# 12. Sampling and Test of Significance

Sampling is a method of ***collection of data.***

***Sampling*** is *the process of getting a representative fraction of a population.*

***Sample*** is *the representative fraction of a population.*

In sampling method, a small group is taken from a large population. This small group is the ***sample.*** Analysis of the sample gives an ***idea*** of the population.

When the population is very large or infinite, sampling is the suitable method of data collection.

One rice is tested from a pot of boiling rice to arrive at a conclusion.

In an electric bulb factory, the bulbs are tested at intervals how long they will burn. If all are tested there is nothing left for selling.

One grape is tasted before buying a bunch of grapes.

The oxygen content of pond water can be found out by titrating just 100 ml of water.

The length of leaves of a neem tree can be calculated by measuring just 10 leaves.

The food habits of all students of a college can be understood by observing only about 100 students.

There are two types of sampling, namely

1. *Random sampling.*
2. *Non-random sampling.*

## 1. Random Sampling

Random sampling is a method of ***collection of data***

In random sampling, a small group is selected from a large population ***without any aim*** or ***predetermination.*** The small group selected is called a ***sample.***

In this method, each item of the population has an equal and independent chance of being included in the sample.

The random sample is selected by ***lottery method.***

Each individual is given a number. The numbers are written on pieces of papers. The papers are put in a box. About 100 papers are picked out. These 100 individuals form a random sample. The analysis of 100 individuals gives an idea of the entire population.

Random sampling is of 3 types, namely

1. *Simple random sampling*
2. *Stratified random sampling*
3. *Cluster random sampling*

In ***simple random sampling,*** each individual of the population has an equal chance of being included in the sample. In this method, the sample is selected by ***lottery method.***

In ***stratified random sampling,*** the population is divided into groups or strata on the basis of certain characteristics.

Then the samples are selected by simple random sampling.

For example, we want to select a sample of 100 students from a college population of 1000 students

nsisting of 700 girls and 300 boys. The whole college pulation should be divided into two strata. One with 700 ls and other with 300 boys. Now by simple random mpling method select 70 girls from total of 700 girls and boys from the total of 300 boys.

In ***cluster sampling***, the whole population is divided to a number of relatively ***small clusters*** or ***groups***. Then me of the clusters are randomly selected. For example, if e want to survey the general health of the college student a state consisting of 5000 colleges. Here we consider ach college as a ***cluster***. Now we can randomly select veral college and conduct the survey.

### 2. Non-Random Sampling

Non-random sampling is a method of ***collection of ata.*** In this method, *a sample is collected from a large opulation based on the convenience, judgement and onsideration of the investigator.*

In non-random sampling, each individual does not get chance of being included in the sample.

Eg. *If 20 students are selected from a college of 1000 students, the investigator selects 20 representatives.*

### Advantages of Sampling

1. Sampling is an economical method of data collection.
2. It saves time, expenditure and energy.
3. It is reliable.

### Disadvantages

1. Sampling needs skill.
2. It needs experts.
3. All the individuals are not represented.

Let a large random sample be taken from a parent population whose mean is M and standard deviation is σ*. Let the mean of the sample be $\bar{x}$. Then we can test the randomness of the sample by taking the hypothesis "The sample is a random one". In other words, this is a random sample of the population whose mean is M and the difference $|\bar{x}-M|$ is a random error due to sampling.• Can the statistician categorically state that the sample is a random one or not? He cannot. He can either accept or reject the hypothesis with a very great amount of assurance or confidence. This amount of confidence is determined by the probability of the occurrence of the difference between the sample mean and the population mean. i.e., $(\bar{x}-M)$. We know that 95% of the sample mean will occur in the interval from M-1.96σM to M+1.96σM and 99% will occur in the interval from M-2.58σM to M+2.58σM according to the normal law. But to be on the safer side we can take it that in the interval from M-2σM to M+2σM 95% of the sample means fall and in the interval from M-3σM to M+3σM 99% of the sample means fall. Converting this statement into a statement of chance or probability we can say that $\bar{x}$ falls in the first interval in 19 out of 20 cases (95 out of 100) and $\bar{x}$ will fall in the wider second interval in 99 out of 100 cases. We can also state the same in a different way. The probability of $\bar{x}$ falling in the interval from M-2σM to M+2σM is 0.95 and in the interval from M-3σM to M+3σM is 0.99. If now the value $|\bar{x}-M|$ is less than 2σM we see that it is likely to be one of the 95% of cases. So we have no reason to doubt the hypothesis and so we say that the hypothesis is accepted at 5% level of

---

* σ = *Sigma*

• $\bar{x}$ = *x bar*

significance. If $|\bar{x}-M| > 2\sigma M$ but $< 3\sigma M$ our statement becomes less confident. Still it might have arisen due to fluctuations of sampling. Though we reject the hypothesis at 5% level of significance we accept the hypothesis at 1% level of significance. If on the contrary $|\bar{x}-M| > 3\sigma M$ it is highly unlikely that this sample is random as it is one of the 1% of cases (actually it is one of 3 cases in a 1000) which is a comparatively rare event. The difference is said to be highly significant. In other words, the hypothesis is rejected at 1% level of significance and we conclude that bias in the sample is strongly indicated.

Thus, if $\frac{|\bar{x}-M|}{\sigma M} < 2$, sample is random at 5% level of significance

if $\frac{|\bar{x}-M|}{\sigma M} > 2$ but $< 3$, sample is random at 1% level of significance

and if $\frac{|\bar{x}-M|}{\sigma M} > 3$ difference is significant at 1% level we strongly suspect bias in the sample.

This ratio $\frac{|\bar{x}-M|}{\sigma M}$ is called the ***critical Ratio*** (C.R).

## Null Hypothesis

The student might have noticed that the statistician takes care at every stage in his statements. He is testing the hypothesis and gives a probability statement. It may be that in a sample not selected by random methods the critical ratio is satisfied. It does not mean that the sample is random. On the contrary a perfectly random sample may as an extreme case give a critical ratio greater than 3. We cannot at once decide that it is not random. Hence the statistician is guarded in his statement and only expresses his statement as

a probability. If based on his statement one takes a decisio that either the sample is random or not and finds later that h has made a mistake the statistician is not to blame. But or need not fear taking decisions on the basis of the statistician statement as he is likely to be correct in over 99% of h decisions and the other (less than 1%) cases need not com pel him to loss his confidence in the statistician.

The hypothesis we have assumed is said to be the nu hypothesis. It means that the true difference between th mean of the sample and the mean of the population is nil an whatever we notice is only a random sampling error. The te does not completely prove or disprove the hypothesis. Henc based on the level of confidence fostered by the statistician statement it is reasonable to accept or reject the hypothesis The method of assuming a null hypothesis is more scien tific. Hence it is free from prejudice and bias. It is denote by *Ho*.

## Alternative Hypothesis

It is a statement about the population parameter o parameters, which gives an alternative to the null hypothesis within the range of pertinent values of the parameter. i.e., i Ho is accepted, what hypothesis is to be rejected and vice versa. An alternative hypothesis is denoted by HA. The idea of alternative hypothesis was originated by ***Neyman.*** For instance,

if Ho : $\mu = 0$, the alternatives are, HA: $\mu \neq 0$, HA$\mu > 0$ or HA$\mu < 0$

if Ho : $\mu_1 = \mu 2$, the alternatives are, HA : $\mu 1 \neq \mu 2$, HA $\mu 1 > \mu 2$ or HA$\mu 1 < \mu 2$

## Standard Error

The standard deviation of the sampling distribution is called the ***standard error.*** For example, $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4 \ldots$.,etc. are not the means of all the samples drawn from the population. The standard deviation of all these means is the standard error of the mean. The formula for this is $\frac{\sigma}{\sqrt{n}}$

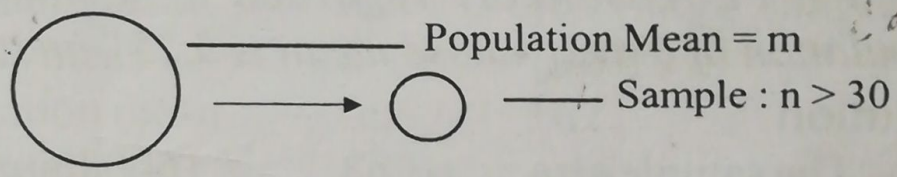
## Large Samples and Small Samples

If the sample size is greater than 30 i.e., if $n > 30$, then those samples may be regarded as large samples. On the other hand if the sample size is less than 30 i.e., if $n < 30$, then those samples may be regarded as small samples.

## Tests of Significance for Large Samples

**Explanation : I Based on Mean**

**I Population and one sample**




**Test :** Whether the sample is random drawn from the population with mean m.

Consider a sample of size n (greater than 30 i.e., a large sample) drawn from a population with mean 'm' and S.D 'σ'. Let $\bar{x}$ be the sample mean (statistic). This sample mean has got a distribution. In this case, since the sample is large the mean $\bar{x}$ is following a normal distribution with mean m (the same as the population mean) and S.D is $\frac{\sigma}{\sqrt{n}}$ whatever be the distribution of the population.

$$\therefore \bar{x} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{\bar{x}-m}{\sigma/\sqrt{n}} \sim N(0,1)$$

**Hypothesis**

The sample is random drawn from the population with mean 'm' or the difference between the sample mean and population mean is nil. It is called ***null hypothesis.***

$$\text{Critical Ratio } (C.R) = \frac{|\bar{x} - m|}{\sigma/\sqrt{n}}$$

where $\bar{x}$ = sample mean
$n$ = size of sample
$m$ = mean of population
$\sigma$ = S.D. of population

If the value of C.R is less than 2, it is accepted at 5% level of significance. If C.R. is greater than 2 and less than 3 it is accepted at 1% level of significance and if C.R is more than 3 the hypothesis is rejected.

**Illustration : 1**

*A sample of 1600 leaves has a mean length of 5.4" could it be reasonably regarded as a sample from a population of leaves whose mean is 5.25" and S.D = 2.6"*

**Solution**

| | | |
|---|---|---|
| The sample size | $n$ | = 1600 (large sample) |
| Sample mean | $\bar{x}$ | = 5.4" |
| Population mean | m | = 5.25" ($\mu$) |
| Population | S.D $\sigma$ | = 2.6" |

**Null hypothesis**

The sample is random drawn from the population with mean 5.25"

or

The difference between the sample mean and the population mean is not significant.

$$C.R. = \frac{|\bar{x} - m|}{\sigma/\sqrt{n}}$$

$$= \frac{5.4 - 5.25}{2.6/\sqrt{1600}}$$

$$= \frac{0.15}{2.6/40} = \frac{0.15}{0.065} = 2.308$$

**Inference**

C.R. value is more than 2 and less than 3, so it is accepted at 1% level of significance.

**Illustration-2:** *A sample of 100 individuals is drawn from a population whose mean is 162 and S.D is 4.16. The mean of the sample is 162.23. Does the sample mean represent a significance divergence from the population mean. Explain.*

**Solution**

| | | |
|---|---|---|
| The sample size | $n_1$ | = 100 (large sample) |
| Sample mean | $\bar{x}$ | = 162.23 |
| Population mean | m | = 162 |
| Population | S.D $\sigma$ | = 4.16 |

**Null hypothesis**

The difference between the sample mean and the population mean is not significant.

$$C.R. = \frac{|\bar{x} - m|}{\sigma/\sqrt{n}}$$

$$= \frac{162.23 - 162}{4.16/\sqrt{100}}$$

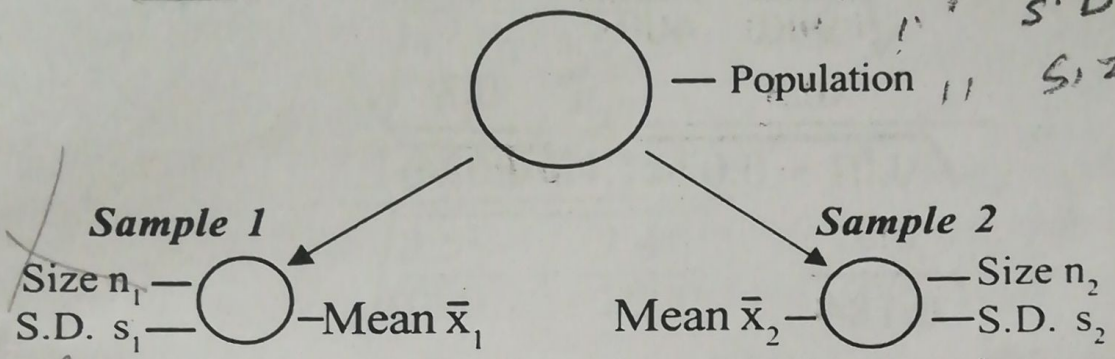
$$= \frac{0.23}{4.16/10}$$

$$= \frac{0.23}{0.416} = 0.5528$$

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

**Explanation: III Based on mean and S.D**

**III Population and two samples**




In this case $\sigma$ is not given, but $s_1$ and $s_2$ are given.

In case, the standard deviation of the population is not known we make use of the sample variance themselves to obtain an estimate of the population variance.

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

**Illustration -3:** *Two random samples of sizes 900 and 400 have means 21 and 21.3 with S.D 3 and 3.1 respectively. These two samples are drawn from the same population or different populations. Explain.*

**Solution**

| Sample 1 | Sample 2 |
|---|---|
| $n_1 = 900$ | $n_2 = 400$ |
| $\bar{x}_1 = 21$ | $\bar{x}_2 = 21.3$ |
| $s_1 = 3$ | $s_2 = 3.1$ |

**Null hypothesis**

The two samples are drawn from the same population.

**Level of significance:** 5% or 1%

**Test statistic**

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{|21 - 21.3|}{\sqrt{\frac{3^2}{900} + \frac{3.1^2}{400}}}$$

$$= \frac{0.3}{\sqrt{0.01 + 0.024}} = \frac{0.3}{\sqrt{0.034}}$$

$$= \frac{0.3}{0.184}$$

$$= 1.63$$

**Inference**

The value is less than 2, so the hypothesis is accepted at 5% level of significance i.e., the two samples are drawn from the same population.

**Illustration-4:** *Random sample of height of 1000 South Indian boys between the ages 18-20 has the mean 64.08" and S.D 2.5". While the sample of 2000 North Indian boys on the same age group has mean 64.15" and S.D = 2.4". Can be suppose that the North Indian boys are on the average taller than the South Indian boys and the real difference is hidden in these two samples due to change of fluctuations.*

**Solution**

| Sample 1 | Sample 2 |
|---|---|
| S. Indian boys | N. Indian boys |
| $n_1 = 1000$ | $n_2 = 2000$ |
| $\bar{x}_1 = 64.08"$ | $\bar{x}_2 = 64.15"$ |
| $s_1 = 2.5"$ | $s_2 = 2.4"$ |

**Null hypothesis:** The difference between the two samples are not significant i.e., these two samples are drawn from the same population.

**Level of significance :** 5% or 1%

**Test statistic**

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{|64.08 - 64.15|}{\sqrt{\frac{2.5^2}{1000} + \frac{2.4^2}{2000}}}$$

$$= \frac{0.07}{\sqrt{0.006 + 0.003}} = \frac{0.07}{\sqrt{0.009}}$$

$$= \frac{0.07}{0.094}$$

$$= 0.74$$

**Inference**

C.R value is less than 2, so the hypothesis is accepted at 5% level of significance i.e., the difference between the two samples are not significant i.e., the two samples are belonging to the same population.

**Illustration-5:** *Two diets are compared by conducting an experiment on two samples of 40 and 50 animals. The averages for the increase in weight due to the diets A and B are 10 and 12 lbs. With S.D.2 and 2.5 lbs respectively. Is the difference between the averages for the increase a weight significant.*

**Solution**

**Null hypothesis:** There is no significant difference between the average for the increase in weight in two samples.

**Level of significance :** 5% or 1%

**Test statistic**

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

**Calculation**

| Sample : 1 | Sample : 2 |
|---|---|
| $n_1 = 40$ | $n_2 = 50$ |
| $x_1 = 10$ | $x_2 = 12$ |
| $s_1 = 2$ | $s_2 = 2.5$ |

$$C.R. = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{|10-12|}{\sqrt{\frac{2^2}{40} + \frac{2.5^2}{50}}}$$

$$= \frac{2}{\sqrt{\frac{4}{40} + \frac{6.25}{50}}}$$

$$= \frac{2}{\sqrt{0.1 + 0.125}}$$

$$= \frac{2}{\sqrt{0.225}}$$

$$= \frac{2}{0.474} = 4.21$$

**Inference**

C.R value is more than 3. So the hypothesis is rejected i.e., there is a significant difference between the average for the increase in weight in two samples.

**Illustration-6:** *Two random samples of 1000 and 2000 forms have average yield 1920 lbs and 1955 lbs o*

*orn per acre. If the S.D of corn yield in the country is assume s 100 lbs. Can you conclude the two samples differ ignificantly?*

**Solution**

**Null hypothesis:** These two samples are random drawn rom the same population.

**Level of significance:** 5% or 1%

**Test statistic**

$$= \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

**Calculation**

Sample 1
$n_1 = 1000$
$\bar{x}_1 = 1920$ lbs
$\sigma = 100$

Sample 2
$n_2 = 2000$
$\bar{x}_2 = 1955$ lbs

$$C.R = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{|1920 - 1955|}{100\sqrt{\frac{1}{1000} + \frac{1}{2000}}}$$

$$= \frac{35}{100\sqrt{0.001 + 0.0005}}$$

$$= \frac{35}{100\sqrt{0.0015}}$$

$$= \frac{35}{100 \times 0.0387}$$

$$= \frac{35}{3.87}$$

$$= 9{:}043$$

$$= \frac{0.6378}{\sqrt{0.034 + 0.010}}$$

$$= \frac{0.6378}{\sqrt{0.044}}$$

$$= \frac{0.6378}{0.2097}$$

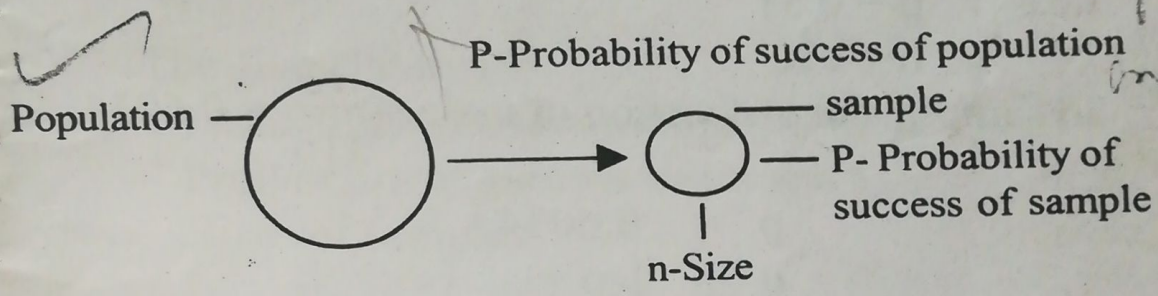
$$= 3.041$$

**Inference**

C.R value is more than 2, so the hypothesis is rejected at 5% level of significance. i.e., the difference between the correlation coefficient of two samples are significant.

**Explanation VIII: Based on the proportion of success**

**VIII : Population with one sample**




In this case, the given items are

P = probability of success of population

p = probability of success of sample

n = size of the sample

$$C.R = \frac{|p-P|}{\sqrt{\frac{pq}{n}}}$$

**Illustration-11**: *The Presidents of the U.S.A have a total of 70 sons and 46 daughters. Is this and unusual proportion if the ratio of male birth to total births in the population at large is 0.51?*

**Solution**

**Null hypothesis:** The difference between the proportion of success of sample and population is not significant. i.e., the sample is random drawn from the population with the proportion of success - P.

**Level of significance:** 5% or 1%

**Test statistic :** $$\frac{|p-P|}{\sqrt{\frac{pq}{n}}}$$

**Calculation**

Sons : 70
Daughters : 46
Total : 116

Proportion of male in population = 0.51

i.e., $p = 0.51$

$n = 116$

In sample, the proportion of male $= \frac{70}{116} = 0.6034$

$\therefore$ $p = 0.6034$

$q = 1-p$

$= 1-0.6034$

$= 0.3966$

$$C.R = \frac{|p-P|}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{|0.6034-0.51|}{\sqrt{\frac{0.6034 \times 0.3966}{116}}}$$

$$= \frac{0.0934}{\sqrt{0.0020}}$$

$$= \frac{0.0934}{0.0447}$$

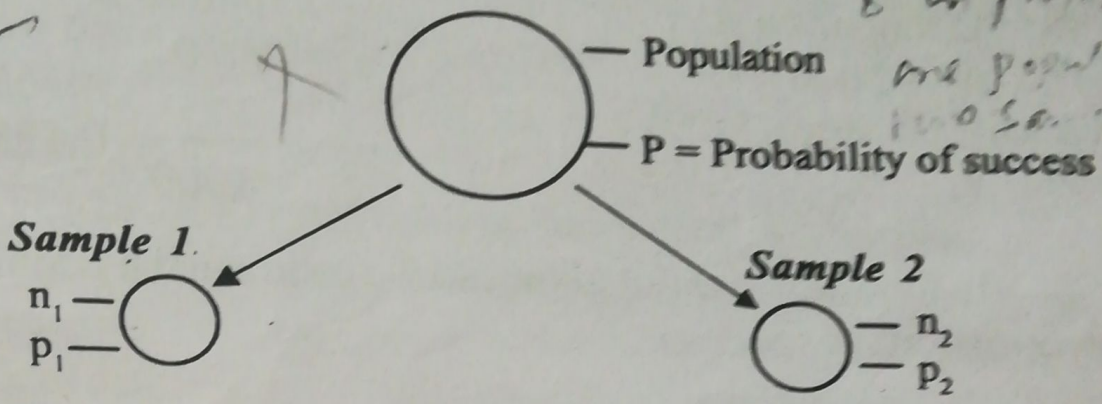
$$= 2.089$$

**Inference**

The C.R value is more than 2 and less than 3, so the hypothesis is accepted at 1% level of significance. i.e., The difference between the proportion of success of sample and population is not significant.

**Explanation: IX: Based on the proportion of success**

**IX: Population with two samples**




The given items are

Probability of success of population = P

Probability of success of samples = $P_1$ & $P_2$

Sizes of samples = $n_1$ & $n_2$

$$C.R = \frac{|p_1 - p_2|}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

If p is not given, we estimate the population proportion of success

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

**Illustration-12:** *Out of 2000 mean in occupation A, 400 die before attaining the age of 50 while out of 1000 men in occupation B, 175 die before attaining the age of 50. Is the difference in proportion is significant?*

**Solution**

**Null hypothesis:** The difference between the proportion of success of two samples are not significant. The two samples are belong to the same population.

**Level of significance:** 5% or 1%

**Test statistic :**

$$\frac{|p_1-p_2|}{\sqrt{\frac{pq}{n_1}+\frac{pq}{n_2}}}$$

**Calculation**

Occupation A

$n_1 = 2000$

$p_1 = \frac{400}{2000} = 0.2$

Occupation B

$n_2 = 1000$

$p_2 = \frac{175}{1000} = 0.175$

In this case, p is not given, so we estimate the population proportion of success.

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

$$= \frac{[2000 \times 0.2] + [1000 \times 0.175]}{2000 + 1000}$$

$$= \frac{400 + 175}{3000} = \frac{575}{3000}$$

$p = 0.1917$

$q = 1-p = 1-0.1917 = 0.8083$

$p = 0.1917$

$q = 0.8083$

$$C.R. = \frac{|p_1-p_2|}{\sqrt{\frac{pq}{n_1}+\frac{pq}{n_2}}}$$

$$= \frac{|0.2-0.175|}{\sqrt{\frac{0.1917 \times 0.8083}{2000} + \frac{0.1917 \times 0.8083}{1000}}}$$

$$= \frac{0.025}{\sqrt{0.00022}}$$

$$= \frac{0.025}{0.015}$$

$$= 1.666$$

**Inference**

The C.R. value is less than 2, so the hypothesis is accepted at, 5% level of significance i.e., the difference between the proportion of success of two samples are not significant. The two samples are belong to the same population.

## Tests of Significance for Small Samples

If the sample size is less than 30 i.e., $n < 30$ then those samples may be regarded as small samples. Principles of statistical inference are the same as in large samples but the techniques differ in the case of small samples. Here Student's 't' can be used.

### Student's t-test

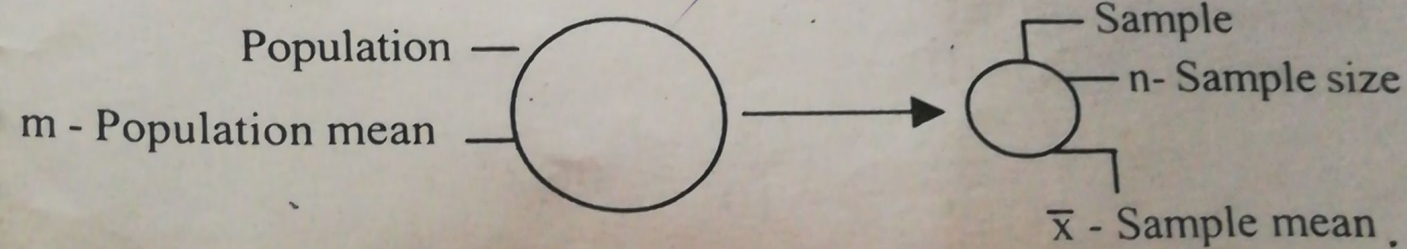
***Sir William Gosset*** contributed a lot to the theory of small samples. i.e., $n < 30$. ***Gosset*** published his discovery in 1905 under the pen name ***'Student'*** and it is popularly known as ***t-test*** or t-distribution or Student's distribution. It is used when the sample is in small size and the population standard deviation is unknown.

### Application of the t-test

1. To test the significance of mean of random sample.
2. To test the difference between the means of two samples.

### Explanation: 1 Based on mean

### I: Population and one sample



In this case, the given items are

Population mean = $m - \mu$

Sample mean = $\bar{x}$

Sample size = $n$

$$C.R = t = \frac{\bar{x} - m}{\sigma E / \sqrt{n}} = \left[\frac{\bar{x} - \mu}{s} x \sqrt{n}\right]$$

Where s = S.D. of sample

$\sigma$ E is unbiased estimate of the standard deviation of the population

$$\sigma E = \sqrt{\frac{ns^2}{n-1}} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

Where $s^2$ = sample variance

**Illustration-13:** *In a random samples of 10 persons selected from a population their heights noted to be*

| Height's in inches | 63 | 63 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 |
|---|---|---|---|---|---|---|---|---|---|---|

Discuss the suggestion that the mean height of the population is 66".

**Solution**

**Null hypothesis:** The sample is random drawn from the population with mean m. The difference between the sample mean and the population mean is not significant.

**Level of significance :** 5% or 1%

**Test statistic**

$$t = \frac{\bar{x} - m}{\sigma E / \sqrt{n}}$$

**Calculation**

$m = 66"$

$n = 10$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{682}{10} = 68.2"$$

$$\sigma E = \sqrt{\frac{ns^2}{n-1}}$$

| x | $x^2$ |
|---|---|
| 63 | 3969 |
| 63 | 3969 |
| 66 | 4356 |
| 67 | 4489 |
| 68 | 4624 |
| 69 | 4761 |
| 70 | 4900 |
| 71 | 5041 |
| 72 | 5184 |
| 73 | 5329 |
| $\Sigma x^2$ | 46622 |

$$\left[ s^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 \right]$$

$$\therefore \quad s^2 = \frac{46622}{10} - (68.2)^2$$

$$= 4662.2 - 4651$$

$$= 11.2$$

$$\therefore \sigma E = \sqrt{\frac{ns^2}{n-1}}$$

$$= \sqrt{\frac{10 \times 11.2}{10-1}}$$

$$= \sqrt{\frac{112}{9}} = 3.528$$

$$t = \frac{\bar{x} - m}{\sigma E / \sqrt{n}}$$

$$= \frac{68.2-66}{3.528/\sqrt{10}} = \frac{2.2}{3.528 / 3.162}$$

$$= \frac{2.2}{1.116}$$

$$= 1.971$$

t 0.05 for (n-1) degrees of freedom

**Table Value**

t 0.05 for (10-1) i.e., 9 degrees of freedom = 2.26

**Inference**

The calculated value 1.971 is less than that of table value 2.26.

So the hypothesis is accepted at 5% level of significance. i.e., the sample is random drawn from the population with mean m. The difference between the sample mean and the population mean is not significant.

**Illustration-14:** *A random sample of size 10 had mean $\bar{x} = 14.3$ and S.D = 1.44. Test at the 5% level of significance that the mean of the population $\mu = 15$.*

**Solution**

**Null hypothesis:** The sample is random drawn from the population with mean $\mu = 15$

**Level of significance** = 5%

**Test statistic** $t = \frac{\bar{x} - \mu}{s} \times \sqrt{n}$

**Calculation**

$\bar{x} = 14.3$

$\mu = 15$

$s = 1.44$

$n = 10$

$\therefore t = \frac{14.3 - 15}{1.44} \times \sqrt{10}$

$= \frac{0.7}{1.44} \times 3.162$

$= 1.54$

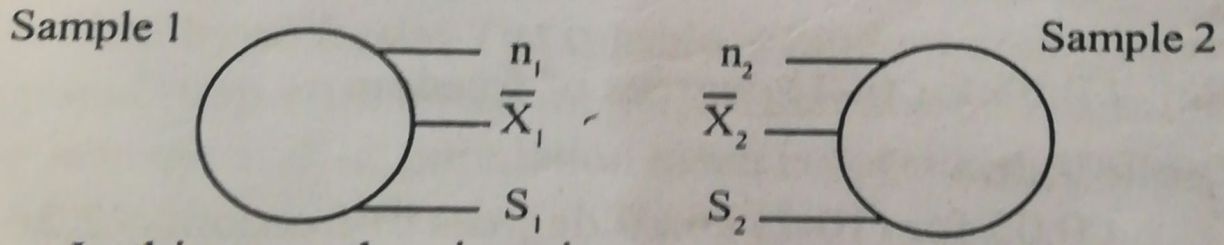
Df, i.e., degrees of freedom = 10-1 = 9

For table value; t 0.05 for 9 degrees of freedom = 2.26

**Inference**

The calculated value 1.54 is less than that of table value 2.26. So the hypothesis is accepted at 5% level of significance. i.e., the sample is random drawn from the population with mean 15.

**Explanation: II Based on mean with two samples:**




In this case, the given items are

The size of the samples = $n_1$ & $n_2$

The mean of the samples = $\bar{x}_1$ & $\bar{x}_2$

The S.D of the sample = $s_1$ & $s_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$$

Where s is combined S.D.

**Illustration-15:** *A group of seven week old chickens reared on a high protein diet weigh 13, 16, 12, 17, 15, 15 and 17 ounces, a second group of 5 chickens similarly treated except that they receive a low protein diet weigh 9, 11, 15, 11 and 14 ounces. Test whether there is significant evidence that additional protein has increased the weight of chickens (The table value of t at 10 of at 5% level of significance is 2.23).*

**Solution**

**Null hypothesis:** The additional protein has not increased the weight of chickens.

**Level of significance :** 5%

**Test statistic:** $t = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$

**Calculation**

$n_1 = 7; n_2 = 5$

| $x_1$ | $(x_1-\bar{x}_1)$ | $(x_1-\bar{x}_1)^2$ | $x_2$ | $(x_2-\bar{x}_2)$ | $(x_2-\bar{x}_2)^2$ |
|---|---|---|---|---|---|
| 13 | -2 | 4 | 9 | -3 | 9 |
| 16 | 1 | 1 | 11 | -1 | 1 |
| 12 | -3 | 9 | 15 | 3 | 9 |
| 17 | 2 | 4 | 11 | -1 | 1 |
| 15 | 0 | 0 | 14 | 2 | 4 |
| 15 | 0 | 0 | | | |
| 17 | 2 | 4 | | | |
| $\sum x_1 = 105$<br>$\bar{x}_1 = \frac{105}{7} = 15$ | $\sum (x_1-\bar{x}) = 0$ | $\sum (x_1-\bar{x})^2 = 22$ | $\sum x_2 = 60$<br>$\bar{x}_2 = \frac{60}{5} = 12$ | $\sum (x_2-\bar{x}) = 0$ | $\sum (x_2-\bar{x})^2 = 24$ |

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\Sigma (x_1-\bar{x}_1)^2 + \Sigma (x_2-\bar{x}_2)^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{22+24}{7+5-2}} = \sqrt{\frac{46}{10}} = 2.14$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \sqrt{\frac{n_1 \times n_2}{n_1+n_2}}$$

$$= \frac{15-12}{2.14} \times \sqrt{\frac{7 \times 5}{7+5}}$$

$$= \frac{3}{2.14} \times \sqrt{\frac{35}{12}}$$

$$= \frac{3}{2.14} \times 1.71 = 2.397$$

Degrees of freedom : $(n_1+n_2-2)$

$= 7+5-2 = 12-2 = 10$

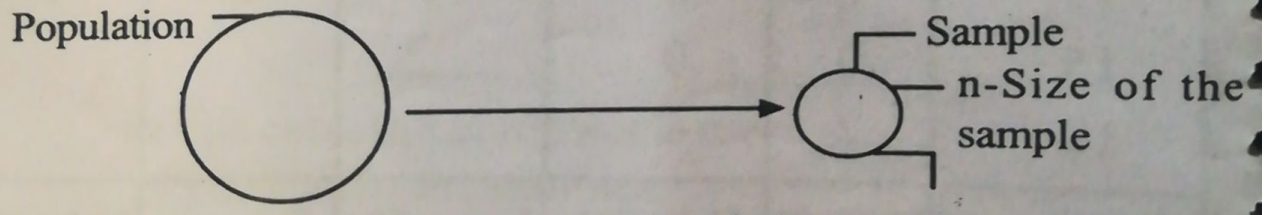
t 0.05 for df 10 the table value is 2.23.

**Inference**

The calculated value 2.397 is more than the table value 2.23. So the hypothesis is rejected at 5% level of significance. i.e., the additional protein has increased the weight of chickens.

**Explanation: III Based on correlation coefficient**

**III Population with one sample**




r - Correlation coefficient of sample

In this case, the given items are

r - correlation coefficient of sample

n - size of the sample

Here degrees of freedom df = (n-2)

# 13. Chi-Square Test and Goodness of Fit

Chi-square test ($\chi^2$) is applied in biostatistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is a measure to study the difference of actual and expected frequencies. It has great use in biostatistics especially in sampling studies.

In sampling studies, we never expect that there will be perfect coincidence between expected and observed frequencies. Since chi-square measures the difference between the expected and observed frequencies. If there is no difference between the actual and expected frequencies, $\chi^2$ is zero. Thus, the chi-square test describes the discrepancy between theory and observation.

## Characteristics of $\chi^2$ Test

1. The test is based on events or frequencies and not based on mean or S.D, etc.

2. The test can be used between the entire set of observed and expected frequencies.

3. To draw inferences, this test is applied, especially testing the hypothesis.

4. It is a general test and is highly useful in research.

## Assumptions

1. The observations must be large.
2. All the observations must be independent.
3. All the events must be mutually exclusive.
4. For comparison purposes, the data, must be in original units.

## Degree of Freedom (df)

When we compare the computed value of $\chi^2$ with the table value, the degree of freedom is evident. The degree of freedom means the number of classes to which values can be assigned. If we have n observed frequencies, the corresponding $\chi^2$ distribution will have (n-1) degrees of freedom. For example, in the case of tossing the coins, there are two possibilities or classes, namely ***head*** and ***tail***. Here df = n-1 i.e. n = Head and tail $\therefore$ df = 2-1 = 1. In such a way according to the classes we fix df, namely n-1.

## Application of Chi-square Test

It is used to test the goodness of fit. The test enables to find out whether the difference between the expected and observed values is significant or not. If the difference is little then the fit is good, otherwise the fit is poor.

## Definition

The $\chi^2$ may be defined as

$$\chi^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$$

where O = Observed frequencies
E = Expected frequencies
$\Sigma$ = Sum of

## Steps

1. A hypothesis is established i.e. Null hypothesis.
2. Calculate the difference between observed value and expected value (O-E).
3. Square the deviations calculated $(O-E)^2$.
4. Divide the $(O-E)^2$ by its expected frequency $(O-E)^2/E$.

5. Add all the values obtained in step 4. $\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$

6. Find the chi-square from $\chi^2$ table at certain level of significance, usually 5% or 1% level.

**Inference**

If the calculated value of $\chi^2$ is greater than the table value of $\chi^2$ at certain degree of level of significance, we reject the hypothesis. If the calculated value of $\chi^2$ is zero, the observed values and expected values completely coincide. If the calculated value of $\chi^2$ is less than table value at certain degree of level of significance, it is said to be non-significant. This implies that the difference between the observed and expected frequencies may be due to fluctuations in sampling.

**Illustration - 1:** *A coin is tossed 100 times of which head comes 60 times and tail 40 times. Would you accept the hypothesis that the coin is normal having no bias for either head or tail.*

**Solution**

**Steps**

**1. Null hypothesis-** i.e. the coin is normal having no bias for either head or tail.

2. Level of significance 5%.

3. Determining expected frequencies (E).

| Possibilities | Observed frequencies (O) | Expected frequencies (E) |
|---|---|---|
| Head | 60 | 50 |
| Tail | 40 | 50 |

4. Fixing the degrees of freedom df = n-1

n = number of events or possibilities i.e. head and tail

n = 2-1 = 1

**6. Calculation**

$$\chi^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$$

where O = Observed value
E = Expected value

| Possi-bilities | Observed frequency (O) | Expected frequency (E) | (O-E) | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| Head | 60 | 50 | 60-50=10 | $(10)^2=100$ | $\frac{100}{50}=2.0$ |
| Tail | 40 | 50 | 40-50=-10 | $(-10)^2=100$ | $\frac{100}{50}=2.0$ |

$$= \Sigma \left[ \frac{(O-E)^2}{E} \right] = 4.00$$

Calculated $\chi^2$ value = 4.00

Table value at 5% level for one degree of freedom is 3.84.

**Inference**

The calculated $\chi^2$ value (4.00) is greater than the table value (3.84). Therefore the hypothesis is rejected. In other words, the coin is defective with bias for head.

**Illustration - 2:** *A dice is tossed 120 times with the following results:*

| No. turned up | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| Frequency | 30 | 25 | 18 | 10 | 22 | 15 | 120 |

Test the hypothesis that the dice is unbiased.

**Solution**

**Steps**

1. **Null hypothesis** - i.e. the dice is an unbiased one.
2. Level of significance 5%.

3. Determining expected frequencies (E).

The expected frequency is $[120 \times \frac{1}{6}] = 20$

4. Fixing the degrees of freedom

$df = n-1$

i.e. $= 6-1 = 5$

**5. Calculation**

$$\chi^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$$

where O = Observed value

E = Expected value

| No. turned up | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| 1 | 30 | 20 | 10 | 100 | 5.00 |
| 2 | 25 | 20 | 5 | 25 | 1.25 |
| 3 | 18 | 20 | -2 | 4 | 0.20 |
| 4 | 10 | 20 | -10 | 100 | 5.00 |
| 5 | 22 | 20 | 2 | 4 | 0.20 |
| 6 | 15 | 20 | -5 | 25 | 1.25 |
| | | | | | $\Sigma = 12.90$ |

$$= \Sigma \left[ \frac{(O-E)^2}{E} \right] = 12.90$$

Calculated $\chi^2$ value = 12.90

For 5df, at 5% level of significance, the table value = 11.07

**Inference**

The calculated $\chi^2$ value (12.90) is greater than the table value (11.07). Therefore the hypothesis is rejected. In otherwords, the dice is biased one.

**Illustration - 3:** *A cross involving different genes gave rise to $F_2$ generation of tall and dwarf in the ratio of 110:90. Test by means of chi-square whether this value is deviated from the Mendel's monohybrid ratio 3:1.*

**Solution**

**Steps**

1. **Null hypothesis:** There is no difference between 110:90 and Mendel's monohybrid ratio 3:1.
2. Level of significance 5%.
3. Determining expected frequencies (E).

Mendel's monohybrid ratio Tall: Dwarf = 3:1

Observed total number = 110+90 = 200

Expected = Tall and dwarf 3 : 1

= 150 : 50 = 200

4. Fixing the degrees of freedom

df = n-1

= 2-1 = 1

**Calculation**

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$$

where O = Observed value

E = Expected value

| Variables | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| Tall | 110 | 150 | -40 | 1600 | 10.6 |
| Dwarf | 90 | 50 | 40 | 1600 | 32.0 |
| | | | | | $\Sigma = 42.6$ |

$$= \Sigma \left[\frac{(O-E)^2}{E}\right] = 42.6$$

Calculated $\chi^2$ value = 42.6

For 1 df, at 5% level of significance the table value = 3.84

**Inference**

The calculated $\chi^2$ value (42.6) is greater than the table value (3.84). Therefore the hypothesis is rejected. In otherwords the value 110:90 is deviated from Mendel's monohybrid ratio 3:1

**Illustration - 4:** *When two heterozygous pea plants are crossed, 1600 plants are produced in the $F_2$ generation out of which 940 are yellow round, 260 are yellow wrinkled, 340 are green round and 60 are green wrinkled. By means of chi-square test whether these values are deviated from Mendel's dihybrid ratio 9 : 3 : 3 : 1. (or By means of chi-square test whether there is real independent assortment).*

**Solution**

**Steps**

1. **Null hypothesis:** There is real independent assortment i.e. there is no difference between observed values and Mendel's dihybrid ratio 9 : 3 : 3 : 1

2. Level of significance 5%.

3. Determining expected frequencies (E) : Mendel's dihybrid ratio 9 : 3 : 3 : 1

Yellow Round = 9 Total 1600 $\therefore E = \frac{9}{16} \times 1600 = 900$

Yellow Wrinkled = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green Round = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green Wrinkled = $\frac{1}{16}$ Total 1600 $\therefore E = \frac{1}{16} \times 1600 = \frac{100}{1600}$

4. Fixing the degrees of freedom df = n-1
= 4-1 = 3

**Calculation**

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$$

where O = Observed value
E = Expected value

| Variables | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| Yellow Round | 940 | 900 | 40 | 1600 | 1.77 |
| Yellow Wrinkled | 260 | 300 | -40 | 1600 | 5.33 |
| Green Round | 340 | 300 | 40 | 1600 | 5.33 |
| Green Wrinkled | 60 | 100 | -40 | 1600 | 16.00 |
| | | | | | $\Sigma$=28.43 |

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right] = 28.43$$

Calculated $\chi^2$ value = 28.43

For 3df, at 5% level of significance,

Table $\chi^2$ value = 7.81

**Inference**

The calculated $\chi^2$ value (28.43) is greater than the table $\chi^2$ value (7.81). Therefore the hypothesis is rejected. In otherwords, there is no real independent assortment or the observed values are deviated from Mendel's dihybrid ratio 9 : 3 : 3 : 1.

**Illustration - 5:** *When a black rat (heterozygous) is crossed with another heterozygous black rat, 43 black, 15 cream and 22 albino offspring are produced in the $F_2$ generation. Using chi-square, test the genetic hypothesis 9:3:4 is consistent with the data.*

**Solution**

**Steps**

1. **Null hypothesis:** The genetic hypothesis 9 : 3 : 4 is consistent with the data.

2. Level of significance 5%.

3. Determining expected frequencies (E).

Genetic hypothesis = 9 : 3 : 4.

$\therefore$ Black = 9 Total offspring = 80 $\therefore E = \frac{9}{16} \times 80 = 45$

Cream = 3 " = 80 $\therefore E = \frac{3}{16} \times 80 = 15$

Albino = $\frac{4}{16}$ " = 80 $\therefore E = \frac{4}{16} \times 80 = \frac{20}{80}$

4. Fixing the degrees of freedom df = n-1

= 3-1=2

**Calculation**

$$\chi^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$$

where O = Observed value

E = Expected value

| Variables | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| Black | 43 | 45 | -2 | 4 | 0.08 |
| Cream | 15 | 15 | 0 | 0 | 0 |
| Albino | 22 | 20 | 2 | 4 | 0.20 |
| | | | | | $\Sigma = 0.28$ |

$$= \Sigma \left[\frac{(O-E)^2}{E}\right] = 0.28$$

Calculated $\chi^2$ value = 0.28

For 2df, at 5% level of significance

the table $\chi^2$ value = 5.96

**Inference**

The calculated $\chi^2$ value (0.28) is less than the table $\chi^2$ value (5.96). Therefore the hypothesis is accepted. In otherwords the observed value is consistent with the ratio 9 : 3 : 4.

**Illustration - 6:** *A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficacy against typhoid. The results are given below: Find out the effectiveness of the drug against the disease (The table value of $\chi^2$ for 1df at 5% level of significance is 3.84).*

| | Typhoid | No typhoid | Total |
|---|---|---|---|
| Administering the drug | 200 | 300 | 500 |
| Without administering the drug | 280 | 20 | 300 |
| Total | 480 | 320 | 800 |

**Solution**

**Steps**

1. **Null hypothesis** i.e. the drug is not effective in preventing typhoid.

2. Level of significance 5%.

3. Preparing 2x2 contingency table (observed value) [O]:

| | Typhoid | No typhoid | Total |
|---|---|---|---|
| Drug | 200 | 300 | 500 |
| No Drug | 280 | 20 | 300 |
| Total | 480 | 320 | 800 |

4. Preparing table for expected frequencies (E).

| | Typhoid | No Typhoid | Total |
|---|---|---|---|
| Drug | $\frac{480 \times 500}{800} = 300$ | $\frac{320 \times 500}{800} = 200$ | 500 |
| No.Drug | $\frac{480 \times 300}{800} = 180$ | $\frac{320 \times 300}{800} = 120$ | 300 |
| Total | 480 | 320 | 800 |

**N.B:** Alternatively, after finding out the first value, the remaining values can be obtained easily in the following manner:

| | Typhoid | No Typhoid | Total |
|---|---|---|---|
| Drug | $\frac{480 \times 500}{800} = 300$ | 500-300 = 200 | 500 |
| No.Drug | 480 - 300 = 180 | 320 - 200 = 120<br>[300-180=120] | 300 |
| Total | 480 | 320 | 800 |

5. Fixing the degrees of freedom

df = (r-1) (c-1)

where r = row

c = column

= (2-1) (2-1)

= 1

**6. Calculation**

$$\chi^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$$

| O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|
| 200 | 300 | -100 | 10000 | 33.33 |
| 280 | 180 | 100 | 10000 | 55.55 |
| 300 | 200 | 100 | 10000 | 50.00 |
| 20 | 120 | -100 | 10000 | 83.33 |
| | | | | $\Sigma = 222.21$ |

$$= \Sigma \left[ \frac{(O-E)^2}{E} \right] = 222.21$$

Calculated $\chi^2$ value = 222.21

For 1df, at 5% level of significance

the table $\chi^2$ value = 3.84

**Inference**

The calculated $\chi^2$ value (222.21) is greater than the table $\chi^2$ value (3.84). Therefore the null hypothesis is rejected. In otherwords the drug is effective in preventing typhoid.