \frac{L_q}{\lambda} = \frac{3.2}{12} = 0.26$ hrs

Average waiting time in the system $W_s = \frac{L_s}{\lambda} = \frac{4}{12} = 0.33$ hrs

EXERCISE

1. Define stochastic process and classify.
2. Differentiate between Markov Process and Markov chain.
3. What is steady state distribution?
4. Discuss Binomial process and Poisson process.
5. Describe Bernoulli single server queuing system.
6. What is queuing? Describe M/M/1 queuing system.
7. Define the terms; transition probability, Markov process, Markov chain, n step transition probability.
8. A computer system can operate in two different modes. Every hour it remains in the same mode or switches to a different mode according to the transition probability matrix $\begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$.
   (i) Compute 2 step transition probability matrix
   (ii) If the system is in mode I at 5:30 in a particular time then what is probability that it will be on same mode on the same day at 8:30. **Ans:** $\begin{bmatrix} 0.37 & 0.63 \\ 0.36 & 0.64 \end{bmatrix}$, 0.363
9. An offspring of a tall man is short with probability 0.4 and tall with probability 0.6. An offspring of short man is tall with probability 0.3 and short with probability 0.7.
   (i) Write transition probability matrix of the Markov chain
   (ii) What is probability that grand child of tall man is short? **Ans:** $\begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$, 0.52
10. A Markov chain has transition probability matrix $\begin{bmatrix} 0.4 & \dots & 0 \\ 0 & 0.7 & \dots \\ 0.2 & \dots & 0.6 \end{bmatrix}$
   (i) Fill in the blanks
   (ii) Compute steady state probabilities **Ans:** 0.6, 0.3, 0.2; 0.16, 0.48, 0.36
11. A Markov chain has 3 possible states A, B and C. Every hour it makes a transition to a different state. From state A transition to State B and C are equally likely. From state B transitions to state A only and from state C transition to state A and B are equally likely. Find steady states distribution of states. Ans: 4/9, 3/9, 2/9
12. Obtain transition probability matrix from the state transition diagram.



Ans: $\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/3 & 0 & 2/3 \\ 1/2 & 1/2 & 0 \end{bmatrix}$

13. Tasks are sent to computers at an average rate of 2 tasks per minute. Arrival of tasks module by binomial counting process with $\Delta = 2$ second frames
   (i) Find probability of more than 3 tasks sent during 8 seconds?
   (ii) Find probability of more than 10 tasks sent during 50 seconds.
14. Printing jobs are sent to printer at the average rate of 3 jobs per minute. Using binomial counting process (i) what frame length gives probability 0.2? (ii) Using frame length compute expectation and standard deviation for the number of printing jobs sent to printer during an minute period? **Ans:** $\frac{1}{15}, \frac{1}{3}$, 0.29

15. For a Binomial counting process with 2-second frames and the arrival rate of 10 arrivals per hour, calculate the probability of at least three new arrivals during an interval of 15 minutes. **Ans:** 0.2643

16. Customers of certain internet service provider connect to the internet at the average rate of 10 new connections per minute. Connection are modeled by binomial counting process,

(i) What frame length gives the probability 0.1 of an arrival during given frame?

(ii) What is expectation and standard deviation for the number of seconds between two consecutive connections?

Ans: $\frac{1}{6000}, \frac{1}{600}, 0.0015$

17. Messages arrive at message center according to binomial counting process with average inter arrival time of 10 seconds. Choosing frame size of 2 seconds compute probability that during 100 minutes operation no more than 500 messages arrive. **Ans:** 0

18. Messages arrive at mail server at the average rate of 5 messages every 10 minutes. The arrival of message is modeled by binomial counting process.

(i) What frame length makes the probability of a new message arrival during a given frame 0.15.

(ii) If 20 messages arrive during an hour. Find arrival rate is increased or decreased as compared to frame of (i). **Ans:** 0.3 min, decreased

19. Telephone calls arrive at customer service center according to Poisson process with rate of 2 calls every 5 minutes. Find the probability of getting more than 10 calls during next 15 minutes.

20. The number of baseball games rained out in Mudville is a Poisson process with the arrival rate of 5 per 30 days. a) Find the probability that there are more than 5 rained-out games in 15 days. b) Find the probability that there are no rained-out games in seven days. **Ans:** 0.042, 0.311

21. An internet service provider offers discount on connecting internet. If customers connect internet according to Poisson process with rate of 2 customers per minute. (i) What is probability that no offer is made during first 3 minutes (ii) Find expectation and variance of the time of the first offer. **Ans:** 0.0024, 6, 6

22. Customers arrive at ATM at the rate of 9 customers per hour and spend 3 minutes on average. The system is modeled by Bernoulli single server queuing system with 15 seconds frame. Find the transition probability matrix for the number of customers at the ATM at the end of each frame.

$$\textbf{Ans: } \begin{bmatrix} 0.9625 & 0.0375 & 0 & 0 & 0 & - \\ 0.08 & 0.88511 & 0.034 & 0 & 0 & - \\ 0 & 0.08 & 0.88511 & 0.034 & 0 & - \\ 0 & 0 & 0.08 & 0.8854 & 0.034 & - \\ 0 & 0 & 0 & 0.08 & 0.8854 & 0.0343 \end{bmatrix}$$

23. A printer is a single-server queuing system because it can process only one job at a time while other jobs are stored in a queue. Suppose the jobs are sent to a printer at the rate of 20 per hour, and that it takes an average of 40 seconds to print each job. Currently a printer is printing a job, and there is another job stored in a queue. Assuming Bernoulli single-server queuing process with 20-second frames. Find transition probability matrix.

$$\textbf{Ans: } \begin{bmatrix} 8/9 & 1/9 & 0 & 0 & 0 & - \\ 4/9 & 9/18 & 1/18 & 0 & 0 & - \\ 0 & 4/9 & 9/18 & 1/18 & 0 & - \\ 0 & 0 & 4/9 & 9/18 & 1/18 & - \end{bmatrix}$$

24. Jobs arrive at the server at the rate of 8 jobs per hour. The service takes 3 minutes, on the average. This system is modeled by the single-server Bernoulli queuing process with 5-second frames and capacity limited by 3 jobs. Write the transition probability matrix for the number of jobs in the system at the end of each frame.

$$\textbf{Ans: }\begin{bmatrix} 0.988 & 0.012 & 0 & 0 \\ 0.027 & 0.962 & 0.011 & 0 \\ 0 & 0.027 & 0.962 & 0.011 \\ 0 & 0 & 0.027 & 0.973 \end{bmatrix}$$

25. A customer service with limited capacity C = 2 can have one person getting service and one more person" on hold". When the capacity is reached and someone tries to call, (s)he gets a busy signal or voice mail. Suppose the representative gets an average of 10 calls per hour, and the average phone conversation lasts 4 minutes. Modeling this by a Bernoulli single-server queuing process with limited capacity and 1-minute frames, a) Compute the steady-state distribution and interpret it. **Ans:** 0.438, 0.3512, 0.2106

26. In queuing system with single server average inter arrival time is 6 minutes and average service time is 4 minutes. Find (i) expected response time (ii) fraction of time when the server is busy. **Ans:** 12 min, 2/3

27. For an M/M/1 queuing system with the average inter arrival time of 5 minutes and the average service time of 3 minutes, compute a) the expected time; b) the customer spent in system probability that there are fewer than 2 jobs in the system; c) the fraction of customers who have to wait before their service starts. **Ans:** 7.5 min, 0.64, 0.9

28. Cars arrive at fast food drive through window according to a Poisson process with the average rate of 1 car every 10 minutes. The time each customer spends ordering and getting food is Exponential with the average time of 3 minutes. When a customer is served, the other arrived customers stay in a line waiting for their turn. Compute (a) the expected number of cars in the line at any time. (b) The proportion of time when nobody is served at the drive through window. (c) The expected time it takes to follow the drive through lane, from arrival till departure. **Ans:** 9/70, 0.7, 4.28 min

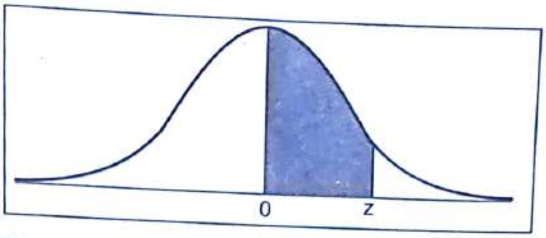
29. Jobs sent to a printer are held in a buffer until they can be printed. Jobs are printed sequentially on a first-come, first-serve basis. Jobs arrive at the printer at the rate of four per minute. The average time to print a job is 10 seconds. Assuming an M/M/1 system, (a) Find the expected value and standard deviation of the number of jobs in this system at any time. (b) When a job is submitted, what is the probability that it will begin printing immediately? **Ans:** 2.44, 0.33

30. In a bank cheques are cashed at a single teller counter. Customers arrive at the counter in a Poisson manner at an average rate of 30 customers per hour. The teller takes on average a minute and half to cash cheque. The service time has been shown to be exponentially distributed. (i) Find the percentage of time teller is busy. (ii) Find average time a customer is expected to spend in system. **Ans:** 75%, 3, 6

31. In a service department managed by a server, on average one customer arrive per 10 minutes. It has been found that each customer requires 6 minutes to be served. Find out (i) Average queue length (ii) Average time spent in the system (iii) Probability that there will be two customers in the queue. **Ans:** 0.9, 15 min, 0.144

❑❑❑

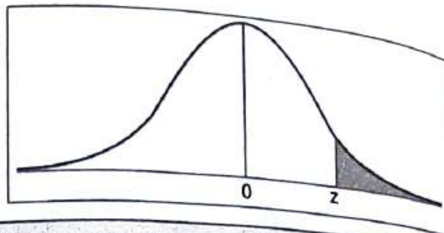
Areas of a Standard Normal Distribution



The table entries represent the area under the standard normal curve from 0 to the specified value of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999

Areas of a Standard Normal Distribution (Right)



The table entries represent the area under the standard normal curve more than specified value of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Critical Values of Student's t-distribution

d.f.	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of significance for two-tailed test					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Critical Values of Chi-Square

The values of χ^2 correspond to a specific right-tail area and specific number of degrees of freedom df.

Degree of Freedom	Level of Significance					
	0.2	0.1	0.05	0.02	0.01	0.001
1	1.642	2.706	3.841	5.412	6.635	10.828
2	3.219	4.605	5.991	7.824	9.210	13.816
3	4.642	6.251	7.815	9.837	11.345	16.266
4	5.989	7.779	9.488	11.668	13.277	18.467
5	7.289	9.236	11.070	13.388	15.086	20.515
6	8.558	10.645	12.592	15.033	16.812	22.458
7	9.803	12.017	14.067	16.622	18.475	24.322
8	11.030	13.362	15.507	18.168	20.090	26.124
9	12.242	14.684	16.919	19.679	21.666	27.877
10	13.442	15.987	18.307	21.161	23.209	29.588
11	14.631	17.275	19.675	22.618	24.725	31.264
12	15.812	18.549	21.026	24.054	26.217	32.909
13	16.985	19.812	22.362	25.472	27.688	34.528
14	18.151	21.064	23.685	26.873	29.141	36.123
15	19.311	22.307	24.996	28.259	30.578	37.697
16	20.465	23.542	26.296	29.633	32.000	39.252
17	21.615	24.769	27.587	30.995	33.409	40.790
18	22.760	25.989	28.869	32.346	34.805	42.312
19	23.900	27.204	30.144	33.687	36.191	43.820
20	25.038	28.412	31.410	35.020	37.566	45.315
21	26.171	29.615	32.671	36.343	38.932	46.797
22	27.301	30.813	33.924	37.659	40.289	48.268
23	28.429	32.007	35.172	38.968	41.638	49.728
24	29.553	33.196	36.415	40.270	42.980	51.179
25	30.675	34.382	37.652	41.566	44.314	52.620
26	31.795	35.563	38.885	42.856	45.642	54.052
27	32.912	36.741	40.113	44.140	46.963	55.476
28	34.027	37.916	41.337	45.419	48.278	56.892
29	35.139	39.087	42.557	46.693	49.588	58.301
30	36.250	40.256	43.773	47.962	50.892	59.703
40	47.269	51.805	55.758	60.436	63.691	73.402
60	68.972	74.397	79.082	84.580	88.379	99.607
120	132.806	140.233	146.567	153.918	158.950	173.617

F Table for $\alpha = 0.10$

Degree of Freedom for Numerator

Rows: Degree of Freedom for Denumerator

df_{2} \ df_{1}	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.863	49.500	53.593	55.832	57.240	58.204	58.905	59.438	59.857	60.194	60.705	61.220	61.740	62.002	62.264	62.529	62.794	63.060	63.328
2	8.526	9.000	9.161	9.243	9.292	9.325	9.349	9.366	9.380	9.391	9.408	9.424	9.441	9.449	9.457	9.466	9.474	9.482	9.491
3	5.538	5.462	5.390	5.342	5.309	5.284	5.266	5.251	5.240	5.230	5.215	5.200	5.184	5.176	5.168	5.159	5.151	5.142	5.133
4	4.544	4.324	4.190	4.107	4.050	4.009	3.978	3.954	3.935	3.919	3.895	3.870	3.844	3.830	3.817	3.803	3.789	3.775	3.760
5	4.060	3.779	3.619	3.520	3.452	3.404	3.367	3.339	3.316	3.297	3.268	3.238	3.206	3.190	3.174	3.157	3.140	3.122	3.105
6	3.775	3.463	3.288	3.180	3.107	3.054	3.014	2.983	2.957	2.936	2.904	2.871	2.836	2.818	2.799	2.781	2.761	2.742	2.722
7	3.589	3.257	3.074	2.960	2.883	2.827	2.784	2.751	2.724	2.702	2.668	2.632	2.594	2.575	2.555	2.535	2.514	2.492	2.470
8	3.457	3.113	2.923	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.501	2.464	2.424	2.404	2.383	2.361	2.339	2.316	2.292
9	3.360	3.006	2.812	2.692	2.610	2.550	2.505	2.469	2.440	2.416	2.378	2.339	2.298	2.276	2.254	2.231	2.208	2.184	2.159
10	3.285	2.924	2.727	2.605	2.521	2.460	2.413	2.377	2.347	2.322	2.284	2.243	2.200	2.178	2.155	2.131	2.107	2.081	2.055
11	3.225	2.859	2.660	2.536	2.451	2.389	2.341	2.304	2.273	2.248	2.208	2.167	2.123	2.100	2.076	2.051	2.026	1.999	1.972
12	3.176	2.806	2.605	2.480	2.394	2.331	2.282	2.244	2.213	2.187	2.147	2.104	2.059	2.035	2.011	1.986	1.959	1.932	1.903
13	3.136	2.763	2.560	2.433	2.346	2.282	2.234	2.195	2.163	2.137	2.096	2.053	2.006	1.982	1.957	1.931	1.904	1.875	1.846
14	3.102	2.726	2.522	2.394	2.306	2.242	2.193	2.153	2.121	2.095	2.053	2.009	1.962	1.937	1.911	1.885	1.857	1.828	1.797
15	3.073	2.695	2.489	2.361	2.273	2.208	2.158	2.118	2.086	2.059	2.017	1.972	1.924	1.899	1.872	1.845	1.816	1.786	1.755
16	3.048	2.668	2.461	2.332	2.243	2.178	2.128	2.087	2.055	2.028	1.985	1.939	1.891	1.865	1.838	1.810	1.781	1.750	1.718
17	3.026	2.644	2.437	2.307	2.218	2.152	2.101	2.061	2.028	2.000	1.957	1.911	1.862	1.836	1.809	1.780	1.750	1.719	1.685
18	3.006	2.623	2.416	2.285	2.195	2.129	2.078	2.037	2.004	1.976	1.933	1.886	1.836	1.810	1.782	1.753	1.723	1.690	1.656
19	2.989	2.605	2.397	2.266	2.175	2.109	2.058	2.017	1.983	1.955	1.911	1.864	1.814	1.787	1.759	1.729	1.698	1.665	1.630
20	2.974	2.589	2.380	2.248	2.158	2.091	2.039	1.998	1.964	1.936	1.892	1.844	1.793	1.766	1.738	1.708	1.676	1.643	1.607
21	2.960	2.574	2.364	2.233	2.142	2.075	2.023	1.981	1.947	1.919	1.874	1.827	1.775	1.748	1.719	1.688	1.656	1.622	1.586
22	2.948	2.561	2.351	2.219	2.127	2.060	2.008	1.966	1.932	1.904	1.859	1.811	1.758	1.731	1.702	1.671	1.638	1.604	1.566
23	2.937	2.549	2.338	2.206	2.114	2.047	1.994	1.953	1.918	1.890	1.844	1.796	1.743	1.715	1.686	1.655	1.622	1.587	1.549
24	2.927	2.538	2.327	2.194	2.103	2.035	1.982	1.940	1.906	1.877	1.831	1.783	1.730	1.701	1.672	1.640	1.607	1.571	1.532
25	2.917	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.894	1.865	1.820	1.770	1.717	1.688	1.658	1.627	1.593	1.557	1.517
26	2.909	2.519	2.307	2.174	2.082	2.013	1.961	1.918	1.884	1.855	1.809	1.759	1.705	1.677	1.646	1.614	1.580	1.543	1.503
27	2.901	2.510	2.298	2.165	2.072	2.004	1.951	1.909	1.874	1.845	1.798	1.749	1.695	1.666	1.635	1.603	1.568	1.531	1.490
28	2.893	2.502	2.290	2.157	2.064	1.995	1.942	1.900	1.865	1.835	1.789	1.739	1.685	1.656	1.625	1.592	1.557	1.519	1.478
29	2.887	2.495	2.283	2.149	2.056	1.987	1.934	1.891	1.856	1.827	1.780	1.730	1.675	1.646	1.615	1.582	1.547	1.508	1.467
30	2.880	2.488	2.276	2.142	2.049	1.980	1.926	1.884	1.848	1.819	1.772	1.722	1.667	1.637	1.606	1.573	1.537	1.498	1.456
40	2.835	2.440	2.226	2.090	1.996	1.926	1.872	1.828	1.792	1.762	1.714	1.662	1.605	1.574	1.541	1.505	1.467	1.424	1.376
60	2.791	2.393	2.177	2.040	1.945	1.874	1.819	1.774	1.738	1.707	1.657	1.603	1.543	1.510	1.475	1.437	1.395	1.347	1.291
120	2.747	2.347	2.129	1.992	1.895	1.823	1.767	1.721	1.684	1.652	1.601	1.545	1.482	1.447	1.409	1.367	1.320	1.264	1.192
∞	2.705	2.302	2.083	1.944	1.847	1.774	1.716	1.670	1.631	1.598	1.545	1.487	1.420	1.383	1.341	1.295	1.239	1.168	1.000

F Table for $\alpha = 0.05$

Degree of Freedom for Numerator

Degree of Freedom for Denumerator

$df_2 \downarrow$ \ $df_1 \rightarrow$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.447	199.500	215.707	224.583	230.161	233.986	236.768	238.882	240.543	241.881	243.906	245.949	248.013	249.051	250.095	251.143	252.195	253.252	254.314
2	18.512	19.000	19.164	19.246	19.296	19.329	19.353	19.371	19.384	19.395	19.412	19.429	19.445	19.454	19.462	19.470	19.479	19.487	19.495
3	10.128	9.552	9.276	9.117	9.013	8.940	8.886	8.845	8.812	8.785	8.744	8.702	8.660	8.638	8.616	8.594	8.572	8.549	8.526
4	7.708	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.998	5.964	5.911	5.857	5.802	5.774	5.745	5.717	5.687	5.658	5.628
5	6.607	5.786	5.409	5.192	5.050	4.950	4.875	4.818	4.772	4.735	4.677	4.618	4.558	4.527	4.495	4.463	4.431	4.398	4.365
6	5.987	5.143	4.757	4.533	4.387	4.283	4.206	4.146	4.099	4.060	3.999	3.938	3.874	3.841	3.808	3.774	3.739	3.704	3.668
7	5.591	4.737	4.346	4.120	3.971	3.866	3.787	3.725	3.676	3.636	3.574	3.510	3.444	3.410	3.375	3.340	3.304	3.267	3.229
8	5.317	4.459	4.066	3.837	3.687	3.580	3.500	3.438	3.388	3.347	3.283	3.218	3.150	3.115	3.079	3.042	3.005	2.966	2.927
9	5.117	4.256	3.862	3.633	3.481	3.373	3.292	3.229	3.178	3.137	3.072	3.006	2.936	2.900	2.863	2.825	2.787	2.747	2.706
10	4.964	4.102	3.708	3.478	3.325	3.217	3.135	3.071	3.020	2.978	2.913	2.845	2.774	2.737	2.699	2.660	2.621	2.580	2.537
11	4.844	3.982	3.587	3.356	3.203	3.094	3.012	2.948	2.896	2.853	2.787	2.718	2.646	2.609	2.570	2.530	2.490	2.448	2.404
12	4.747	3.885	3.490	3.259	3.105	2.996	2.913	2.848	2.796	2.753	2.686	2.616	2.543	2.505	2.466	2.425	2.384	2.341	2.296
13	4.667	3.805	3.410	3.179	3.025	2.915	2.832	2.766	2.714	2.671	2.603	2.533	2.458	2.420	2.380	2.339	2.296	2.252	2.206
14	4.600	3.738	3.343	3.112	2.958	2.847	2.764	2.698	2.645	2.602	2.534	2.463	2.387	2.348	2.308	2.266	2.222	2.177	2.130
15	4.543	3.682	3.287	3.055	2.901	2.790	2.706	2.640	2.587	2.543	2.475	2.403	2.327	2.287	2.246	2.204	2.160	2.114	2.065
16	4.494	3.633	3.238	3.006	2.852	2.741	2.657	2.591	2.537	2.493	2.424	2.352	2.275	2.235	2.193	2.150	2.105	2.058	2.009
17	4.451	3.591	3.196	2.964	2.810	2.698	2.614	2.548	2.494	2.449	2.380	2.307	2.230	2.189	2.147	2.104	2.058	2.010	1.960
18	4.413	3.554	3.159	2.927	2.772	2.661	2.576	2.510	2.456	2.411	2.342	2.268	2.190	2.149	2.107	2.062	2.016	1.968	1.916
19	4.380	3.521	3.127	2.895	2.740	2.628	2.543	2.476	2.422	2.377	2.308	2.234	2.155	2.114	2.071	2.026	1.979	1.930	1.878
20	4.351	3.492	3.098	2.866	2.710	2.599	2.514	2.447	2.392	2.347	2.277	2.203	2.124	2.082	2.039	1.993	1.946	1.896	1.843
21	4.324	3.466	3.072	2.840	2.684	2.572	2.487	2.420	2.366	2.321	2.250	2.175	2.096	2.054	2.010	1.964	1.916	1.865	1.811
22	4.300	3.443	3.049	2.816	2.661	2.549	2.463	2.396	2.341	2.296	2.225	2.150	2.070	2.028	1.984	1.938	1.889	1.838	1.783
23	4.279	3.422	3.028	2.795	2.640	2.527	2.442	2.374	2.320	2.274	2.203	2.128	2.047	2.005	1.960	1.913	1.864	1.812	1.757
24	4.259	3.402	3.008	2.776	2.620	2.508	2.422	2.355	2.300	2.254	2.183	2.107	2.026	1.983	1.939	1.892	1.842	1.789	1.733
25	4.241	3.385	2.991	2.758	2.603	2.490	2.404	2.337	2.282	2.236	2.164	2.088	2.007	1.964	1.919	1.871	1.821	1.768	1.711
26	4.225	3.369	2.975	2.742	2.586	2.474	2.388	2.320	2.265	2.219	2.147	2.071	1.989	1.946	1.901	1.853	1.802	1.748	1.690
27	4.210	3.354	2.960	2.727	2.571	2.459	2.373	2.305	2.250	2.204	2.132	2.055	1.973	1.929	1.884	1.836	1.785	1.730	1.671
28	4.196	3.340	2.946	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.117	2.041	1.958	1.914	1.868	1.820	1.768	1.713	1.654
29	4.183	3.327	2.934	2.701	2.545	2.432	2.346	2.278	2.222	2.176	2.104	2.027	1.944	1.900	1.854	1.805	1.753	1.698	1.637
30	4.170	3.315	2.922	2.689	2.533	2.420	2.334	2.266	2.210	2.164	2.092	2.014	1.931	1.887	1.840	1.791	1.739	1.683	1.622
40	4.084	3.231	2.838	2.606	2.449	2.335	2.249	2.180	2.124	2.077	2.003	1.924	1.838	1.792	1.744	1.692	1.637	1.576	1.508
60	4.001	3.150	2.758	2.525	2.368	2.254	2.166	2.097	2.040	1.992	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	3.920	3.071	2.680	2.447	2.289	2.175	2.086	2.016	1.958	1.910	1.833	1.750	1.658	1.608	1.554	1.495	1.429	1.351	1.253
∞	3.841	2.995	2.604	2.371	2.214	2.098	2.009	1.938	1.879	1.830	1.752	1.666	1.570	1.517	1.459	1.394	1.318	1.221	1.000

F Table for α = 0.01

Degree of Freedom for Numerator

Degree of Freedom for Denumerator

$df_2 \downarrow$ \ $df_1 \rightarrow$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.864
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.020
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1.000

Critical Values of T in the Wilcoxon Matched-Pairs Signed-Ranks Test.

n	Level of significance for one-tailed test			
	0.05	.025	.01	.005
	Level of significance for two-tailed test			
	0.10	.05	.02	.01
5	1	–	–	–
6	2	1	–	–
7	4	2	0	–
8	6	4	2	0
9	8	6	3	2
10	11	8	5	3
11	14	11	7	6
12	17	14	10	7
13	21	17	13	10
14	26	21	16	13
15	30	25	20	16
16	36	30	24	20
17	41	35	28	23
18	47	40	33	28
19	54	46	38	32
20	60	52	43	38
21	68	59	49	43
22	75	66	56	49
23	83	73	62	55
24	92	81	69	61
25	101	90	77	68

Critical values of the Kolmogorov-Smirnov One Sample Test statistics D_n.
This table gives the values of $D^+_{n,\alpha}$ and $D^-_{n,\alpha}$ for which $\alpha \geq P\{D^+_n > D^+_{n,\alpha}\}$ and $\alpha \geq P\{D_n > D_{n,\alpha}\}$ for some selected values of n and α.

One-Sided Test: α =	.10	.05	.025	.01	.005
Two-Sided Test: α =	**.20**	**.10**	**.05**	**.02**	**.01**
n = 1	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.708	.785	.829
4	.493	.565	.524	.689	.734
5	.447	.509	.563	.627	.669
6	.410	.468	.519	.577	.617
7	.381	.436	.483	.538	.576
8	.358	.410	.454	.507	.542
9	.339	.387	.430	.480	.513
10	.323	.369	.409	.457	.489
11	.308	.352	.391	.437	.468
12	.296	.338	.375	.419	.449
13	.285	.325	.361	.404	.432
14	.275	.314	.349	.390	.418
15	.266	.304	.338	.377	.404
16	.258	.295	.327	.366	.392
17	.250	.286	.318	.355	.381
18	.244	.279	.309	.346	.371
19	.237	.271	.301	.337	.361
20	.232	.265	.294	.329	.352
21	.226	.259	.287	.321	.344
22	.221	.253	.281	.314	.337
23	.216	.247	.275	.307	.330
24	.212	.242	.269	.301	.323
25	.208	.238	.264	.295	.317
26	.204	.233	.259	.290	.311
27	.200	.229	.254	.284	.305
28	.197	.225	.250	.279	.300
29	.193	.221	.246	.275	.295
30	.190	.218	.242	.270	.290
31	.187	.214	.238	.266	.285
32	.184	.211	.234	.262	.281
33	.182	.208	.231	.258	.277
34	.179	.205	.227	.254	.273
35	.177	.202	.224	.251	.269
36	.174	.199	.221	.247	.265
37	.172	.196	.218	.244	.262
38	.170	.194	.215	.241	.258
39	.168	.191	.213	.238	.255
40	.165	.189	.210	.235	.252
appro. for n > 40	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Critical values of the Kolmogorov-Smirnov. Test statistics for two samples of equal size

This table gives the values of $D^{+}_{n,\,n,\,\alpha}$ and $D_{n,\,n,\,\alpha}$ for which $\alpha \geq P\{D^{+}_{n,\,n} > D^{+}_{n,\,n,\alpha}\}$ and $\alpha \geq P\{D_{n,n} > D_{n,n,\alpha}\}$ for some selected values of n and α.

One-Sided Test:					
α =	.10	.05	.025	.01	.005
Two-Sided Test:					
α =	.20	.10	.05	.02	.01
n = 3	2/3	2/3			
4	3/4	3/4	3/4		
5	3/5	3/5	4/5	4/5	4/5
6	3/6	4/6	4/6	5/6	5/6
7	4/7	4/7	5/7	5/7	5/7
8	4/8	4/8	5/8	5/8	6/8
9	4/9	5/9	5/9	6/9	6/9
10	4/10	5/10	6/10	6/10	7/10
11	5/11	5/11	6/11	7/11	7/11
12	5/12	5/12	6/12	7/12	7/12
13	5/13	6/13	6/13	7/13	8/13
14	5/14	6/14	7/14	7/14	8/14
15	5/15	6/15	7/15	8/15	8/15
16	6/16	6/16	7/16	8/16	9/16
17	6/17	7/17	7/17	8/17	9/17
18	6/18	7/18	8/18	9/18	9/18
19	6/19	7/19	8/19	9/19	9/19
20	6/20	7/20	8/20	9/20	10/20
21	6/21	7/21	8/21	9/21	10/21
22	7/22	8/22	8/22	10/22	10/22
23	7/23	8/23	9/23	10/23	10/23
24	7/24	8/24	9/24	10/24	11/24
25	7/25	8/25	9/25	10/25	11/25
26	7/26	8/26	9/26	10/26	11/26
27	7/26	8/27	9/27	11/27	11/27
28	8/28	9/28	10/28	11/28	12/28
29	8/29	9/29	10/29	11/29	12/29
30	8/30	9/30	10/30	11/30	12/30
31	8/31	9/31	10/31	11/31	12/31
32	8/32	9/32	10/32	12/32	12/32
34	8/34	10/34	11/34	12/34	13/34
36	9/36	10/36	11/36	12/36	13/36
38	9/38	10/38	11/38	13/38	14/38
40	9/40	10/40	12/40	13/40	14/40
appro. for n > 40	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$	$\frac{2.30}{\sqrt{n}}$

Critical values of the Kolmogorov-Smirnov Test Statistic for Two Samples of Unequal Size.
This table gives the values of $D^{+}_{n_1, n_2, \alpha}$ and $D_{n_1, n_2, \alpha}$ for which $\alpha \geq P[D^{+}_{n_1, n_2} > D^{+}_{n_1, n_2, \alpha}]$ and $\alpha \geq P[D_{n_1, n_2} > D_{n_1, n_2}]$ for some selected values of n_1 = smaller sample size n_2 = larger sample size, and α.

One-Sided Test:						
	α =	.10	.05	.025	.01	.005
Two-Sided Test:						
	α =	.20	.10	.05	.02	.01
$n_1 = 1$	$n_2 = 9$	17/18				
	10	9/10				
$n_1 = 2$	$n_2 = 3$	5/6				
	4	3/4				
	5	4/5	4/5			
	6	5/6	5/6			
	7	5/7	6/7			
	8	3/4	7/8	7/8		
	9	7/9	8/9	8/9		
	10	7/10	4/5	9/10		
$n_1 = 3$	$n_2 = 4$	3/4	3/4			
	5	2/3	4/5	4/5		
	6	2/3	2/3	5/6		
	7	2/3	5/7	6/7	6/7	
	8	5/8	3/4	3/4	7/8	
	9	2/3	2/3	7/9	8/9	8/9
	10	3/5	7/10	4/5	9/10	9/10
	12	7/12	2/3	3/4	5/6	11/12
$n_1 = 4$	$n_2 = 5$	3/5	3/4	4/5	4/5	
	6	7/12	2/3	3/4	5/6	5/6
	7	17/28	5/7	3/4	6/7	6/7
	8	5/8	5/8	3/4	7/8	7/8
	9	5/9	2/3	3/4	7/9	8/9
	10	11/20	13/20	7/10	4/5	4/5
	12	7/12	2/3	2/3	3/4	5/6
	16	9/16	5/8	11/16	3/4	13/16
$n_1 = 5$	$n_2 = 6$	3/5	2/3	2/3	5/6	5/6
	7	4/7	23/35	5/7	29/35	6/7
	8	11/20	5/8	27/40	4/5	4/5
	9	5/9	3/5	31/45	7/9	4/5
	10	1/2	3/5	7/10	7/10	4/5
	15	8/15	3/5	2/3	11/15	11/15
	20	1/2	11/20	3/5	7/10	3/4

$n_1 = 6$	$n_2 = 7$	23/42	4/7	29/42	5/7	5/6
	8	1/2	7/12	2/3	3/4	3/4
	9	1/2	5/9	2/3	13/18	7/9
	10	1/2	17/30	19/30	7/10	11/15
	12	1/2	7/12	7/12	2/3	3/4
	18	4/9	5/9	11/18	2/3	13/18
	24	11/24	1/2	7/12	5/8	2/3
$n_1 = 7$	$n_2 = 8$	27/56	33/56	5/8	41/56	3/4
	9	31/63	5/9	40/63	5/7	47/63
	10	33/70	39/70	43/70	7/10	5/7
	14	3/7	1/2	4/7	9/14	5/7
	28	3/7	13/28	15/28	17/28	9/14
$n_1 = 8$	$n_2 = 9$	4/9	13/24	5/8	2/3	¾
	10	19/40	21/40	23/40	27/40	7/10
	12	11/24	½	7/12	5/8	2/3
	16	7/16	½	9/16	5/8	5/8
	32	13/32	7/16	1/2	9/16	19/32
$n_1 = 9$	$n_2 = 10$	7/15	½	26/45	2/3	31/45
	12	4/9	½	5/9	11/18	2/3
	15	19/45	22/45	8/15	3/5	29/45
	18	7/18	4/9	½	5/9	11/18
	36	13/36	5/12	17/36	19/36	5/9
$n_1 = 10$	$n_2 = 15$	2/5	7/15	½	17/30	19/30
	20	2/5	9/20	½	11/20	3/5
	40	7/20	2/5	9/20	1/2	
$n_1 = 12$	$n_2 = 15$	23/60	9/20	½	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	11/36	5/12	17/36	19/36	5/9
	20	13/30	5/12	7/15	31/60	17/30
$n_1 = 15$	$n_2 = 20$	7/20	2/5	13/30	29/60	31/60
$n_1 = 16$	$n_2 = 20$	27/80	31/80	17/40	19/40	41/80
Large sample approx.		$1.07\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1.22\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1.36\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1.52\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1.63\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

Probabilities associated with values as large as observed values of Friedman Statistic F_r i.e. $p_0 = P(F_r > F_r^*)$ where F_r^* is the calculated value of F_r.

Table for k = 3

n = 2		n = 3		n = 4		n = 5	
F_r	p	F_r	p	F_r	p	F_r	p
0	1.000	.000	1.000	.0	1.000	.0	1.000
1	.833	.667	.994	.5	.931	.4	.954
3	.500	2.000	.528	1.5	.653	1.2	.691
4	.167	2.667	.361	2.0	.431	1.6	.522
		4.667	.194	3.5	.273	2.8	.367
		6.000	.028	4.5	.125	3.6	.182
				6.0	.069	4.8	.124
				6.5	.042	5.2	.093
				8.0	.0046	6.4	.039
						7.6	.024
						8.4	.0085
						10.0	.00077

n = 6		n = 7		n = 8		n = 9	
F_r	p	F_r	p	F_r	p	F_r	p
.00	1.000	.000	1.000	.00	1.000	.000	1.000
.33	.956	.286	.964	.25	.967	.222	.971
1.00	.740	.857	.768	.75	.794	.667	.814
1.33	.570	1.143	.620	1.00	.654	.889	.865
2.33	.430	2.000	.486	1.75	.531	1.556	.569
3.00	.252	2.571	.305	2.25	.355	2.000	.398
4.20	.184	3.429	.237	3.00	.285	2.667	.328
4.33	.142	3.714	.192	3.25	.236	2.889	.278
5.33	.072	4.571	.112	4.00	.149	3.556	.187
6.33	.052	5.429	.085	4.75	.120	4.222	.154
7.00	.029	6.000	.052	5.25	.079	4.667	.107
8.33	.012	7.143	.027	6.25	.047	5.556	.069
9.00	.0081	7.714	.021	6.75	.038	6.000	.057
9.33	.0055	8.000	.016	7.00	.030	6.222	.048
10.33	.0017	8.857	.0084	7.75	.018	6.889	.031
12.00	.00013	10.286	.0036	9.00	.0099	8.000	.019
		10.571	.0027	9.25	.0080	8.222	.016
		11.143	.0012	9.75	.0048	8.667	.010
		12.286	.00032	10.75	.0024	9.556	.0060
		14.000	.000021	12.00	.0011	10.667	.0035
				12.25	.00086	10.889	.0029
				13.00	.00026	11.556	.0013
				14.25	.000061	12.667	.00066
				16.00	.0000036	13.556	.00035
						14.000	.00020
						14.222	.000097
						14.889	.000054
						16.222	.000011
						18.000	.0000006

Table for k = 4

n = 2		n = 3		n = 4		n = 5	
F_r	p	F_r	p	F_r	p	F_r	p
.0	1.000	.2	1.000	.0	1.000	5.7	.141
.6	.958	.6	.958	.3	.992	6.0	.105
1.2	.834	1.0	.910	.6	.928	6.3	.094
1.8	.792	1.8	.727	.9	.900	6.6	.077
2.4	.625	2.2	.608	1.2	.800	6.9	.068
3.0	.542	2.6	.524	1.5	.754	7.2	.054
3.6	.458	3.4	.446	1.8	.677	7.5	.052
4.2	.375	3.8	.342	2.1	.649	7.8	.036
4.8	.208	4.2	.300	2.4	.524	8.1	.033
5.4	.167	5.0	.207	2.7	.508	8.4	.019
6.0	.042	5.4	.175	3.0	.432	8.7	.014
		5.8	.148	3.3	.389	9.3	.012
		6.6	.075	3.6	.355	9.6	.0069
		7.0	.054	3.9	.324	9.9	.0062
		7.4	.033	4.5	.242	10.2	.0027
		8.2	.017	4.8	.200	10.8	.0016
		9.0	.0017	5.1	.190	11.1	.00094
				5.4	.158	12.0	.000072

Critical values for total number of runs 'r' at $\alpha = 0.05$ for two tailed test.

The smaller critical value for a left-hand critical region, the larger for a right-hand critical region. For a one tailed test $\alpha = 0.025$ and use only-one of the critical values of r.

The Smaller of n_1 and n_2	The larger of n_1 and n_2															
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2								2, 6	2, 6	2, 6	2, 6	2, 6	2, 6	2, 6	2, 6	2, 6
3		2, 8	2, 8	2, 8	2, 8	2, 8	2, 8	2, 8	2, 8	2, 8	3, 8	3, 8	3, 8	3, 8	3, 8	3, 8
4	2, 9	2, 9	2, 10	3, 10	3, 10	3, 10	3, 10	3, 10	3, 10	3, 10	3, 10	4, 10	4, 10	4, 10	4, 10	4, 10
5	2, 10	3, 10	3, 11	3, 11	3, 12	3, 12	4, 12	4, 12	4, 12	4, 12	4, 12	4, 12	4, 12	5, 12	5, 12	5, 12
6		3, 11	3, 12	3, 12	4, 13	4, 13	4, 13	4, 13	5, 14	5, 14	5, 14	5, 14	5, 14	5, 14	6, 14	6, 14
7			3, 13	4, 13	4, 14	5, 14	5, 14	5, 14	5, 15	5, 15	6, 15	6, 16	6, 16	6, 16	6, 16	6, 16
8				4, 14	5, 14	5, 15	5, 15	6, 16	6, 16	6, 16	6, 16	6, 17	7, 17	7, 17	7, 17	7, 17
9					5, 15	5, 16	6, 16	6, 16	6, 17	7, 17	7, 18	7, 18	7, 18	7, 18	7, 18	7, 18
10						6, 16	6, 17	7, 17	7, 18	7, 18	7, 18	8, 19	8, 19	8, 19	8, 20	9, 20
11							7, 17	7, 18	7, 19	8, 19	8, 19	8, 20	9, 20	9, 20	9, 21	9, 21
12								7, 19	8, 19	8, 20	8, 20	9, 21	9, 21	9, 21	10, 22	10, 22
13									8, 20	9, 20	9, 21	9, 21	10, 22	10, 22	10, 23	10, 23
14										9, 21	9, 22	10, 22	10, 23	10, 23	11, 23	11, 24
15											10, 22	10, 23	11, 23	11, 24	11, 24	12, 25
16												11, 23	11, 24	11, 25	12, 25	12, 25
17													11, 25	12, 25	12, 26	13, 26
18														12, 26	13, 26	13, 27
19															13, 27	13, 27
20																14, 28

Distribution function of U i.e. $P(U \leq U_0) = p_0$, $n_1 \leq n_2$ and $3 \leq n_2 \leq 10$

The probabilities associated with the values as small as observed value of U in the Mann-Whitney test.

$n_2 = 3$

	n_1		
U_0	1	2	3
0	.25	.10	.05
1	.50	.20	.10
2		.40	.20
3		.60	.35
4			.50

$n_2 = 4$

	n_1			
U_0	1	2	3	4
0	.2000	.0667	.0286	.0143
1	.4000	.1333	.0571	.0286
2	.6000	.2667	.1143	.0571
3		.4000	.2000	.1000
4		.6000	.3143	.1714
5			.4286	.2429
6			.5714	.3429
7				.4429
8				.5571

$n_2 = 5$

	n_1				
U_0	1	2	3	4	5
0	.1667	.0476	.0179	.0079	.0040
1	.3333	.0952	.0357	.0159	.0079
2	.5000	.1905	.0714	.0317	.0159
3		.2857	.1250	.0556	.0278
4		.4286	.1964	.0952	.0476
5		.5714	.2857	.1429	.0754
6			.3929	.2063	.1111
7			.5000	.2778	.1548
8				.3651	.2103
9				.4524	.2738
10				.5476	.3452
11					.4206
12					.5000

$n_2 = 6$

	n_1					
U_0	1	2	3	4	5	6
0	.1429	.0357	.0119	.0048	.0022	.0011
1	.2857	.0714	.0238	.0095	.0043	.0022
2	.4286	.1429	.0476	.0190	.0087	.0043
3	.5714	.2143	.0833	.0333	.0152	.0076
4		.3214	.1310	.0571	.0260	.0130
5		.4286	.1905	.0857	.0411	.0206
6		.5714	.2738	.1286	.0628	.0325
7			.3571	.1762	.0887	.0465
8			.4524	.2381	.1234	.0660
9			.5476	.3048	.1645	.0898
10				.3810	.2143	.1201
11				.4571	.2684	.1548
12				.5429	.3312	.1970
13					.3961	.2424
14					.4654	.2944
15					.5346	.3496
16						.4091
17						.4686
18						.5314

$n_2 = 7$

	n_1						
U_0	1	2	3	4	5	6	7
0	.1250	.0278	.0083	.0030	.0013	.0006	.0003
1	.2500	.0556	.0167	.0061	.0025	.0012	.0006
2	.3750	.1111	.0333	.0121	.0051	.0023	.0012
3	.5000	.1667	.0583	.0212	.0088	.0041	.0020
4		.2500	.0917	.0364	.0152	.0070	.0035
5		.3333	.1333	.0545	.0240	.0111	.0055
6		.4444	.1917	.0818	.0366	.0175	.0087
7		.5556	.2583	.1152	.0530	.0256	.0131
8			.3333	.1576	.0745	.0367	.0189
9			.4167	.2061	.1010	.0507	.0265
10			.5000	.2636	.1338	.0688	.0364
11				.3242	.1717	.0903	.0487
12				.3939	.2159	.1171	.0641
13				.4636	.2652	.1474	.0825
14				.5364	.3194	.1830	.1043
15					.3775	.2226	.1297
16					.4381	.2669	.1588
17					.5000	.3141	.1914
18						.3654	.2279
19						.4178	.2675
20						.4726	.3100
21						.5274	.3552
22							.4024
23							.4508
24							.5000

$n_2 = 8$

	n_1							
U_0	1	2	3	4	5	6	7	8
0	.1111	.0222	.0061	.0020	.0008	.0003	.0002	.0001
1.	.2222	.0444	.0121	.0040	.0016	.0007	.0003	.0002
2.	.3333	.0889	.0242	.0081	.0031	.0013	.0006	.0003
3.	.4444	.1333	.0424	.0141	.0054	.0023	.0011	.0005
4.	.5556	.2000	.0667	.0242	.0093	.0040	.0019	.0009
5.		.2667	.0970	.0364	.0148	.0063	.0030	.0015
6.		.3556	.1394	.0545	.0225	.0100	.0047	.0023
7.		.4444	.1879	.0768	.0326	.0147	.0070	.0035
8.		.5556	.2485	.1071	.0466	.0213	.0103	.0052
9.			.3152	.1414	.0637	.0296	.0145	.0074
10.			.3879	.1838	.0855	.0406	.0200	.0103
11.			.4606	.2303	.1111	.0539	.0270	.0141
12.			.5394	.2848	.1422	.0709	.0361	.0190
13.				.3414	.1772	.0906	.0469	.0249
14.				.4040	.2176	.1142	.0603	.0325
15.				.4667	.2618	.1412	.0760	.0415
16.				.5333	.3108	.1725	.0946	.0524
17.					.3621	.2068	.1159	.0652
18.					.4165	.2454	.1405	.0803
19.					.4716	.2864	.1678	.0974
20.					.5284	.3310	.1984	.1172
21.						.3773	.2317	.1393
22.						.4259	.2679	.1641
23.						.4749	.3063	.1911
24.						.5251	.3472	.2209
25.							.3894	.2527
26.							.4333	.2869
27.							.4775	.3227
28.							.5225	.3605
29.								.3992
30.								.4392
31.								.4796
32.								.5204

$n_2 = 9$

	n_1								
U_0	1	2	3	4	5	6	7	8	9
0	.1000	.0182	.0045	.0014	.0005	.0002	.0001	.0000	.0000
1.	.2000	.0364	.0091	.0028	.0010	.0004	.0002	.0001	.0000
2.	.3000	.0727	.0182	.0056	.0020	.0008	.0003	.0002	.0001
3.	.4000	.1091	.0318	.0098	.0035	.0014	.0006	.0003	.0001
4.	.5000	.1636	.0500	.0168	.0060	.0024	.0010	.0005	.0002
5.		.2182	.0727	.0252	.0095	.0038	.0017	.0008	.0004
6.		.2909	.1045	.0378	.0145	.0060	.0026	.0012	.0006
7.		.3636	.1409	.0531	.0210	.0088	.0039	.0019	.0009
8.		.4545	.1864	.0741	.0300	.0128	.0058	.0028	.0014
9.		.5455	.2409	.0993	.0415	.0180	.0082	.0039	.0020

10.			.3000	.1301	.0559	.0248	.0115	.0056	.0028
11.			.3636	.1650	.0734	.0332	.0156	.0076	.0039
12.			.4318	.2070	.0949	.0440	.0209	.0103	.0053
13.			.5000	.2517	.1199	.0567	.0274	.0137	.0071
14.				.3021	.1489	.0723	.0356	.0180	.0094
15.				.3552	.1818	.0905	.0454	.0232	.0122
16.				.4126	.2188	.1119	.0571	.0296	.0157
17.				.4699	.2592	.1361	.0708	.0372	.0200
18.				.5301	.3032	.1638	.0869	.0464	.0252
19.					.3497	.1942	.1052	.0570	.0313
20.					.3986	.2280	.1261	.0694	.0385
21.					.4491	.2643	.1496	.0836	.0470
22.					.5000	.3035	.1755	.0998	.0567
23.						.3445	.2039	.1179	.0680
24.						.3878	.2349	.1383	.0807
25.						.4320	.2680	.1606	.0951
26.						.4773	.3032	.1852	.1112
27.						.5227	.3403	.2117	.1290
28.							.3788	.2404	.1487
29.							.4185	.2707	.1701
30.							.4591	.3029	.1933
31.							.5000	.3365	.2181
32.								.3715	.2447
33.								.4074	.2729
34.								.4442	.3024
35.								.4813	.3332
36.								.5187	.3652
37.									.3981
38.									.4317
39.									.4657
40.									.5000

$n_2 = 10$

	n_1									
U_0	1	2	3	4	5	6	7	8	9	10
0	.0909	.0152	.0035	.0010	.0003	.0001	.0001	.0000	.0000	.0000
1.	.1818	.0303	.0070	.0020	.0007	.0002	.0001	.0000	.0000	.0000
2.	.2727	.0606	.0140	.0040	.0013	.0005	.0002	.0001	.0000	.0000
3.	.3636	.0909	.0245	.0070	.0023	.0009	.0004	.0002	.0001	.0000
4.	.4545	.1364	.0385	.0120	.0040	.0015	.0006	.0003	.0001	.0001
5.	.5455	.1818	.0559	.0180	.0063	.0024	.0010	.0004	.0002	.0001
6.		.2424	.0804	.0270	.0097	.0037	.0015	.0007	.0003	.0002
7.		.3030	.1084	.0380	.0140	.0055	.0023	.0010	.0005	.0002
8.		.3788	.1434	.0529	.0200	.0080	.0034	.0015	.0007	.0004
9.		.4545	.1853	.0709	.0276	.0112	.0048	.0022	.0011	.0005
10.		.5455	.2343	.0939	.0376	.0156	.0068	.0031	.0015	.0008
11.			.2867	.1199	.0496	.0210	.0093	.0043	.0021	.0010

12.			.3462	.1518	.0646	.0280	.0125	.0058	.0028	.0014
13.			.4056	.1868	.0823	.0363	.0165	.0078	.0038	.0019
14.			.4685	.2268	.1032	.0467	.0215	.0103	.0051	.0026
15.			.5315	.2697	.1272	.0589	.0277	.0133	0.066	.0034
16.				.3177	.1548	.0736	.0351	.0171	.0086	.0045
17.				.3666	.1855	.0903	.0439	.0217	.0110	.0057
18.				.4196	.2198	.1099	.0544	.0273	.0140	.0073
19.				.4725	.2567	.1317	.0665	.0338	.0175	.0093
20.				.5275	.2970	.1566	.0806	.0416	.0217	.0116
21.					.3393	.1838	.0966	.0506	.0267	.0144
22.					.3839	.2139	.1148	.0610	.0326	.0177
23.					.4296	.2461	.1349	.0729	.0394	.0216
24.					.4765	.2811	.1574	.0864	.0474	.0262
25.					.5235	.3177	.1819	.1015	.0564	.0315
26.						.3564	.2087	.1185	.0667	.0376
27.						.3962	.2374	.1371	.0782	.0446
28.						.4374	.2681	.1577	.0912	.0526
29.						.4789	.3004	.1800	.1055	.0615
30.						.5211	.3345	.2041	.1214	.0716
31.							.3698	.2299	.1388	.0827
32.							.4063	.2574	.1577	.0952
33.							.4434	.2863	.1781	.1088
34.							.4811	.3167	.2001	.1237
35.							.5189	.3482	.2235	.1399
36.								.3809	.2483	.1575
37.								.4143	.2745	.1763
38.								.4484	.3019	.1965
39.								.4827	.3304	.2179
40.								.5173	.3598	.2406
41.									.3901	.2644
42.									.4211	.2894
43.									.4524	.3153
44.									.4841	.3421
45.									.5159	.3697
46.										.3980
47.										.4267
48.										.4559
49.										.4853
50.										.5147

Critical value of U in the Mann-Whitney Test

a. Critical values of U for a one-tailed test at 0.025 or for a two-tailed test at 0.05

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1								0	0	0	0	1	1	1	1	1	2	2	2	2
2					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
3				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
4			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
5			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
6			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
7		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
8		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
9		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
10		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
11																				
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	[illegible]	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	[illegible]	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	103	112	119	127

b. Critical values of U for a one-tailed test at 0.05 or for a two-tailed test at 0.10

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																			0	0
2					0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
3			0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
4			0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5		0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6		0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7		0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
8		1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
9		1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
10		1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
11		1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
12		2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
13		2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
14		2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
15		3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
16		3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
17		3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
18		4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

Probabilities associated with values as large as observed values of Kruskul-Wallis H Statistic. i.e. $p_0 = P(H > H^*)$ where $H^* = H_{cal.}$

Sample sizes			H	p
n_1	n_2	n_3		
2	1	1	2.7000	.500
2	2	2	3.6000	.200
2	2	2	4.5714	.067
			3.7143	.200
3	1	1	3.2000	.300
3	2	1	4.2857	.100
			3.8571	.100
3	2	2	5.3572	.029
			4.7143	.048
			4.5000	.067
			4.4643	.105
3	3	1	5.1429	.043
			4.5714	.100
			4.0000	.129
3	3	2	6.2500	.011
			5.3611	.032
			5.1389	.061
			4.5556	.100
			4.2500	.121
3	3	3	7.2000	.004
			6.4889	.011
			5.6889	.029
			5.6000	.050
			5.0667	.086
			4.6222	.100
4	1	1	3.5714	.200
4	2	1	4.8214	.057
			4.5000	.076
			4.0179	.114
4	2	2	6.0000	.014
			5.3333	.033
			5.1250	.052
			4.4583	.100
			4.1667	.105
4	3	1	5.8333	.021
			5.2083	.050
			5.0000	.057
			4.0556	.093
4	3	2	6.4444	.008
			6.3000	.011
			5.4444	.046
			5.4000	.051
			4.5111	.098
			4.4444	.102
4	3	3	6.7455	.010
			6.7455	.010
			6.7091	.013
			5.7909	.046
			4.7091	.092
			4.7000	.101
4	4	1	6.6667	.010
			6.1667	.022
			4.9667	.048
			4.8667	.054
			4.1667	.082
			4.0667	.102
4	4	2	7.0364	.006
			6.8727	.011
			5.4545	.046
			5.2364	.052
			4.5545	.098
			4.4455	.103
4	4	3	7.1439	.010
			7.1364	.011
			5.5985	.049
			5.5758	.051
			4.5455	.099
			4.4773	.102
4	4	4	7.6538	.008
			7.5385	.011
			5.6923	.049
			5.6538	.054
			4.6539	.097
			4.5001	.104
5	1	1	3.8571	.143
5	2	1	5.2500	.036
			5.0000	.048
			4.4500	.071
			4.2000	.095

			3.8889	.129
5	2	2	6.5333	.008
			6.1333	.013
			5.1600	.034
			5.0400	.056
			4.3733	.090
			4.2933	.122
5	3	1	6.4000	.012
			4.9600	.048
			4.8711	.052
			4.0178	.095
			3.8400	.123
5	3	2	6.9091	.009
			6.8281	.010
			5.2509	.049
			5.1055	.052
			4.6509	.091
			4.4945	.101
5	3	3	7.0788	.009
			6.9818	.011
			5.6485	.049
			5.5152	.051
			4.5333	.097
			4.4121	.109
5	4	1	6.9545	.008
			6.8400	.011
			4.9855	.044
			4.8600	.056
			3.9873	.098
			3.9600	.102
5	4	2	7.2045	.009
			7.1182	.010
			5.2727	.049
			5.2682	.050
			4.5409	.098
			4.5182	.101
5	4	3	7.4449	.010
			7.3949	.011
			5.6564	.049
			5.6308	.050
			4.5487	.099
			4.5231	.103
			4.0500	.119
5	4	4	7.7604	.009
			7.7440	.011
			5.6571	.049
			5.6176	.050
			4.6187	.100
			4.5527	.102
5	5	1	7.3091	.009
			6.8364	.011
			5.1273	.046
			4.9091	.53
			4.1091	.086
			4.0364	.105
5	5	2	7.3385	.010
			7.2692	.010
			5.3385	.047
			5.2464	.051
			4.6231	.097
			4.5077	.100
5	5	3	7.5780	.010
			7.5429	.010
			5.7055	.046
			5.6264	.051
			4.5451	.100
			4.5363	.102
5	5	4	7.8229	.010
			7.7914	.010
			5.6657	.049
			5.6429	.050
			4.5229	.099
			4.5200	.101
5	5	5	8.0000	.009
			7.9800	.010
			5.7800	.049
			5.6600	.051
			4.5600	.100
			4.5000	.102

Probabilities associated with values as small as observed values of x in the binomial distribution with parameter n and p = ½

$$\text{i.e. } p_0 = P(X \leq x) = \sum_{x=0}^{x=k} \binom{n}{x} \left(\frac{1}{2}\right)^n ; k = 0, 1, 2, \ldots (n-1).$$

n \ x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	5	*														
2	250	500	*													
3	125	500	875	*												
4	063	313	688	938	*											
5	031	188	500	812	969	*										
6	016	109	344	656	891	984	*									
7	008	062	227	500	773	938	992	*								
8	004	035	145	363	637	855	965	996	*							
9	002	020	090	254	500	746	910	980	998	*						
10	001	011	055	172	377	623	828	945	989	999	*					
11		006	033	113	274	500	726	887	967	994	*	*				
12		003	019	073	194	387	613	806	927	981	997	*	*			
13		002	011	046	133	291	500	709	867	954	989	998	*	*		
14		001	006	029	090	212	395	605	788	910	971	994	999	*	*	
15			004	018	059	151	304	500	696	849	941	982	996	*	*	*
16			002	011	038	105	227	402	598	773	895	962	989	998	*	*
17			001	006	025	072	166	315	500	685	834	928	975	994	999	*
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

Note: To save space decimal points are omitted in the p's.

** = 1 or approximately 1.0*

Tribhuvan University

Institute of Science and Technology

Model Questions

Bachelor Level/Second Year/Third Semester/Science — Full Marks 60

Computer Science and Information Technology (STA 210) — Pass Marks: 24

(Statistics II) — Time: 3 hours

Candidates are required to give their answers in their own words as for as practicable.

All notations have the usual meanings

Group 'A'

Attempt any Two questions **(2 × 10 = 20)**

1. Suppose a population of 4 computers with their lifetimes 3, 5, 7 & 9 years. Comment on the population distribution. Assuming that you sample with replacement, select all possible samples of n = 2, construct sampling distribution of mean and compare the population distribution and sampling distribution of mean. Compare population mean versus mean of all sample means, and population variance versus variance of sample means and comment on them with the support of theoretical consideration if any.

2. A computer manager is keenly interested to know how efficiency of her new computer program depends on the size of incoming data and data structure. Efficiency will be measured by the number of processed requests per hour. Data structure may be measured on how many tables were used to arrange each data set. All the information was put together as follows.

Data size (Giga bytes)	6	7	7	8	10	10	15
Number of tables	4	20	20	10	10	2	1
Processed requests	40	55	50	41	17	26	16

Identify which one is dependent variable? Fit the appropriate multiple regression model and provide problem specific interpretations of the fitted regression coefficients.

3. State and explain the mathematical model for randomized complete block design. Explain all the steps to be adopted to carry out the analysis and finally prepare the ANOVA table.

Group 'B'

Attempt any Eight questions **(8 × 5 = 40)**

4. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7, with a sample standard deviation of 9.2. At the 1s% level of significance, do these data provide considerable evidence that the mean number of concurrent users is greater than 35? Draw your conclusion based on your result.

5. A sample of 250 items from lot A contains 10 defective items, and a sample of 300 items from lot B is found to contain 18 defective items.

At a significance level $\alpha = 0.05$, is there a significance difference between the quality of the two lots?

6. Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam, and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains images or not. The following were collected on $n = 1000$ random email message.

Spam Status	Image Containing status		
Spam	160	240	400
No spam	140	460	600
Total	300	700	1000

Assess whether being spam and containing images are independent factors at 1% level of significance.

7. Two computer marker, A and B, complete for a certain market. Their users rank the quality of computers on a 4-points scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality, will recommend to other. The following counts were observed:

Computer maker	Not satisfied	Satisfied	Good quality	Excellent quality
A	20	40	70	20
B	10	30	40	20

Is there a significant difference in customer satisfaction of the computers produced by A and by B using Mann-Whitney U test at 5% level of significance.

8. Defined queuing systems with suitable examples. Also explain the main components of queuing system in brief.

9. In some town, each day is either sunny day is followed by another sunny day with probability 0.7, whereas a rainy day is followed by a sunny day with probability 0.4. Weather conditions in this problem represents a homogeneous Markov chain with 2 states; state I = "Sunny" and state 2 = "rainy." Transition probability matrix of sunny and rainy days is given below.

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.5 \end{pmatrix}$$

Compute the probability of sunny days and rainy days using the steady-state equation for this Markov Chain.

10. Consider a completely randomized design with 4 treatments with 7 observations in each. For the ANOVA summary table below, fill in the missing results. Also indicate your statistical decision.

Source	Degrees of freedom	Sum of squares	Mean sum of squares	F-ratio
Treatment?	?	SSA = ?	70	F
Error	?	SSE = 590	40	20
Total	?	SST = ?		

11. Following are the score obtained by 10 university staffs on the computer proficiency skills before training and after training. It was assumed that the proficiency of computer skills is expected to be increased after training.

Staff	**Score**	
1	Before training	After training
2	50	55
3	30	40
4	15	30
5	22	30
6	34	36
7	45	45
8	40	41
9	10	30
10	26	40

Test at 5% level of significance whether the training is effective to improve the computer proficiency skills applying appropriate statistical test. Assume that the given score follows normal distribution.

12. Write short notes on the follows:

a. Concept of Latin Square Design

b. Multiple correlation

Tribhuvan University

Institute of Science and Technology

2075

Bachelor Level/Second Year/Third Semester/Science
Computer Science and Information Technology (STA 210)
(Statistics II)
(New Course)

Full Marks 60
Pass Marks: 24
Time: 3 hours

Candidates are required to give their answers in their own words as for as practicable.

All notations have the usual meanings

The figures in the margin indicate full marks.

Group 'A'

Attempt any Two questions (2 × 10 = 20)

1. What is multiple Linear Regression (MLR)? From following information of variables X_1, X_2 and Y.

   $\Sigma X_1 = 272$, $\Sigma X_2 = 441$, $\Sigma Y = 147$, $\Sigma X_1^2 = 7428$, $\Sigma X_2^2 = 19461$, $\Sigma Y^2 = 2173$, $\Sigma X_1Y = 4013$, $\Sigma X_1X_2 = 12005$, $\Sigma X_2Y = 6485$, $n = 10$. Fit regression equation Y on X_1 and X_2 Interpret the regression coefficient.

2. What do you mean by Latin Square Design? Write down its merit and demerit. Set up the analysis of variance for the following of design.

A (10)	B (15)	C(20)
B (25)	C (10)	A (15)
C (25)	A (20)	B (15)

3. What do you mean by hypothesis? Describe null and alternative hypothesis. A company claims that its light bulbs are superior to those of the competitor on the basis of study which showed that a sample of 40 of its bulbs had an average life time 628 hours of continuous use with a standard deviation of 27 hours. While sample of 30 bulbs made by the competitor had and average life time 619 hours of continuous use with a standard deviation of 25 hours. Test at 5% level of significance, whether this claim is justified.

Group 'B'

Attempt any Eight questions: (8 × 5 = 40)

4. Suppose we are given following information with n = 7, multiple regression mode is $\hat{Y} = 8.15 + 0.56X_1 + 0.54X_2$

   Here,
   Total sum of square = 1493,
   Sum of square due to error = 91

   Find i) R^2 and interpret it. ii) Test the overall significance of model.

5. The following data related to the number of children classified according to the type of feed and the nature of teeth.

Type of feed	Nature of Teeth	
	Normal	Defective
Breast	18	12
Bottle	2	13

Do the information provide sufficient evidence to conclude that type of feeding and nature of teeth are dependent? Use chi square test at 5% level of significance

6. Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% level of confidence when coefficient of variation is 60%.

7. A manufacture of computer paper has a production process that operates continuously throughout an entire production shift. The paper is expected to have an average length of 11 inches and standard deviation is known to be 0.01 inch. Suppose random sample of 100 sheets is selected and the average paper length is found to be 10.68 inches. Set up 95% and 90% confidence interval estimate of the population average paper length.

8. A chemist use three catalyst for distilling alcohol and lay out were tabulated below

Catalyst	Alcohol (in cc)				
C_1	380	430	410		
C_2	290	350	270	250	270
C_3	400	380	450		

Are there any significant difference between catalyst? Text at 5% level of significance. use Kruskal Walli's H test.

9. Consider the partially completed ANOVA table below. Complete the ANOVA table and answer the following:

Source of Variation	Sum of Square	Degree of freedom	Mean sum of square	F value
Column	72	?	?	2
Rows	?	?	36	?
Treatments	180	3	?	?
Error	?	6	12	
Total	?	?		

i. What design was employed?

ii. How many treatments were compared?

10. Defined main component of queuing system.
11. Jobs are sent to main farm computer at a rate of 4 jobs per minute. Arrivals are modeled by a binomial process.
    i. Choose a frame size that makes the probability of a new received during each frame equal to 0.1.
    ii. using the chosen frame compute the probability of more than 4 jobs received during one minute.
    iii. Compute mean and variance of inter arrival time?
12. Write short notes of the following:
    i. Need of non parametric statistical methods.
    ii. Efficiency of Randomized Block Design relative to completely Randomized Design.