MODULE - 05

JOINT PROBABILITY DISTRIBUTION

INTRODUCTION

We have discussed probability distribution associated with a single random variable. The same can be generalized for two or more random variables. We discuss probability distributions associated with two random variables referred to as a joint distribution.

JOINT DISTRIBUTION AND JOINT PROBABILITY DISTRIBUTION

If X & Y are two discrete random variables, we define the joint probability function of X & Y by

$P(X = x, Y = y) = f(x,y)$

Where f(x,y) satisfy conditions

$f(x,y) \geq 0$ and $\sum_x \sum_y f(x,y) = 1$

The second condition means that the sum over all the values of x and y is equal to one.

Suppose $X = \{x_1, x_2, \ldots x_m\}$ and $Y = \{y_1, y_2, \ldots y_n\}$ then $P(X = x_i, Y = y_i)$ denoted by J_{ij}.

It should be observed that f is a function on the Cartesian product of the sets X and Y as we have

$X \times Y = \{(x_1, y_1), (x_2, y_2) \ldots (x_m, y_n)\}$

f is also referred to as joint probability density function of X and Y in the respective order. The set of values of this function $f(x_i, y_i) = J_{ij}$ for i = 1,2,...m, j = 1,2,...n is called the joint probability distribution of X and Y.These values are presented in the form of a two way table called the joint probability table.

X \ Y	y_1	y_2	...	y_n	Sum
x_1	J_{11}	J_{12}		J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	...	J_{2n}	$f(x_2)$
...	...		...	...	...
x_m	J_{m1}	J_{m2}	...	J_{mn}	$f(x_m)$
sum	$g(y_1)$	$g(y_2)$		$g(y_n)$	1

MARGINAL PROBABILITY DISTRIBUTION

In the joint probability table $\{f(x1), f(x2) \ldots f(xm)\}$ are the sum of horizontal entries and $\{g(y1), g(y2), \ldots g(yn)\}$ are the sum of vertical entries in the joint probability distribution table. These are called marginal probability distribution of X and Y respectively.

INDEPENDENT RANDOM VARIABLES

The discrete random variable X and Y are said to be independent random variables if $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_i)$

i.e $f(x_i)\, g(y_j) = J_{ij}$

Expectation, Variance, Covariance and Correlation

Expectation

$\mu_X = E(X) = \sum_x \sum_y x\, f(x,y) = \sum_i x_i f(x_i)$

$\mu_Y = E(Y) = \sum_x \sum_y y\, f(x,y) = \sum_j y_j g(y_j)$

$\mu_{XY} = E(XY) = \sum_i \sum_j x_i y_j J_{ij}$

Variance

$\sigma_X^2 = E(X^2) - [E(X)]^2$

$\sigma_Y^2 = E(Y^2) - [E(Y)]^2$

Covariance

$COV(X,Y) = E(XY) - E(X) \cdot E(Y)$

Correlation

Correlation of X and Y = $\rho(X,Y) = \frac{COV\ (X,Y)}{\sigma_X \sigma_Y}$

NOTE:

If X and Y are independents, E(X , Y) = E(X)· E(Y) and hence

COV (X , Y) = 0 = $\rho(X,Y)$

PROBLEMS

1. The joint probability distribution of two random variables X and Y ia as follows.

X \ Y	-4	2	7
1	1/8	1//4	1/8
5	1/4	1/8	1/8

Compute the following

(a) E(X) and E(Y) (b) E(XY) (c) σ_X and σ_y (d) COV (X , Y) (e) $\rho(X,Y)$

Solu: The distribution is obtained adding the all the respective row entries and also the respective coloumn entries.

Distribution of X :

x_i	1	5
$f(x_i)$	1/2	1/2

Distribution of Y :

y_j	-4	2	7
$g(y_j)$	3/8	3/8	1/4

a) E(X) = $\sum x_i f(x_i)$ = (1)(1/2) + 5 (1/2) = 3 = μ_x

E(Y) = $\sum y_j g(y_j)$ = (-4)(3/8) + 2(3/8) + 7(1/4) = 1 = μ_y

b) E (XY) = $\sum x_i y_j J_{ij}$= (1)(-4)(1/8) + (1) (2) (1/4) + (1) (7) (1/8)

+(5) (-4) (1/4)+ (5) (2) (1/8)+ (5) (7) (1/8)

= 3/2

c) σ_X^2 = E (X^2) – $[E(X)]^2$ and σ_Y^2 = E (Y^2) – $[E(Y)]^2$

Now E (X^2) = $\sum x^2 f(x_i)$ = (1)(1/2)+25(1/2) = 13

E (Y^2) = $\sum y^2 g(y_j)$ = (16)(3/8)+(4)(3/8) + (48)(1/4) = 79/4

Hence $\sigma_X^2 = 13 - (3)^2 = 4$ and $\sigma_Y^2 = (79/4) - (1)^2 = 75/4$

Thus $\sigma_X = 2$ and $\sigma_y = \sqrt{(\frac{75}{4})} = 4.33$

d) COV (X,Y) = E (XY) – E(X)· E(Y)

= (3/2) – 3 (1) = - 3/2

e) $\rho(X,Y) = \frac{COV\ (X,Y)}{\sigma_X \sigma_Y} = \frac{(-\frac{3}{2})}{(2)\sqrt{(\frac{75}{4})}} = -\ 0.\ 1732.$

2. The joint probability distribution table for two random variables X and Y is as follows.

X \ Y	-2	-1	4	5
1	0.1	0.2	0	0.3
2	0.2	0.1	0.`1	0

Determine the marginal probability distributions of X and Y. Also compute

(a) Expectations of X , Y and XY

(b) S.D's of X,Y

(c) covariance of X and Y (d) Correlation of X and Y

Further verify that X and Y are dependent random variables

Solu: Marginal distributions of X and Y are got by adding all the respective row entries and the respective column entries.

x_i	1	2
$f(x_i)$	0.6	0.4

y_j	-2	-1	4	5
$g(y_j)$	0.3	0.3	0.1	0.3

(a)

= E(X) = $\sum x_i f(x_i)$ = (1)(0.6) + (2)(0.4) = 1.4

μ_y = E(Y) = $\sum y_j g(y_j)$ = (-2) (0.3) + (-1)(0.3) + 4 (0.1) + 5 (0.3) = 1

E (XY) = $\sum x_i y_j J_{ij}$ = (1)(-2)(0.1) + (1) (-1) (0.2) + (1) (4) (0) + (1) (5) (0.3)

+ (2) (-2) (0.2) + (2) (-1) (0.1) + (2) (4) (0.1) + (2)(5) (0)

= 0.9

b) σ_X^2 = E (X^2) – $[E(X)]^2$ and σ_Y^2 = E (Y^2) – $[E(Y)]^2$

Now E (X^2) = $\sum x^2 f(x_i)$ = (1)(0.6) + (4) (0.4) = 2.2

E (Y^2) = $\sum y^2 g(y_j)$ = (4)(0.3) + 1(0.3) + 16(0.1) + 25(0.3) = 10.6

Hence σ_X^2 = 2.2 – $(1.4)^2$ = 0.245 and σ_Y^2 = (10.6) – $(1)^2$= 9.6

Thus σ_X = 0,49 and σ_y = 3.1

c) COV (X,Y) = E (XY) – E(X)· E(Y)

= 0.9 – 1.4(1) = - 0.5

e) $\rho(X,Y) = \frac{COV\ (X,Y)}{\sigma_X \sigma_Y} = \frac{(-0.5)}{(0.49)(3.1)}$ = - 0.3.

If X and Yare independent random variables we must have

$f(x_i)\ g(y_j) = J_{ij}$

It can be seen that $f(x_1)g(y_1)$ = (0.6)(0.3) = 0.18 and J_{11} = 0.1

i.e $f(x_1)g(y_1) \neq J_{11}$

Hence we conclude that X and Y are dependent random variables.

3.The joint probability distribution of two discrete random variables X and Y is given by f(x,y)= k(2x + y) where x and y are integers such that $0 \leq x \leq 2$, $0 \leq y \leq 3$.

(a) Find the value of the constant k

(b) Find the marginal probability distributions of X and Y

(C) Show that the random variable X and Y are dependent.

Solu: X = $\{x_i\}$ = {0,1,2} and Y = $\{y_j\}$ = {0,1,2,3}

f(x,y) = k(2x+y) and the joint probability distribution table is

formed as follows.

X / Y	0	1	2	3	Sum
0	0	k	2k	3k	6k
1	2k	3k	4k	5k	14k
2	4k	5k	6k	7k	22
Sum	6k	9k	12k	15k	42k

a) We must have $42k = 1$

$\therefore k = 1/42$

b) Marginal probability distribution is as follows.

x_i	0	1	2
$f(x_i)$	6/42 = 1/7	4/42 =1/3	22/42 =11/21

y_j	0	1	2	3
$g(y_j)$	6/42 =1/7	9/42 =3/14	12/4 =2/7	15/42 =5/14

c) It can be easily seen that $f(x_i)\, g(y_j) \neq J_{ij}$

Hence the random variables ade dependent.

4. A fair coin is tossed thrice. The random variables X and Y are defined as follows. X = 0 or 1 according as head or tail occurs on the first toss.

Y = Number of heads

(a) Determine the distribution of X and Y

(b)Determine the joint distribution of X and Y

(C) Obtain the expectations of X,Y and XY. Also find S.Ds of X and Y

(d) Compute Covariance and Correlation of X and Y.

Solu. The sample space S and the association of random variables X and Y is given by the following table

S	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	0	0	0	0	1	1	1	1
Y	3	2	2	1	2	1	1	0

(a) The probability distribution of X and Y is found as follows.

X = $\{x_i\}$ = {0,1} and Y = $\{y_j\}$ = {0,1,2,3}

P(X=0) is 4/8 = 1/2, P(X = 1) is 4/8 = 1/2

P(Y=0) is 1/8 , P(Y = 1) is 3/8

P(Y=2) is 3/8 , P(Y = 3) is 1/8

Thus we have the following probability distribution of X and Y

x_i	0	`1
$f(x_i)$	1/2	1/2

y_j	0	1	2	3
$g(y_j)$	1/8	3/8	3/8	1/8

(b)The joint distribution of X and Y is found by computing

J_{ij} = P(X = x_i , Y = y_j) where we have

X_1 = 0, X_2 = 1 and y_1 = 0, y_2 = 1 , y_3 = 2 , y_4 = 3

J_{11} = P (X = 0, Y = 0) = 0

(X = 0 implies that their is a head turn out and Y the total

number of heads 0 is impossible)

J_{12} = P (X = 0, Y = 1) = 1/8 correspondings to the outcome HTT

J_{13} = P (X = 0, Y = 2) = 2/8=1/4; outcomes are HHT and HTH

J_{14} = P (X = 0, Y = 3) = 1/8;outcome is HHH

J_{21} = P (X = 1, Y = 0) = 1/8,outcome is TTT

J_{22} = P (X = 1, Y = 1) = 2/8=1/4; outcomes are THT ,TTH

J_{23} = P (X = 1, Y = 2) = 1/8,outcome is THH

J_{24} = P (X = 1, Y = 3) = 0 since the outcome is impossible.

(These values can be written quickly by looking at the table of S , X,Y)

The required joint probability distribution of X and Y is as follows.

X / Y	0	1	2	3	Sum
0	0	1/8	1/4	1/8	1/2
1	1/8	1/4	1/8	0	1/2
Sum	1/8	3/8	3/8	1/8	1

(c) μ_x = E(X) = $\sum x_i f(x_i)$ = (0) (1/2) + (1)(1/2) = 1/2

μ_y = E(Y) = $\sum y_j g(y_j)$ = (0) (1/8) + (1)(3/8) + 2 (3/8) + 3 (1/8) = 12/8 =3/2

E (XY) = $\sum x_i y_j J_{ij}$ = 0 + (0+ ¼+ 2/8 + 0) = 1/2

σ_X^2 = E (X^2) – $[E(X)]^2$ and σ_Y^2 = E (Y^2) – $[E(Y)]^2$

σ_X^2 = (0 + ½) – ¼ = ¼ σ_Y^2 = (0+3/8 + 3/2 + 9/8) – (9/4) = 3- (9/4) = 3/4

Thus σ_X = 1/2 and $\sigma_y = \sqrt{3/2}$

c) COV (X,Y) = E (XY) – E(X)· E(Y)

= ½ - ¾ = - 1/4

$\rho(X,Y) = \frac{COV\ (X,Y)}{\sigma_X \sigma_Y} = \frac{(-1/4)}{\sqrt{3/4}} = -\frac{1}{\sqrt{3}}$

SAMPLING THEORY

INTRODUCTION

Statistical Inference is a branch of Statistics which uses probability concepts to deal with uncertainty in decision making. There are a number of situations where in we come across problems involving decision making. For example, consider the problem of buying 1 kilogram of rice, when we visit the shop, we do not check each and every rice grains stored in a gunny bag; rather we put our hand inside the bag and collect a sample of rice grains. Then analysis takes place. Based on this, we decide to buy or not. Thus, the problem involves studying whole rice stored in a bag using only a sample of rice grains.

This topic considers two different classes of problems

1. Hypothesis testing – we test a statement about the population parameter from which the sample is drawn.
2. Estimation – A statistic obtained from the sample collected is used to estimate the population parameter.

First what is meant by hypothesis testing?

This means that testing of hypothetical statement about a parameter of population.

Conventional approach to testing:

The procedure involves the following:

1. First we set up a definite statement about the population parameter which we call it as null hypothesis, denoted by H_0. According to Professor R. A. Fisher,

Null Hypothesis is the statement which is tested for possible rejection under the assumption that it is true. Next we set up another hypothesis called alternate statement which is just opposite of null statement; denoted by H_1 which is just complimentary to the null hypothesis. Therefore, if we start with $H_0 : \mu = \mu_0$ then alternate hypothesis may be considered as either one of the following statements; $H_1 : \mu \neq \mu_0$, or $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$.

As we are studying population parameter based on some sample study, one can not do the job with 100% accuracy since sample is drawn from the population and possible sample may not represent the whole population. Therefore, usually we conduct analysis at certain level of significance (lower than 100%. The possible choices include 99%, or 95% or 98% or 90%. **Usually we conduct analysis at 99% or 95% level of significance**, denoted by the symbol α. We test H_0 against H_1 at certain level of significance. The confidence with which a person rejects or accepts H_0 depends upon the significance level adopted. It is usually expressed in percentage forms such as 5% or 1% etc. Note that when α is set as 5%, then probability of rejecting null hypothesis when it is true is only 5%. It also means that when the hypothesis in question is accepted at 5% level of significance, then statistician runs the risk of taking wrong decisions, in the long run, is only 5%. The above is called II step of hypothesis testing.

Critical values or Fiducial limit values for a two tailed test:

Sl. No	Level of significance	Theoretical Value
1	$\alpha = 1\%$	2.58
2	$\alpha = 2\%$	2.33
3	$\alpha = 5\%$	1.96

Critical values or Fiducial limit values for a single tailed test (right and test)

Tabulated value	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
Right – tailed test	2.33	1.645	1.28
Left tailed test	-2.33	-1.645	-1.28

Setting a test criterion: The third step in hypothesis testing procedure is to construct a test criterion. This involves selecting an appropriate probability distribution for the particular test i.e. a proper probability distribution function to be chosen. Some of the distribution functions used are t, F, when the sample size is small (size lower than 30).

However, for large samples, normal distribution function is preferred. Next step is the computation of statistic using the sample items drawn from the population. Usually, samples are drawn from the population by a procedure called random, where in each and every data of the population has the same chance of being included in the sample. Then the computed value of the test criterion is compared with the tabular value; as long the calculated value is lower then or equal to tabulated value, we accept the null hypothesis, otherwise, we reject null hypothesis and accept the alternate hypothesis. Decisions are valid only at the particular level significance of level adopted.

During the course of analysis, there are two types of errors bound to occur. These are **(i) Type – I error and (ii) Type – II error.**

Type – I error: This error usually occurs in a situation, when the null hypothesis is true, but we reject it i.e. rejection of a correct/true hypothesis constitute type I error.

Type – II error: Here, null hypothesis is actually false, but we accept it. Equivalently, accepting a hypothesis which is wrong results in a type – II error. The probability of committing a type – I error is denoted by α where

α = Probability of making type I error = Probability [Rejecting H_0 | H_0 is true]

On the other hand, type – II error is committed by not rejecting a hypothesis when it is false. The probability of committing this error is denoted by β . Note that

β = Probability of making type II error = Probability [Accepting H_1 | H_1 is false]

Critical region:

A region in a sample space S which amounts to Rejection of H_0 is termed as critical region.

One tailed test and two tailed test:

This depends upon the setting up of both null and alternative hypothesis.

A note on computed test criterion value:

1. When the sampling distribution is based on population of proportions/Means, then test criterion may be given as

$$Z_{cal} = \frac{\text{Expected results - Observed results}}{\text{Standard error of the distribution}}$$

Application of standard error:

1. S.E. enables us to determine the probable limit within which the population parameter may be expected to lie. For example, the probable limits for population of proportion are given by $p \pm 3\sqrt{pqn}$. Here, p represents the chance of achieving a success in a single trial, q stands for the chance that there is a failure in the trial and n refers to the size of the sample.
2. The magnitude of standard error gives an index of the precision of the parameter.

ILLUSTRATIVE EXAMPLES

1. **A coin is tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is un– biased?**

Solution: First we construct null and alternate hypotheses set up H_0 : The coin is not a biased one. Set up H_1 : Yes, the coin is biased. As the coin is assumed be fair and it is tossed 400 times, clearly we must expect 200 times heads occurring and 200 times tails. Thus, expected number of heads is 200. But the observed result is 216. There is a difference of 16. Further, standard error is $\sigma = \sqrt{npq}$. With p = ½, q = ½ and n = 400, clearly $\sigma = 10$. The test criterion is $z_{cal} = \dfrac{\text{difference}}{\text{standard error}} = \left|\dfrac{216-240}{10}\right| = 1.6$

If we choose $\alpha = 5\%$, then the tabulated value for a two tailed test is 1.96. Since, the calculated value is lower than the tabulated value; we accept the null hypothesis that coin is un – biased.

2. **A person throws a 10 dice 500 times and obtains 2560 times 4, 5, or 6. Can this be attributed to fluctuations in sampling?**

Solution: As in the previous problem first we shall set up H_0: The die is fair and H_1: The die is unfair. We consider that problem is based on a two – tailed test. Let us choose level of significance as $\alpha = 5\%$ then, the tabulated value is 1.96. Consider computing test criterion, $z_{cal} = \left|\dfrac{\text{Expected value - observed result}}{\text{standard error}}\right|$; here, as the dice is tossed by a person 5000 times, and on the basis that die is fair, then chance of getting any of

the 6 numbers is 1/6. Thus, chance of getting either 4 or 5, or 6 is p = ½. Also, q = ½. With n = 5000, standard error, $\sigma = \sqrt{npq}$ = 35.36. Further, expected value of obtaining 4 or 5 or 6 is 2500. Hence, $z_{cal} = \left|\frac{2500 - 2560}{35.36}\right| = 1.7$ which is lower than 1.96. Hence, we conclude that die is a fair one.

3. A sample of 1000 days is taken from meteorological records of a certain district and 120 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district?

Solution: Let p denote the probability that a day is foggy in nature in a district as reported by meteorological records. Clearly, $p = \frac{120}{1000} = 0.12$ and q = 0.88. With n = 1000, the probable limits to the percentage of foggy days is given by $p \pm 3\sqrt{pqn}$. Using the data available in this problem, one obtains the answer as $0.12 \pm 3\sqrt{0.12 \cdot 88 \cdot 1000}$. Equivalently, 8.91% to 15.07%.

4. A die was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate that die is biased? (Model Question Paper Problem)

Solution: We set up the null hypothesis as H_0 : Die is un - biased. Also, H_1 : Die is biased. Let us take level of significance as $\alpha = 5\%$. Based on the assumption that distribution is normally distributed, the tabulated value is 1.96. The chance of getting each of the 6 numbers is same and it equals to 1/6 therefore chance of getting either 5 or 6 is 1/3. In a throw of 9000 times, getting the numbers either 5 or 6 is $\frac{1}{3} \times 9000 = 3000$. Now the difference in these two results is 240. With p = 1/3, q = 2/3, n = 9000, $S.E. = \sqrt{npq}$ = 44.72. Now consider the test criterion $z_{cal} = \frac{\text{Difference}}{\text{S.E.}}$ = $= \frac{240}{44.72} = 5.367$ which is again more than the tabulated value. Therefore, we reject null hypothesis and accept the alternate that die is highly biased.

Test for significance for large samples

In the previous section, we discussed problems pertaining to sampling of attributes. It is time to think of sampling of other variables one may come across in a practical situation such as height weight etc. We say that a sample is small when the size is usually lower than 30, otherwise it is called a large one.

The study here is based on the following assumptions: (i) the random sampling distribution of a statistic is approximately normal and (ii) values given by the samples are sufficiently close to the population value and can be used in its place for calculating standard error. When the standard deviation of population is known, then $S.E(\overline{X}) = \frac{\sigma_p}{\sqrt{n}}$ where σ_p denotes the standard deviation of population. When the standard deviation of the population is unknown, then $S.E(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the sample.

Fiducial limits of population mean are:

95% fiducial limits of population mean are $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$

99% fiducial limits of population mean are $\overline{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$. Further, test criterion $z_{cal} = \left|\frac{\overline{x} - \mu}{S.E.}\right|$

ILLUSTRATIVE EXAMPLES

1. **A sample of 100 tyres is taken from a lot. The mean life of tyres is found to be 39, 350 kilo meters with a standard deviation of 3, 260. Could the sample come from a population with mean life of 40, 000 kilometers? Establish 99% confidence limits within which the mean life of tyres is expected to lie.**

Solution: First we shall set up null hypothesis, $H_0 : \mu = 40{,}000$, alternate hypothesis as $H_1 : \mu \neq 40{,}000$. We consider that the problem follows a two tailed test and chose $\alpha = 5\%$. Then corresponding to this, tabulated value is 1.96. Consider the expression for finding test criterion, $z_{cal} = \left|\frac{\overline{x} - \mu}{S.E.}\right|$. Here, μ=40, 000, $\overline{x} = 39, 350$ and

$\sigma = 3,260$, n = 100. S.E. $= \frac{\sigma}{\sqrt{n}} = \frac{3,260}{\sqrt{100}} = 326$. Thus, $z_{cal} = 1.994$. As this value is slightly greater than 1.96, we reject the null hypothesis and conclude that sample has not come from a population of 40, 000 kilometers.

The 99% confidence limits within which population mean is expected to lie is given as $\overline{x} \pm 2.58 \times$S.E. i.e. $39,350 \pm 2.58 \times 326 = (38, 509, 40, 191)$.

2. **The mean life time of a sample of 400 fluorescent light bulbs produced by a company is found to be 1, 570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of bulbs is 1600 hours against the alternative hypothesis that it is greater than 1, 600 hours at 1% and 5% level of significance.**

Solution: First we shall set up null hypothesis, $H_0 : \mu = 1,600$ hours, alternate hypothesis as $H_1 : \mu > 1,600$ hours. We consider that the problem follows a two tailed test and chose $\alpha = 5\%$. Then corresponding to this, tabulated value is 1.96. Consider the expression for finding test criterion, $z_{cal} = \left| \frac{\overline{x} - \mu}{\text{S.E.}} \right|$. Here, μ=1, 600, $\overline{x} = 1, 570$, n = 400, $\sigma = 150$ hours so that using all these values above, it can be seen that $z_{cal} = 4.0$ which is really greater than 1.96. Hence, we have to reject null hypothesis and to accept the alternate hypothesis.

Test of significance of difference between the means of two samples

Consider two populations P1 and P2. Let S_1 and S_2 be two samples drawn at random from these two different populations. Suppose we have the following data about these two samples, say

Samples/Data	Sample size	Mean	Standard Deviation
S_1	n_1	$\overline{x_1}$	σ_1
S_2	n_2	$\overline{x_2}$	σ_2

then standard error of difference between the means of two samples S_1 and S_2 is $S.E = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ and the test criterion is $Z_{cal} = \frac{\text{Difference of sample means}}{\text{Standard error}}$. The rest of the analysis is same as in the preceding sections.

When the two samples are drawn from the same population, then standard error is

$S.E = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and test criterion is $Z_{cal} = \frac{\text{Difference of sample means}}{\text{Standard error}}$.

When the standard deviations are un – known, then standard deviations of the two samples must be replaced. Thus, $S.E = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where s_1 and s_2 are standard deviations of the two samples considered in the problem.

ILLUSTRATIVE EXAMPLES

1. Intelligence test on two groups of boys and girls gave the following data:

Data	Mean	Standard deviation	Sample size
Boys	75	15	150
Girls	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls?

Solution: We set up null hypothesis as H_0 : there is no significant difference between the mean scores obtained by boys and girls. The alternate hypothesis is considered as H_1 : Yes, there is a significant difference in the mean scores obtained by boys and girls. We choose level of significance as $\alpha = 5\%$ so that tabulated value is 1.96. Consider $z_{cal} = \frac{\text{Difference of means}}{\text{Standard Error}}$. The standard error may be calculated as $S.E = \sqrt{\frac{15^2}{150} + \frac{20^2}{250}} = 1.761$, The test criterion is $z_{cal} = \frac{75 - 70}{1.761} = 2.84$. As 2.84 is more than 1.96, we have to reject null hypothesis and to accept alternate hypothesis that there are some significant difference in the mean marks scored by boys and girls.

2. A man buys 50 electric bulbs of "Philips" and 50 bulbs of "Surya". He finds that Philips bulbs give an average life of 1,500 hours with a standard deviation of 60 hours and Surya bulbs gave an average life of 1, 512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbs?

Solution: we set up null hypothesis, H_0 : there is no significant difference between the bulbs made by the two companies, the alternate hypothesis can be set as H_1 : Yes, and there could be some significant difference in the mean life of bulbs. Taking $\alpha = 1\%$ and $\alpha = 5\%$, the respective tabulated values are 2.58 and 1.96. Consider standard error is $S.E = \sqrt{\frac{60^2}{50} + \frac{80^2}{50}} = 14.14$ so that $z_{cal} = \frac{1512 - 1500}{14.14} = 0.849$. Since the calculated value is certainly lower than the two tabulated values, we accept the hypothesis there is no significant difference in the make of the two bulbs produced by the companies.

A discussion on tests of significance for small samples

So far the problem of testing a hypothesis about a population parameter was based on the assumption that sample drawn from population is large in size (more than 30) and the probability distribution is normally distributed. However, when the size of the sample is small, (say < 30) tests considered above are not suitable because the assumptions on which they are based generally do not hold good in the case of small samples. IN particular, here one cannot assume that the problem follows a normal distribution function and those values given by sample data are sufficiently close to the population values and can be used in their place for the calculation of standard error. Thus, it is a necessity to develop some alternative strategies to deal with problems having sample size relatively small. Also, we do see a number of problems involving small samples. With these in view, here, we will initiate a detailed discussion on the same.

Here, too, the problem is about testing a statement about population parameter; i.e. in ascertaining whether observed values could have arisen by sampling fluctuations from some value given in advance. For example, if a sample of 15 gives a correlation coefficient of +0.4, we shall be interested not so much in the value of the correlation in the parent population, but more generally this value could have come from an un – correlated population, i.e. whether it is significant in the parent population. It is widely accepted that when we work with small samples, estimates will vary from sample to sample.

Further, in the theory of small samples also, we begin study by making an assumption that parent population is normally distributed unless otherwise stated. Strictly, whatever the decision one takes in hypothesis testing problems is valid only for normal populations.

Sir William Gosset and R. A. Fisher have contributed a lot to theory of small samples. Sir W. Gosset published his findings in the year 1905 under the pen name "student". He gave a test popularly known as "t – test" and Fisher gave another test known as "z – test". These tests are based on "t distribution and "z – distribution".

Student's t - distribution function

Gosset was employed by the Guinness and Son, Dublin bravery, Ireland which did not permit employees to publish research work under their own names. So Gosset adopted the pen name "student" and published his findings under this name. Thereafter, the t – distribution commonly called student's t – distribution or simply student's distribution.

The t – distribution to be used in a situation when the sample drawn from a population is of size lower than 30 and population standard deviation is un – known. The t – statistic, t_{cal} is defined as $t_{cal} = \left(\frac{\bar{x} - \mu}{S}\right)\cdot\sqrt{n}$ where $S = \sqrt{\frac{\sum_{i=1}^{i=n} x_i - \bar{x}^2}{n-1}}$, $\bar{x}$ is the sample mean, n is the sample size, and x_i are the data items. $\gamma = n - 1 = 12$

The t – distribution function has been derived mathematically under the assumption of a normally distributed population; it has the following form

$f(t) = C\left(1 + \frac{t^2}{\gamma}\right)^{-\left(\frac{\gamma+1}{2}\right)}$ where C is a constant term and γ = n - 1 denotes the number of degrees of freedom. As the p.d.f. of a t – distribution is not suitable for analytical treatment. Therefore, the function is evaluated numerically for various values of t, and for particular values of γ . The t – distribution table normally given in statistics text books gives, over a range of values of γ, the probability values of exceeding by chance value of t at different levels of significance. The t – distribution function has a different value for each degree of freedom and when degrees of freedom approach a large value, t – distribution is equivalent to normal distribution function.

The application of t – distribution includes (i) testing the significance of the mean of a random sample i.e. determining whether the mean of a sample drawn from drawn from a normal population deviates significantly from a stated value (i.e. hypothetical value of the populations mean) and (ii) testing whether difference between means of two independent samples is significant or not i.e. ascertaining whether the two samples comes from the same normal population? (iii) Testing difference between means of two dependent samples is significant? (iv) Testing the significance of on observed correlation coefficient.

Procedures to be followed in testing a hypothesis made about the population parameter using student's t - distribution:

- As usual first set up null hypothesis,
- Then, set up alternate hypothesis,
- Choose a suitable level of significance,

- Note down the sample size, n and the number of degrees of freedom,
- Compute the theoretical value, t_{tab} by using t – distribution table.
- t_{tab} value is to be obtained as follows: If we set up $\alpha = 5\% = 0.05$, suppose $\gamma = 9$ then, t_{tab} is to be obtained by looking in 9th row and in the column $\alpha = 0.025$ (i.e. half of $\alpha = 0.05$) .
- The test criterion is then calculated using the formula, $t_{cal} = \left(\frac{\bar{x} - \mu}{S}\right) \cdot \sqrt{n}$
- Later, the calculated value above is compared with tabulated value. As long as the calculated value matches with the tabulated value, we as usual accept the null hypothesis and on the other hand, when the calculated value becomes more than tabulated value, we reject the null hypothesis and accept the alternate hypothesis.

ILLUSTRATIVE EXAMPLES

1. The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. Random samples of 6 such bulbs have the following values: Life of bulbs in months: 24, 20, 30, 20, 20, and 18. Can you regard the producer's claim to valid at 1% level of significance? (Given that $t_{tab} = 4.032$ corresponding to $\gamma = 5$).

Solution: To solve the problem, we first set up the null hypothesis $H_0 : \mu = 25$ months, alternate hypothesis may be treated as $H_0 : \mu < 25$ months. To set up $\alpha = 1\%$, then tabulated value corresponding to this level of significance is $t_{tab}|_{\alpha=1\% \text{ and } \gamma=5} = 4.032$ (4.032 value has been got by looking in the 5th row) . The test criterion is given by $t_{cal} = \left(\frac{\bar{x} - \mu}{S}\right) \cdot \sqrt{n}$ where $s = \sqrt{\frac{\sum_{i=1}^{n} x_i - \bar{x}^2}{n-1}}$.

Consider

x_i	$\bar{x}$	$x_i - \bar{x}$	$x_i - \bar{x}^2$
24		1	1
26		3	9
30		7	49
20	23	-3	9
20		-3	9
18		-5	25
Total = 138		-	Total = 102

Thus, $S=\sqrt{\frac{102}{5}}=\sqrt{20.4}=4.517$ and $t_{cal}=\left|\frac{23-25}{4.517}\right|\sqrt{6}=1.084$. Since the calculated value, 1.084 is lower than the tabulated value of 4.032; we accept the null hypothesis as mean life of bulbs could be about 25 hours.

2. A certain stimulus administered to each of the 13 patients resulted in the following increase of blood pressure: 5, 2, 8,-1, 3, 0, -2, 1, 5, 0, 4, 6, 8. Can it be concluded that the stimulus, in general, be accompanied by an increase in the blood pressure? (Model Question Paper Problem)

Solution: We shall set up $H_0: \mu_{before}=\mu_{after}$ i.e. there is no significant difference in the blood pressure readings before and after the injection of the drug. The alternate hypothesis is $H_0: \mu_{before}>\mu_{after}$ i.e. the stimulus resulted in an increase in the blood pressure of the patients. Taking $\alpha=1\%$ and $\alpha=5\%$, as n = 13, $\gamma=n-1=12$, respective tabulated values are $t_{tab}|_{\alpha=1\% \text{ and } \gamma=12}=3.055$ and $t_{tab}|_{\alpha=5\% \text{ and } \gamma=12}=2.179$. Now, we compute the value of test criterion. For this, consider

x_i	$\bar{x}$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
5		2	4
2		-1	1
8		5	25
-1		-4	16
3		0	0
0		-3	9
-2		-5	25
1		-2	4
5		2	4
0		-3	9
4		1	1
6		3	9
8		5	25
Total = 39		-	Total = 132

3

Consider $S=\sqrt{\dfrac{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\dfrac{132}{12}}=\sqrt{11}=3.317$. Therefore, $t_{cal}=\left|\dfrac{\bar{x}-\mu}{S}\right|\cdot\sqrt{n}$ may be obtained as $t_{cal}=\left|\dfrac{0-3}{3.317}\right|\sqrt{13}=3.2614$. As the calculated value 3.2614 is more than the tabulated values of 3.055 and 2.179, we accept the alternate hypothesis that after the drug is given to patients, there is an increase in the blood pressure level.

3. the life time of electric bulbs for a random sample of 10 from a large consignment gave the following data: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6 (in '000 hours). Can we accept the hypothesis that the average life time of bulbs is 4, 000 hours?

Solution: Set up $H_0:\mu=4{,}000$ hours , $H_1:\mu<4{,}000$ hours . Let us choose that $\alpha=5\%$. Then tabulated value is $t_{tab}|_{\alpha=5\% \text{ and } \gamma=9}=2.262$. To find the test criterion, consider

x_i	$\bar{x}$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
4.2	4.4	-0.2	0.04
4.6		0.2	0.04
3.9		-0.5	0.25
4.1		-0.3	0.09
5.2		0.8	0.64
3.8		-0.6	0.36
3.9		-0.5	0.25
4.3		-0.1	0.01
4.4		0.0	0.0
5.6		1.2	1.44
Total = 44		-	**Total = 3.12**

Consider $S=\sqrt{\dfrac{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\dfrac{3.12}{9}}=0.589$. Therefore, $t_{cal}=\left|\dfrac{\bar{x}-\mu}{S}\right|\cdot\sqrt{n}$ is computed as $t_{cal}=\left|\dfrac{4.4-4.0}{0.589}\right|\cdot\sqrt{10}=2.148$. As the computed value is lower than the tabulated value of 2.262, we conclude that mean life of time bulbs is about 4, 000 hours.

A discussion on χ^2 test and Goodness of Fit

Recently, we have discussed t – distribution function (i.e. t – test). The study was based on the assumption that the samples were drawn from normally distributed populations, or, more accurately that the sample means were normally distributed. Since test required such an assumption about population parameters. For this reason, A test of this kind is called parametric test. There are situations in which it may not be possible to make any rigid assumption about the distribution of population from which one has to draw a sample.

Thus, there is a need to develop some non – parametric tests which does not require any assumptions about the population parameters.

With this in view, now we shall consider a discussion on χ^2 distribution which does not require any assumption with regard to the population. The test criterion corresponding to this distribution may be given as $\chi^2 = \dfrac{\sum_i O_i - E_i^{\,2}}{E_i}$ where

- $E_i : \dfrac{RT \cdot CT}{N}$ O_i : Observed values , E_i : Expected values.

When Expected values are not given, one can calculate these by using the following relation; $E_i : \dfrac{RT \cdot CT}{N}$. Here, **RT** means the row total for the cell containing the row, **CT** is for the column total for the cell containing columns, and **N** represent the total number of observations in the problem.

The calculated χ^2 value (i.e. test criterion value or calculated value) is compared with the tabular value of χ^2 value for given degree of freedom at a certain prefixed level of significance. Whenever the calculated value is lower than the tabular value, we continue to accept the fact that there is not much significant difference between expected and observed results.
On the other hand, if the calculated value is found to be more than the value suggested in the table, then we have to conclude that there is a significant difference between observed and expected frequencies.

As usual, degrees of freedom are $\gamma = n - k$ where **k** denotes the number of independent constraints. Usually, it is 1 as we will be always testing null hypothesis against only one hypothesis, namely, alternate hypothesis.

This is an approximate test for relatively a large population.

For the usage of test, the following conditions must checked before employing the test. These are:

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any, must be linear.
3. i.e. the sum of all the observed values must match with the sum of all the expected values.
4. **N**, total frequency should be reasonably large
5. No theoretical frequency should be lower than 5.
6. It may be recalled this test is depends on χ^2 test: The set of observed and expected frequencies and on the degrees of freedom, it does not make any assumptions regarding the population.

ILLUSTRATIVE EXAMPLES

1. From the data given below about the treatment of 250 patients suffering from a disease, state whether new treatment is superior to the conventional test.

Data	Number of patients		
	Favourable	Not favorable	Total
New one	140	30	170
Conventional	60	20	80
Total	200	50	280

Solution: We set up null hypothesis as there is no significance in results due to the two procedures adopted. The alternate hypothesis may be assumed as there could be some difference in the results. Set up level of significance as $\alpha = 5\%$ then tabulated value is $\chi^2|_{\alpha=0.05,\,r=1} = 3.841$.

$$+\left(\frac{112-100}{100}\right)^2+\left(\frac{71-50}{50}\right)^2+\left(\frac{32-10}{10}\right)^2$$

Consider finding expected values given by the formula. $\text{Expectation}(AB) = \dfrac{RT \cdot CT}{N}$

where RT means that the row total for the row containing the cell, CT means that the total for the column containing the cell and N, total number of frequencies. Keeping these in view, we find that expected frequencies are

	B		
	136	34	170
A	64	16	80
	200	50	250

Note: $\frac{170 \cdot 200}{250} = 136$; $\frac{170 \cdot 50}{250} = 34$, $\frac{80 \cdot 200}{250} = 64$ and $\frac{80 \cdot 50}{250} = 16$.

O_i	E_i	$O_i - E_i$	$O_i - E_i^2$	$O_i - E_i^2 / E_i$
140	136	4	16	0.118
60	64	-4	16	0.250
30	34	-4	16	0.471
20	16	4	16	1.000
			Total	1.839

As the calculated value 1.839 is lower than the tabulated value $\chi^2|_{\alpha=0.05,\gamma=4} = 3.841$, we accept the null hypothesis, namely, that there is not much significant difference between the two procedures.

2. A set of five similar coins is tossed 320 times and the result is

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution function.

Solution: We shall set up the null hypothesis that data actually follows a binomial distribution. Then alternate hypothesis is, namely, data does not follow binomial distribution. Next, to set up a suitable level of significance, $\alpha = 5\%$, with n = 6, degrees of freedom is $\gamma = 5$. Therefore, the tabulated value is $\chi^2|_{\alpha=0.05,\gamma=5} = 11.07$. Before proceeding to finding test criterion, first we compute the various expected frequencies. As the data is set to be following binomial distribution, clearly probability density function is $b(n,p,k) = \binom{n}{k} p^k q^{n-k}$. Here, $n = 320$, $p = 0.5$, $q = 0.5$, and k takes the values right from 0 up to 5. Hence, the expected frequencies of getting 0, 1, 2, 3, 4, 5 heads are the successive terms of the binomial expansion of $320 \cdot (p+q)^5$. Thus, expected frequencies E_i are 10, 50, 100, 100, 50, 10. Consider the test criterion given by $\chi^2|_{cal} = \frac{\sum_i O_i - E_i^2}{E_i}$:

Here, observed values are: O_i: 6, 27, 72, 112, 71, 32

The expected values are: E_i: 10, 50, 100, 100, 50, 10. Consider

$$\chi^2_{cal} = \left(\frac{6-10}{10}\right)^2 + \left(\frac{27-50}{50}\right)^2 + \left(\frac{72-100}{100}\right)^2 + \left(\frac{112-100}{100}\right)^2 + \left(\frac{71-50}{50}\right)^2 + \left(\frac{32-10}{10}\right)^2 = 78.68.$$

As the calculated value is very much higher than the tabulated value of 3.841, we reject the null hypothesis and accept the alternate hypothesis that data does not follow the binomial distribution.