SAMPLING METHODS

1 POPULATION (UNIVERSE)

Before giving the notion of sampling, we will first define *population.* The group of individuals under study is called *population* or *universe.* It may be finite or infinite.

2 SAMPLING

A part selected from the population is called *a sample.* The process of selection of a sample is called sampling. A *Random sample* is one in which each member of population has an equal chance of being included in it. There are ${}^{N}C_{n}$ different samples of size n that can be picked up from a population of size N.

3 PARAMETERS AND STATISTICS

The statistical constants of the population such as mean (μ), standard deviation (σ) are called parameters. Parameters are denoted by Greek letters.

The mean $(\bar{x})$, standard deviation $|S|$ of a sample are known as statistics. Statistics are denoted by Roman letters.

Symbols for Population and Samples

Characteristic	Population	Sample
	Parameter	Statistic
Symbols	population size = N population mean = μ population standard deviation = σ population proportion = p	sample size = n sample mean = $\bar{x}$ sample standard deviation = s sample proportion = $\tilde{p}$

4 AIMS OF A SAMPLE

The population parameters are not known generally. Then the sample characteristics are utilised to approximately determine or estimate of the population. Thus, static is an estimate of the parameter. To what extent can we depend on the sample estimates?

The estimate of mean and standard deviation of the population is a primary purpose of all scientific experimentation. The logic of the sampling theory is the logic of *induction.* In induction, we pass from a particular (sample) to general (population). This type of generalization here is known as *statistical inference.* The conclusion in the sampling studies are based not on certainties but on probabilities.

.5 TYPES OF SAMPLING

Following types of sampling are common:

(1) Purposive sampling (2) Random sampling (3) Stratified sampling (4) Systematic sampling

.6 SAMPLING DISTRIBUTION

From a population a number of samples are drawn of equal size n. Find out the mean of each sample. The means of samples are not equal. The means with their respective frequencies are grouped. The frequency distribution so formed is known as *sampling distribution of the mean.* Similarly, sampling distribution of standard deviation we can have.

7 STANDARD ERROR (S.E.)

is the standard deviation of the sampling distribution. For assessing the difference between the expected value and observed value, standard error is used. Reciprocal of standard error is known as *precision.*

8 SAMPLING DISTRIBUTION OF MEANS FROM INFINITE POPULATION

Let the population be infinitely large and having a population mean of μ and a population variance of σ^2. If x is a random variable denoting the measurement of the characteristic, then

Expected value of x, $E(x) = \mu$

Variance of x, Var $(x) = \sigma^2$

The sample mean $\bar{x}$ is the sum of n random variables, *viz.*, $x_1, x_2, \ldots, x_n$, each being divided by n. Here, $x_1, x_2, \ldots, x_n$ are independent random variables from the infinitely large population.

$\therefore$ $E(x_1) = \mu$ and $Var\ (x_1) = \sigma^2$

$E(x_2) = \mu$ and $Var\ (x_2) = \sigma^2$ and so on

Finally $E(\bar{x}) = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right] = \frac{1}{n}Ex_1 + \frac{1}{n}E(x_2) + \ldots + \frac{1}{n}E(x_n) = \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu = \mu$

and $Var(\bar{x}) = Var\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right] = Var\left(\frac{x_1}{n}\right) + Var\left(\frac{x_2}{n}\right) + \ldots + Var\left(\frac{x_n}{n}\right)$

$$= \frac{1}{n^2}Var(x_1) + \frac{1}{n^2}Var\ (x_2) + \ldots + \frac{1}{n^2}Var(x_n) = \frac{1}{n^2}.\sigma^2 + \frac{1}{n^2}.\sigma^2 + \ldots + \frac{1}{n^2}.\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

The expected value of the sample mean is the same as population mean. The variance of the sample mean is the variance of the population divided by the sample size.

The average value of the sample tends to true population mean. If sample size *(n)* is increased then variance of $\bar{x}$, $\left(\frac{\sigma^2}{n}\right)$ gets reduced, by taking large value of n, the variance $\left(\frac{\sigma^2}{n}\right)$ of $\bar{x}$ can be made as small as desired. The standard deviation $\left(\frac{\sigma}{\sqrt{n}}\right)$ of $\bar{x}$ is also called **standard error of the mean.** It is denoted by $\sigma_{\bar{x}}$.

Sampling with Replacement

When the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. We have

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

$$Var\ (\bar{x}) = \frac{\sigma^2}{n} \text{ or } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling without replacement from Finite population

When a sample is picked up without replacement from a finite population, the probability distribution of second random variable depends on the outcome of the first pick up. n sample members do not remain independent. Now we have

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

and $$Var(\bar{x}) = \sigma_{\bar{x}}^{2} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \quad \text{or} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{\sigma}{\sqrt{n}} \text{ app} \qquad \left(\text{if } \frac{n}{N} \text{ is very small}\right)$$

Sampling from Normal Population

If $x \sim N\left(\mu,\ \sigma^2\right)$ then it follows that $\bar{x} \sim N\ (\mu,\ \frac{\sigma^2}{n})$

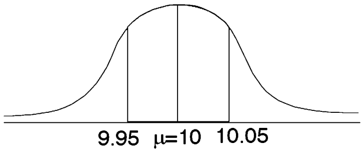
Example 1. *The diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of* 0.1 *mm. If we pick up a random sample of size 5, what is the probability that the same mean will be between 9.95 and 10.05 mm?*

Solution. Let x be a random variable representing the diameter of one component picked up at random.

Here $x \sim$ N (10, 0.01), Therefore, $\bar{x} \sim N\left(10, \frac{0.01}{5}\right)$ $\qquad \left[\bar{x} = N\left(\bar{x}, \frac{\sigma^2}{n}\right)\right]$

$$Pr\left\{9.95 \le \bar{x} \le 10.05\right\} = 2 \times Pr\left\{10 \le \bar{x} \le 10.05\right\} \qquad \left\{z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right\}$$

$$= 2 \times Pr\left\{\frac{10-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{10.05-\mu}{\frac{\sigma}{\sqrt{n}}}\right\}$$



$$= 2 \times Pr\left\{0 \le z \le \frac{10.05-10}{\frac{0.1}{\sqrt{5}}}\right\} = 2 \times Pr\{0 \le z \le 1.12\} = 2 \times 0.3686 = 0.7372 \qquad \textbf{Ans.}$$

Similar Question

A sample of size 25 is picked up at random from a population which is normally distributed with a mean 100 and a variance of 36. Calculate (*a*) $Pr\ \{\ \bar{x} \le 99\}$, (*b*) $Pr\{98 \le \bar{x} \le 100\}$

Ans. (*a*) 0.2023 (*b*) 0.4522

9 SAMPLING DISTRIBUTION OF THE VARIANCE

We use a sample statistic called the sample variance to estimate the population variance. The sample variance is usually denoted by s^2

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

10 TESTING A HYPOTHESIS

On the basis of sample information, we make certain decisions about the population. In taking such decisions we make certain assumptions. These assumptions are known as *statistical hypothesis.* These hypothesis are tested. Assuming the hypothesis correct we calculate the probability of getting the observed sample. If this probability is less than a certain assigned value, the hypothesis is to be rejected.

11 NULL HYPOTHESIS (H_0)

Null hypothesis is based for analysing the problem. Null hypothesis is the *hypothesis of no difference.* Thus, we shall persume that there is no significant difference between the observed value and expected value. Then, we shall test whether this hypothesis is satisfied by the data or not. If the hypothesis is not approved the difference is considered to be significant. If hypothesis is approved then the difference would be described as due to sampling fluctuation. Null hypothesis is denoted by H_0.

12 ERRORS

In sampling theory to draw valid inferences about the population parameter on the basis of the sample results.

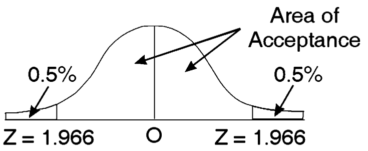
We decide to accept or to reject the lot after examining a sample from it. As such, we are liable to commit the following two types of errors.

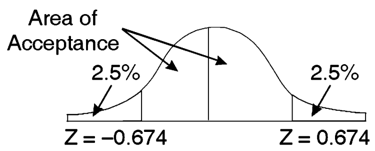
Type 1 Error. If H_0 is rejected while it should have been accepted.

Type II Error. If H_0 is accepted while it should have been rejected.

13 LEVEL OF SIGNIFICANCE

There are two critical regions which cover 5% and 1% areas of the normal curve. The shaded portions are the critical regions.





Thus, the probability of the value of the variate falling in the critical region is the level of significance. If the variate falls in the critical area, the hypothesis is to be rejected.

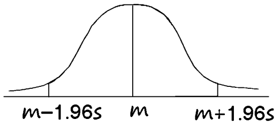
14 TEST OF SIGNIFICANCE

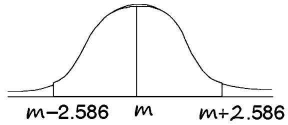
The tests which enables us to decide whether to accept or to reject the null hypothesis is called the tests of significance. If the difference between the sample values and the population values are so large (lies in critical area), it is to be rejected

15 CONFIDENCE LIMITS

$\mu - 1.96\ \sigma, \mu + 1.96\sigma$ are 95% confidence limits as the area between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 95%. If a sample statistics lies in the interval $\mu - 1.96\ \sigma$, $\mu + 1.96\sigma$, we call 95% confidence interval.

Similarly, $\mu - 2.58\sigma$, $\mu + 2.58\sigma$ is 99% confidence limits as the area between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is 99%. The numbers 1.96, 2.58 are called confidence coefficients.





16 TEST OF SIGNIFICANCE OF LARGE SAMPLES (*N* > 30)

Normal distribution is the limiting case of Binomial distribution when n is large enough. For normal distribution 5% of the items lie outside $\mu \pm 1.96\sigma$ while only 1% of the items lie outside $\mu \pm 2.586\ \sigma$.

$$z = \frac{x - \mu}{\sigma}$$

where z is the standard normal variate and x is the observed number of successes.

First we find the value of z. Test of significance depends upon the value of z.

(*i*) (*a*) If $|z| < 1.96$, difference between the observed and expected number of successes is not significant at the 5% level of significance.

(*b*) If $|z| > 1.96$, difference is significant at 5% level of significance.

(*ii*) (*a*) If $|z| < 2.58$, difference between the observed and expected number of successes is not significant at 1% level of significance.

(*b*) If $|z| > 2.58$, difference is significant at 1% level of significance.

Example 2. *A cubical die was thrown 9,000 times and 1 or 6 was obtained 3120 times. Can the deviation from expected value lie due to fluctuations of sampling?*

Solution. Let us consider the hypothesis that the die is an unbiased one and hence the probability of obtaining 1 or $6 = \frac{2}{6} = \frac{1}{3}$ *i.e.,* $p = \frac{1}{3},\ q = \frac{2}{3}$

The expected value of the number of successes = $np = 9000 \times \frac{1}{3} = 3000$

Also $\sigma = \text{S.D.} = \sqrt{npq} = \sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{2000} = 44.72$

$$3\,\sigma = 3 \times 44.72 = 134.16$$

Actual number of successes = 3120

Difference between the actual number of successes and expected number of successes

$$= 3120 - 3000 = 120 \text{ which is } < 3\,\sigma$$

Hence, the hypothesis is correct and the deviation is due to fluctuations of sampling due to randon causes. **Ans.**

17 SAMPLING DISTRIBUTION OF THE PROPORTION

A simple sample of n items is drawn from the population. It is same as a series of n independent trials with the probability p of success. The probabilities of 0, 1, 2, ..., n success are the terms in the binomial expansion of $(q + p)^n$.

Here mean = np and standard deviation = $\sqrt{npq}$.

Let us consider the proportion of successes, then

(*a*) Mean proportion of successes $= \frac{np}{n} = p$

(*b*) Standard deviation (standard error) of proportion of successes $= \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$

(*c*) Precision of the proportion of success $= \frac{1}{\text{S.E.}} = \sqrt{\frac{n}{pq}}$.

Example 3. *A group of scientific mens reported 1705 sons and 1527 daughters. Do these figures conform to the hypothesis that the sex ratio is* $\frac{1}{2}$.

Solution. The total number of observations = 1705 + 1527 = 3232

The number of sons = 1705

Therefore, the observed male ratio $= \frac{1705}{3232} = 0.5275$

On the given hypothesis the male ratio = 0.5000

Thus, the difference between the observed ratio and theoretical ratio

$$= 0.5275 - 0.5000$$

$$= 0.0275$$

The standard deviation of the proportion $= s = \sqrt{\dfrac{pq}{n}} = \sqrt{\dfrac{\dfrac{1}{2} \times \dfrac{1}{2}}{3232}} = 0.0088$

The difference is more than 3 times of standard deviation.

Hence, it can be definitely said that the figures given do not conform to the given hypothesis.

18 ESTIMATION OF THE PARAMETERS OF THE POPULATION

The mean, standard deviation etc. of the population are known as parameters. They are denoted by μ and σ. Their estimates are based on the sample values. The mean and standard deviation of a sample are denoted by $\bar{x}$ and s respectively. Thus, a static is an estimate of the parameter. There are two types of estimates.

(*i*) *Point estimation:* An estimate of a population parameter given by a single number is called a point estimation of the parameter. For example,

$$(\text{S.D.})^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

(*ii*) *Interval estimation:* An interval in which population parameter may be expected to lie with a given degree of confidence. The intervals are

(*i*) $\bar{x} - \sigma_s$ to $\bar{x} + \sigma_s$ (68.27% confidence level)

(*ii*) $\bar{x} - 2\sigma_s$ to $\bar{x} + 2\sigma_s$ (95.45% confidence level)

(*iii*) $\bar{x} - 3\ \sigma_s$ to $\bar{x} + 3\ \sigma_s$ (99.73% confidence level)

$\bar{x}$ and σ_s are the mean and S.D. of the sample.

Similarly, $\bar{x} \pm 1.96\ \sigma_s$ and $\bar{x} \pm 2.58\ \sigma_s$ are 95% and 99% confidence of limits for μ.

$\bar{x} \pm 1.96\ \dfrac{\sigma}{\sqrt{n}}$ and $\bar{x} \pm 2.58\ \dfrac{\sigma}{\sqrt{n}}$ are also the intervals as $\sigma_s = \dfrac{\sigma}{\sqrt{n}}$.

19 COMPARISON OF LARGE SAMPLES

Let two large samples of size n_1, n_2 be drawn from two populations of proportions of attributes A's as P_1, P_2 respectively.

(*i*) *Hypothesis:* As regards the attribute A, the two populations are similar. On combining the two samples

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

where p is the common proportion of attributes.

Let e_1, e_2 be the standard errors in the two samples, then

$$e_1^2 = \frac{pq}{n_1} \text{ and } e_2^2 = \frac{pq}{n_2}$$

If e be the standard error of the combined samples, then

$$e = P_1^2 + P_2^2 = \frac{pq}{n_1} + \frac{pq}{n_2} = pq\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

$$z = \frac{P_1 - P_2}{e}$$

1. If $z > 3$, the difference between P_1 and P_2 is significant.
2. If $z < 2$, the difference may be due to fluctuations of sampling.
3. If $2 < z < 3$, the difference is significant at 5% level of significance.

(*ii*) *Hypothesis.* In the two populations, the proportions of attribute A are not the same, then standard error e of the difference $p_1 - p_2$ is

$$e^2 = p_1 + p_2$$

$$= \frac{P_1 - q_1}{n_1} + \frac{P_2 - q_2}{n_2}, z = \frac{P_1 - P_2}{e} < 3,$$

difference is due to fluctuations of samples.

Example 4. *In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men.*

Solution. $n_1 = 600$ men, Number of smokers= 450, $P_1 = \frac{450}{600} = 0.75$

$n_2 = 900$ men, Number of smokers= 450, $P_2 = \frac{450}{900} = 0.5$

$$P = \frac{n_1P_1 + n_2P_2}{n_1 + n_2} = \frac{600 \times 0.75 + 900 \times 0.5}{600 + 900} = \frac{900}{1500} = 0.60$$

$$q = 1 - P = 1 - 0.6 = 0.4$$

$$e^2 = P_1^2 + P_2^2 = Pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$e^2 = 0.6 \times 0.4\left(\frac{1}{600} + \frac{1}{900}\right) = 0.000667$$

$$e = 0.02582$$

$$z = \frac{P_1 - P_2}{e} = \frac{0.75 - 0.50}{0.02582} = 9.682$$

$z > 3$ so that the difference is significant. **Ans.**

Example 5. *One type of aircraft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a significant difference in the two types of aircrafts so far as engine defects are concerned.*

Solution. $n_1 = 100$ flights, Number of troubled flights = 5, $P_1 = \frac{5}{100} = \frac{1}{20}$

$n_2 = 200$ flights, Number of troubled flights = 7, $P_2 = \frac{7}{200}$

$$e^2 = \frac{P_1q_1}{n_1} + \frac{P_2q_2}{n_2} = \frac{0.05 \times 0.95}{100} + \frac{0.035 \times 0.965}{200}$$

$$= 0.000475 + 0.0001689 = 0.0006439$$

$$e = 0.0254$$

$$z = \frac{0.05 - 0.035}{0.0254} = 0.59$$

$z < 1$, Difference is not significant. **Ans.**

.20 THE t-DISTRIBUTION (FOR SMALL SAMPLE)

The students distribution is used to test the significance of

(*i*) The mean of a small sample.

(*ii*) The difference between the means of two small samples or to compare two small samples.

(*iii*) The correlation coefficient.

Let $x_1, x_2, x_3, ..., x_n$, be the members of random sample drawn from a normal population with mean μ. If $\bar{x}$ be the mean of the sample then

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \text{ where } s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Example 6. *A machine which produces mica insulating washers for use in electric device to turn out washers having a thickness of 10 mm. A sample of 10 washers has an average thickness 9.52 mm with a standard deviation of* 0.6 *mm. Find out t.*

Solution. $\bar{x} = 9.52, M = 10, S' = 0.6, n = 10$

$$t = \frac{\bar{x}-M}{\frac{s}{\sqrt{n}}} = \frac{9.52-10}{\frac{0.6}{\sqrt{10}}} = -\frac{0.48\sqrt{10}}{0.6} = -\frac{4}{5}\sqrt{10}$$

$$= -0.8 \times 3.16 = -2.528$$ **Ans.**

.21 WORKING RULE

To calculate significance of sample mean at 5% level.

Calculate $t = \frac{\bar{x}-\mu}{s}\sqrt{n}$ and compare it to the value of t with (n- 1) degrees of freedom at 5% level, obtained from the table. Let this tabulated value of t be t_1.

If $t < t_1$, then we accept the hypothesis *i.e.*, we say that the sample is drawn from the population.

If $t > t_1$, we compare it with the tabulated value of t at 1% level of significance for (n – 1) degrees of freedom. Denote it by t_2. If $t_1 < \text{t} < t_2$ then we say that the value of t is significant.

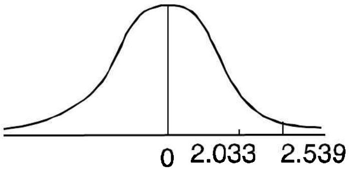
If $t > t_1$, we reject the hypothesis and the sample is not drawn from the population.

Example 7. *A manufacturer intends that his electric bulbs have a life of 1000 hours. He tests a sample of 20 bulbs, drawn at random from a batch and discovers that* the mean life of the sample bulbs is 990 hours with a *S.D of 22 hours. Does this signify that the batch is not up to the standard?*

[**Given:** *The table value of t at 1% level is significance with 19 degrees of freedom is 2.539*]

Solution. $\bar{x} = 990, \sigma = 22, x = 1000$

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{990-1000}{\frac{22}{\sqrt{20}}} = -\frac{10\sqrt{20}}{22} = -\frac{22.36}{11} = -2.033$$



Since the calculated value of t (2.032) is less than the value of t (2.539) from the table. Hence, it is not correct to say that this batch is not upto this standard. **Ans.**

Example 8. *Ten individuals are chosen at random from a population and their heights are found to be in inches 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Discuss the suggestion that the Mean height of universe is 65.*

For 9 degree of freedom t at 5% level of significance = 2.262.

Solution.

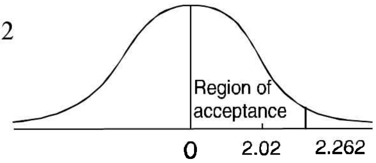
x	$x-67$	$(x-67)^2$
63	–4	16
63	–4	16
64	–3	9
65	–2	4
66	–1	1
69	+2	4
69	+2	4
70	+3	9
70	+3	9
71	+4	16
$\sum x = 670$		$\sum (x-\bar{x})^2 = 88$

$$\bar{x} = \frac{\sum x}{n} = \frac{670}{10} = 67,$$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{88}{9}} = 3.13$$

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{67-65}{\frac{3.13}{\sqrt{10}}} = \frac{2\sqrt{10}}{3.13} =2.02$$

$2.02 < 2.262$



Calculated value of t (2.02) is less than the table value of t (2.262). The hypothesis is accepted the mean height of universe is 65 inches. **Ans.**

Example 9. *The mean life time of sample of 100 fluorescent light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. The company claims that the average life of the bulbs produced by it is 1600 hours. Using the level of significance of 0.05, is the claim acceptable?*

Solution. $\bar{x} = 1570$, $s = 120$, $n = 100$, $\mu = 1600$

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{1570-1600}{\frac{120}{\sqrt{100}}} = \frac{1570-1600}{12} = 2.5$$

At 0.05 the level of significance, $t = 1.96$

Calculated value of t > Table value of t.

$2.5 > 1.96$

Hence the claim is to be rejected. **Ans.**

Example 10. *A sample of 6 persons in an office revealed an average daily smoking of 10, 12, 8, 9, 16, 5 cigarettes. The average level of smoking in the whole office has to be estimated at 90% level of confidence.*

$t = 2.015$ for 5 degree of freedom

Solution.

x	$x-10$	$(x-10)^2$
10	0	0
12	2	4
8	–2	4
9	–1	1
16	+6	36
5	–5	25
Total	0	$\sum (x-10)^2 = 70$

$$\text{Mean} = a + \frac{\Sigma fd}{\Sigma f} = 10 + \frac{0}{6} = 10$$

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{70}{5}} = 3.74$$

At 90% level of confidence, $t = \pm\, 2.015$.

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \quad \Rightarrow \quad \pm\, 2.015 = \frac{10-\mu}{\frac{3.74}{\sqrt{6}}}$$

$$\Rightarrow \quad \mu = 2.015 \times \frac{3.74}{\sqrt{6}} + 10 = 6.92, 13.08$$ **Ans.**

Example 11. *A fertiliser mixing machine is set to give 12 kg of nitrate for quintal bag of fertiliser: Ten 100 kg bags are examined The percentages of nitrate per bag are as follows:*

11, 14, 13, 12, 13, 12, 13, 14, 11, 12

Is there any reason to believe that the machine is defective? Value of t for 9 degrees of freedom is 2.262.

Solution. The calculation of $\bar{x}$ and s is given in the following table:

x	$d = x - 12$	d^2
11	–1	1
14	2	4
13	1	1
12	0	0
13	1	1
12	0	0
13	1	1
14	2	4
11	–1	1
12	0	0
$\sum x = 125$	$\sum d = 5$	$\sum d^2 = 13$

$$\mu = 12 \text{ kg}, \; n = 10, \bar{x} = \frac{\sum x}{n} = \frac{125}{10} = 12.5$$

$$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{13}{10} - \left(\frac{5}{10}\right)^2 = \frac{13}{10} - \frac{1}{4} = \frac{21}{20} = \frac{105}{100}$$

$$s = 1.024$$

Value of t for 9 degrees of freedom = 2.262

Also $t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{12.5 - 12}{1.024}\sqrt{10} = 1.54$

Since the value of t is less than 2.262, there in no reason to believe that machine is defective. **Ans.**

Example 12. *A random sample of size 16 values from a normal population showed a mean of 53 and a sum of squares of deviation from the mean equals to 150. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.*

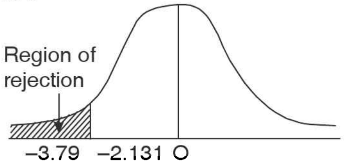
$\gamma = 15, \; \alpha = 0.05, t = 2.131$

$\alpha = 0.01, t = 2.947$

Solution. $\mu = 56, n = 16, \; \bar{x} = 53, \; \sum(x - \bar{x})^2 = 150$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{150}{15} = 10$$

$$s = \sqrt{10}$$



$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{53 - 56}{\frac{\sqrt{10}}{\sqrt{16}}} = \frac{-3 \times 4}{\sqrt{10}} = -3.79$$

$$t = 3.79$$

When $\alpha = 0.5$ then $3.79 > 2.131$

When $\alpha = 0.01$ then $3.79 > 2.947$

Thus, the sample cannot be regarded as taken from the population. **Ans.**

22 TESTING FOR DIFFERENCE BETWEEN MEANS OF TWO SMALL SAMPLES

Let the mean and variance of the first population be μ_1 and σ_1^2 and $\mu_2 . \; \sigma_2^2$ be the mean and variance of the second population.

Let $\bar{x}_1$ be the mean of small sample of size n_1 from first population and $\bar{x}_2$ the mean of a sample of size n_2 from second population.

We know that

$$E(\bar{x}_1) = \mu_1 \text{ and } Var\ (\bar{x}_1) = \frac{\sigma_1^2}{n_1}$$

$$E(\bar{x}_2) = \mu_2 \text{ and } Var\,(\bar{x}_2) = \frac{\sigma_2{}^2}{n_2}$$

If the samples are independent, then $(\bar{x}_1)$ and $(\bar{x}_2)$ are also independent.

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2 \text{ and } Var\,(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}$$

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma^2{}_1}{n_1}\right) \text{and}\, \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma^2{}_2}{n_2}\right) \text{ then } (\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}\right)$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1{}^2}{n_1} + \dfrac{\sigma_2{}^2}{n_2}}}$$

If the population is the same then

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \qquad (\mu_1 - \mu_2 = \mu_1 - \mu_1 = 0)$$

Example 13. *Two independent samples of 8 and 7 items respectively had the following values of the variable (weight in ounces):*

Sample 1: 9 11 13 11 15 9 12 14

Sample 2: 10 12 10 14 9 8 10

Is the difference between the means of the sample significant?

[*Given for $V = 13$, $t_{0.05} = 2.16$*]

Solution.

Assumed mean of x = 12, Assumed mean of y = 10

x	$(x–12)$	$(x–12)^2$	y	$(y–10)$	$(y–10)^2$
9	–3	9	10	0	0
11	–1	1	12	2	4
13	1	1	10	0	0
11	–1	1	14	4	16
15	3	9	9	–1	1
9	–3	9	8	–2	4
12	0	0	10	0	0
14	2	4	–	–	–
94	–2	34	73	3	25

$$\bar{x} = \frac{\sum x}{n} = \frac{94}{8} = 11.75$$

$$\sigma_x{}^2 = \frac{\Sigma(x-12)^2}{n} - \left(\frac{\Sigma(x-12)}{n}\right)^2 = \frac{34}{8} - \left(\frac{-2}{8}\right)^2 = 4.1875$$

$$\bar{y} = \frac{\sum y}{n} = \frac{73}{7} = 10.43$$

$$\sigma_y{}^2 = \frac{\sum(y-10)^2}{n} - \left[\frac{\Sigma(y-10)}{n}\right]^2 = \frac{25}{7} - \left(\frac{3}{7}\right)^2 = 3.388$$

$$s = \sqrt{\frac{(x-\bar{x})^2 + \sum(y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{34+25}{8+7-2}} = \sqrt{\frac{59}{13}} = \sqrt{4.54} = 2.13$$

$$t=\frac{\overline{x}-\overline{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{11.75-10.43}{2.13\sqrt{\frac{1}{8}+\frac{1}{7}}}=\frac{1.32}{2.13\sqrt{0.268}}=\frac{1.32}{2.13\times 0.518}=\frac{1.32}{1.103}=1.2$$

Thus, 5% value of t for 13 degree of freedom is given to be 2.16. Since calculated value of t is 1.2 is less than 2.16, the difference between the means of samples is not significant. **Ans.**

EXERCISE ¯¯.1

1. A random sample of six steel beams has mean compressive strength of 58.392 psi (pounds per square inch) with a standard deviation of s = 648 psi. Test the null hypothesis $H_0 = \mu$ = 58,000 psi against the alternative hypothesis H_1: $\mu > 58{,}000$ psi at 5% level of significance (value for t at 5 degree of freedom and 5% significance level is 2.0157). Here μ denotes the population mean. *(A.M.I.E., Summer 2000)*
2. A certain cubical the was thrown 96 times and shows 2 upwards 184 times. Is the the biased?

   Ans. die is biased.
3. In a sample of 100 residents of a colony 60 are found to be wheat eaters and 40 rice eaters. Can we assume that both food articles are equally popular?
4. Out of 400 children, 150 are found to be under weight. Assuming the conditions of simple sampling, estimate the percentage of children who are underweight in, and assign limits within which the per-centage probably lies.

   Ans. 37.5% approx. Limits = 37.5 ± 3 (2.4)
5. 500 eggs are taken at random from a large consignment, and 50 are found to be bad. Estimate the percentage of bad eggs in the consignment and assign limits within which the percentage probably lies.

   Ans. 10%, 10 ± 3.9
6. A machine puts out 16 imprefect articles in a sample of 500. After the machine is repaired, puts out 3 imprefect articles in a batch of 100. Has the machine been improved?

   Ans. The machine has not been improved.
7. In a city A, 20% of a random sample of 900 school boys had a certain slight physical defect. In another city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant?

   Ans. z = 0.37, Difference between proportions is significant.
8. In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

   Ans. z = 2.5, not hidden at 5% level of significance.
9. One thousand articles from a factory are examined and found to be three percent defective. Fifteen hundred similar articles from a second factory are found to be only 2 percent defective. Can it reasonably be concluded that the product of the first factory is inferior to the second?

   Ans. It cannot be reasonable concluded that the product of the first factory is inferior to that of the second.
10. A manufacturing company claims 90% assurance that the capacitors manufactured by them will show a tolerance of better than 5%. The capacitors are packaged and sold in lots of 10. Show that about 26% of his customers ought to complain that capacitors do not reach the specified standard.

.23 CHI SQUARE TEST

The Chi-square distribution is one of the most extensively used distribution function in statistics. It was first discovered by Helmert in 1875 and later on Karl Pearson's in 1900.

.24 CHI-SQUARE VARIATES

The square of a standard normal variate is known as Chi-square variate (χ^2) with one degree function :

$$z = \frac{x-\mu}{\sigma} \text{ is a normal variate.}$$

$$\text{Hence } \left(\frac{x-\mu}{\alpha}\right)^2 \text{ is a Chi-square variate.}$$

If x be a normaly distributed variate and $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a random sample of n-values from this population then $w = x_1^2 + x_2^2 + \ldots\ldots. + x_n^2$ has χ^2 distribution with n-degree of freedom.

25 CONDITIONS FOR CHI-SQUARE TEST

There are some conditions which are necessary for Chi square test.

1. The sample under study must be large and may be total of cell frequency should not be less than 50.
2. The member of the cells should be independent.
3. The cell frequency of each cell should be greater than 5. If any cell has frequency less than 5 then it should be combined with the next or preceding cell until the total frequency exceeds 5.
4. If there are any constraint on the cell frequencies they should be linear
   i.e.; $\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \ldots\ldots.. + \alpha_i x_i + \ldots\ldots\ldots\ldots + \alpha_n x_n = \lambda$

Note: Cell frequency should not involve any logarithmic, exponential or trigonometric relation.

.26 CHI-SQUARE (χ^2) IS USED AS:

(1) Test of independence (2) Test of goodness of fit

(3) To test if the hypothetical value of the population variate is σ^2

(4) To test the homogeneity of independent estimate of the population variance.

We shall mainly use the first two test

27 CHI-SQUARE TEST OF GOODNESS OF FIT

This test is used to test significance of the discrepancy between theory and experiment. It helps us to find if the deviation of the experiment from the theory is just by chance or it is due to the inadequacy of the theory to fit the observed data.

The theoretical frequencies for various classes are calculated from the assumption of the population. The significant deviation between the observed and theoretical frequencies is tested by means of this test.

χ is calculated by means of the following formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \text{and} \quad \Sigma O_i = \Sigma E_i = N$$

where O_i is the observed frequency E_i is the expected (Theoretical) frequency of the cell.

28 WORKING RULE TO CALCULATE χ^2:

Step 1. Calculate the expected frequencies.

Step 2. Calculate the difference between each observed frequency O_i and the corresponding expected frequency E_i for each class i.e.; to find $O_i - E_i$

Step 3. Square the difference obtained in step 2 for each value i.e.; Calculate $(O_i - E_i)^2$.

Step 4. Divide $(O_i - E_i)^2$ by the expected frequency E to get $\dfrac{(O_i - E_i)^2}{E_i}$

Step 5. Add all these quotients obtained in step 4. Then $\chi^2 = \dfrac{\sum\limits_{i=1}^{n} (O_i - E_i)^2}{E_i}$

It is to be noted

(1) The value of χ^2 is always positive. (2) χ^2 will be zero if each pair is zero.

(3) The value of χ^2 lies between 0 and ∞.

29 DEGREE OF FREEDOM

Case I. If the data is given in the form of a series of variables in a row or column then the degree of freedom = (No. of items in the series) – 1

Case 2. When the number of frequencies are put in cells in a contingency table.

The degree of freedom = (R – 1) (C – 1)

where R is number of rows and C is the number of columns.

Example 14. *A survey of 320 families with 5 children is given below :*

No. of boys	*5*	*4*	*5*	*2*	*1*	*0*	*Total*
No. of girls	*0*	*1*	*2*	*3*	*4*	*5*	
No. of families	*14*	*56*	*110*	*88*	*40*	*12*	*320*

Is this result consistent with hypothesis i.e.; the male and female birth are equally possible.

Solution. Null Hypothesis H_0.

(1) Male and Female birth are equally probable.

Alternate Hypothesis $\mathbf{H_1}$: Male and female birth are not equally probable.

Calculation of expected frequencies $(q + p)^n$

Probability of female birth $= p = \dfrac{1}{2}$

Probability of male birth $= q = \dfrac{1}{2}$

$$(q + p)^n = q^n + {}^nC_1\, p\, q^{n-1} + {}^nC_2\, p^2\, q^{n-2} + {}^nC_3\, p^3\, q^{n-3} + \ldots\ldots. + p^n$$

$$\left(\frac{1}{2}+\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^5 + 5\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 + 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 + 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^5$$

$$\text{No. of girls} = 320\left[\frac{1}{32}+\frac{5}{32}+\frac{10}{32}+\frac{10}{32}+\frac{5}{32}+\frac{1}{32}\right]$$

$$= 320 \times \frac{1}{32} + 320 \times \frac{5}{32} + 320 \times \frac{10}{32} + 320 \times \frac{10}{32} + 320 \times \frac{5}{32} + 320 \times \frac{1}{32}$$

$$= 10 + 50 + 100 + 100 + 50 + 10$$

These are the expected frequencies of the female births.

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14 56 110 88 40 12	10 50 100 100 50 10	4 6 10 – 12 – 10 2	16 36 100 144 100 4	1.60 0.72 1.00 1.44 2.00 0.40
			Total	7.16

Level of significance Let $\alpha = 0.05$

Critical value. The table value of χ^2 at $\alpha = 0.05$ for $(6-1)(2-1) = 5$ degree of freedom is 11.07

Decision Since the calculated value of χ^2 (7.16) < Table value of χ^2 at level of significance 0.05 for $5df = 11.07$

Hence, the null hypothesis is accepted i.e.; the male and female birth is equally probable.

Example 15. *The table below give the number of air craft accidents that occured during the various days of the week. Test whether the accidents are uniformly distributed over the week.*

Days	*Mon.*	*Tue.*	*Wed.*	*Thu.*	*Fri*	*Sat*	*Sun*	*Total no. accidents*
No. of accidents	*14*	*18*	*12*	*11*	*15*	*14*	*14*	*98*

Solution. H_0 : Null Hypothesis : The accidents are uniformly distributed over the week.

The expected frequencies of the accidents on each day = $\frac{98}{7} = 14$

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14 18 12 11 15 14 14	14 14 14 14 14 14 14	0 4 – 2 – 3 1 0 0	0 16 4 9 1 0 0	0 1.14 0.29 0.64 0.07 0 0
98			Total	2.14

Level of significance :

Let $\alpha = 0.05$

Critical value: The table value of χ^2 at $\alpha = 0.05$ is for $(7-1)(2-1)$ i.e.; 6 degree is $\chi^2 = 12.592$

Since the calculated value of χ^2 (7.16) < Table value of χ^2 at level of significance 0.05 for six degree = 12.592

Hence, the null Hypothesis is accepted i.e.; the air craft accident are uniformly distributed over the week. **Ans.**

$$\text{Expected Frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Example 16. *In an investigation into the health and nutrition of two groups of children of different social status the following results are obtained.*

Social Status / *Health*	*Poor*	*Rich*	*Total*
Below Normal	*130*	*20*	*150*
Normal	*102*	*108*	*210*
Above normal	*24*	*96*	*120*
Total	*256*	*224*	*480*

Discuss the relation between the health and their social status.

Solution. H_0. Null Hypothesis : There is no association between health and social status.

H_1. Alternate Hypothesis: there an no association between health and social status.

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Social Status / *Health*	*Poor*	*Rich*	*Total*
Below Normal	$\frac{256 \times 150}{480} = 80$	$\frac{224 \times 150}{480} = 70$	150
Normal	$\frac{256 \times 210}{480} = 112$	$\frac{224 \times 210}{480} = 98$	210
Above normal	$\frac{256 \times 120}{480} = 64$	$\frac{224 \times 120}{480} = 56$	120
Total	256	224	480

Total number of observed frequencies = Total number of expected frequencies = 480

Degree of freedom = (3 – 1) (2 – 1) = 2

Level of significance, take $\alpha = 0.5$

Critical value the table value of χ^2 at $\alpha = 0.05$ for degree of freedom 2 is 5.99.

Decision. Since the calculated value of χ^2 (122.44) > table value of χ^2 at level of significance 0.05 for two 2*d.f.* = 5.991.

Hence, the null hypothesis is rejected i.e., social status and health are associated (Dependent). **Ans.**

Calculation of Chi-square

Observed value (O)	Expected Value (E)	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
130	80	50	2500	$\frac{2500}{80} = 31.25$
102	112	– 10	100	$\frac{100}{112} = 0.89$
24	64	– 40	1600	$\frac{1600}{64} = 25$
20	70	– 50	2500	$\frac{2500}{70} = 35.71$
108	98	10	100	$\frac{100}{98} = 1.02$
96	56	40	1600	$\frac{1600}{56} = 28.57$
			Total	122.44

Example 17. *The I.Q. and economic condition of home of 1000 students of an engineering college, Delhi were noted as given in the table :*

I.Q. / *Economic con.*	*High*	*Low*	*Total*
Rich	*100*	*300*	*400*
Poor	*350*	*250*	*600*
Total	*450*	*550*	*1,000*

Find out whether there is any association between economic condition at home and I.Q. of the students.

Given for 1 *d.f.*, χ^2 *at the level of significance 0.05 is 3.84.*

Solution.

Null Hypothesis H_0: There is no association between economic condition at home and I.Q.

Alternative hypothesis H_1: There is an association between economic condition at home and *I . Q.*

$$\text{Expected frequency E } = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

I.Q. / *Economic cond.*	*High*	*Low*	*Total*
Rich	$\frac{400 \times 450}{1000} = 180$	$\frac{400 \times 550}{1000} = 220$	400
Poor	$\frac{600 \times 450}{1000} = 270$	$\frac{600 \times 550}{1000} = 330$	600
Total	450	550	1000

Calculation of Chi-square

Observed value (O)	Expected Value (E)	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
100	180	– 80	6400	35.5
350	270	80	6400	23.7
300	220	80	6400	29.1
250	330	– 80	6400	19.4
			Total	107.7

Degree of freedom = (R – 1) (C – 1) = (2 – 1) (2 – 1) = 1 given for *d.f.* = 1,
χ^2 at the level of significance 0.05 = 3.84

Decision. The calculated value of χ^2 is greater than Table value of χ^2. Hence, the hypothesis is rejected and the alternative hypothesis is accepted.

Hence, there is an association between economic condition at home and I.Q. **Ans.**

Example 18. *To test the effectiveness of inoculation against cholera, the following table was obtained.*

	Attached	*Not attached*	*Total*
Inoculated	30	160	190
Notinoculated	140	460	600
Total	170	620	790

(*The figures represent the number of persons*)

Use χ^2 - test to defend or refute the statement. The inoculation prevents attack from cholera. (*U.P. III Semester Dec. 2009*)

Solution. **H_0 Null Hypothesis:** No inoculation prevents attack from cholera.

H_1 Alternate Hypothesis:

The inoculation prevents attack from cholera.

Expected frequency $E = \frac{\text{Row total} \times \text{Column Total}}{\text{Grand total}}$

	Attacked	Not Attacked	Total
Inoculated	$\frac{170 \times 190}{790} = 40.9$	$\frac{620 \times 190}{790} = 149.1$	190
Not inoculated	$\frac{170 \times 600}{790} = 129.1$	$\frac{620 \times 600}{790} = 470.9$	600
Total	170	620	790

Total number of observed frequencies
= Total number of expected frequencies = 790

Calculation of Chi-square

Observed value (O)	Expected value (E)	(O – E)	$(O - E)^2$	$\frac{(O-E)^2}{E}$
30	40.9	– 10.9	118.81	2.904
140	129.1	10.9	118.81	0.920
160	149.1	10.9	118.81	0.797
460	470.9	– 10.9	118.81	0.252
			Total	4.873

Degree of freedom = (R – 1) (C – 1) = (2 – 1) (2 – 1) = 1

The critical value of the table value of χ^2 at $\alpha = 0.05$ for 1d.f. is 3.841.

Decision: Since the calculated value of χ^2 (4.873) is greater than the table value (3.841). Thus, the hypothesis is rejected and the alternative hypothesis is accepted. Hence, the inoculation prevents attack from cholera. **Ans.**

EXERCISE ¯¯.2

1. In an experiment immunization of cattle from a disease, the following results are obtain:

	Affected	Unaffected	Total
Inoculated	12	28	40
No Inoculated	13	7	20
Total	25	35	60

Examine the effect of vaccin in controlling the incidence of the disease. **Ans.** Not independent

2. In the contigency table given below use Chi-square test to test for independence of hair colour and eye colour of persons:

Hair colour Eye colour	Light	Dark	Total
Blue	26	9	35
Brown	7	18	25
Total	33	27	60

Ans. Hair colour and eye colour are associated

3. A survey amongst women was conducted to study the family life. The observations are as follows:

Family life

	Happy	Not Happy	Total
Educated	70	30	100
Not educated	60	40	100
Total	130	70	200

Test whether there is any association between family life and education.

Ans. there is no association between family life and education.

4. A certain drug was administrated to 500 people out of a total of 800 included in a sample to test its efficiency against typhoid, the results are given below :

	Typhoid	No Typhoid	Total
Drug	200	300	500
No Drug	280	20	300
	48	320	800

On the basis of the data, can we say that drug is effective in preventing Typhoid.

Ans. $\chi^2 = 222.22$ drug is effective

5. The following table gives the number of person's whose eye sight is attacked and an injection of macugen is injected by Prof. Atul.

	Eye sight Improved	Eye sight not improved	total
Injected	216	145	361
Not injected	105	234	339
Total	321	379	700

Do you think macugen injection can improve the eye sight. **Ans.** By injection, eye sight has improved

6. From the table given below, whether the colour of the sons eyes is associated with that of father's eye.

Eyes colour in sons

		Not light	light	
Eyes colour	Not light	230	148	378
in fathers	light	151	471	622
		381	619	1000

There is an association between the colour of eyes of sons and colours of eye's of fathers.

Ans. Null hypothesis in rejected