4.1. INTRODUCTION

Very often in practice we are called upon to make decisions about populations on the basis of sample information. For example, we may wish to decide on the basis of sample data whether a new drug is really effective in curing a disease, whether one educational procedure is better than another etc. Such decisions are called statistical inferences or statistical decisions. Two major areas of statistical inference are (1) estimation and (2) testing of hypothesis. By “estimate” we mean “judgment or opinion of the approximate size or amount.” In estimation we use a statistic (a function of sample observations only) to make a close guess about the unknown value of a population parameter. The main aim of hypothesis testing is to provide rules that lead to decision resulting in acceptance or rejection of statements about the population parameters.

In this chapter we shall discuss the problem of testing of hypothesis.

4.2 TESTS OF HYPOTHESIS

The principal objective of statistical inference is to draw inferences (or generalize) about the population on the basis of data collected by sampling from population.

On the basis of sample data drawn from the population a sample statistic is obtained. If the statistic differs from the assumptions made about a population parameter, a decision must be made as to whether or not this difference is significant. If it is, the assumption called hypothesis is rejected otherwise it can be accepted. This
procedure of taking decisions is called ‘Tests of hypothesis or tests of significance ’ or rules of decision.

Before developing the necessary rules, let’s explain few related terms and concepts.

4.3. SOME RELATED TERMS AND CONCEPTS

(I) Parameters

Statistical measures or constants of population such as mean ($\mu$), Standard deviation ($\sigma$), correlation coefficient ($\rho$), population proportion ($P$) etc. are called the parameters.

(II) Statistic

Statistic is a real valued function of the random sample. So statistic is a function of one or more random variables not involving any unknown parameter.

Thus, statistic is a function of samples observations only and is itself a random variable. Therefore a statistic must have a probability distribution.

A value obtained for the sample drawn from the given population is called a statistic. Statistics are generally denoted by small (latin) letters. E.g. $\bar{x}$ = sample mean, $s$ = sample standard deviation, $r$ = sample correlation coefficient, $p$ = sample proportion etc.

(III) Hypothesis

In order to arrive at a decision regarding the population through a sample of the population we have to make certain assumption referred as hypothesis which may or may not be true.
(IV) Statistical Hypothesis

It is an assumption or conjecture or guess about the parameter(s) or population distribution(s). Such assumptions may or may not be true, are called statistical hypothesis and in general these are statements about probability distributions of the population.
(V) Sampling Distribution

Suppose random samples of size ‘n’ are drawn from a population of size N. These samples will give different values of a statistic. For example, we calculate mean $\bar{x}$. The means of the samples will not be identical. If we arrange different values of these means according to their frequencies, the frequency distribution so formed is called sampling distribution of the mean. Similarly we can have sampling distribution of the standard deviation etc.

(VI) Standard Error (S.E.) of a Statistic

The standard deviation of the sampling distribution of a statistic is called the standard error of that statistic. It gives an index of the precision of the estimate of the parameters. For example, the standard error (S.E.) of the sample mean is $\frac{\sigma}{\sqrt{n}}$, where $\sigma$ is the standard deviation of the population from which the sample is drawn and n is the sample size. As the sample size n increases, S.E. decreases. Standard error plays an important role in large sample theory and forms the basis in tests of hypothesis. The standard error enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for the population mean $\mu$ are given by $\bar{x} \pm 3 \frac{s}{\sqrt{n}}$, where s is the sample standard deviation.

(VII) Null Hypothesis (N.H.)

A statistical hypothesis is a statement about a population parameter or the nature of the population. Statistical hypothesis which is formulated with a view of testing or verifying its validity is called null hypothesis and is denoted by $H_0$. The null hypothesis is always tested on the basis of sample information which may or
may not be consistent with it. If the sample information is found to be consistent with \( H_0 \), the null hypothesis is rejected and we conclude that it is false. On the other hand, if the sample information is found to be consistent with \( H_0 \), it is accepted even though, we do not conclude that it is true. The reason is that, the sample information is not sufficient to conclude that it is true. We can say at best that \( H_0 \) is not false. Thus, \( H_0 \) asserts that there is no true difference in the sample statistic and population parameter in the particular matter under consideration and that the difference found is random arising out of fluctuations.

**(VIII) Alternative Hypothesis (A.H.)**

The rejection of null hypothesis i.e. \( H_0 \) implies that it is rejected in favour of some other hypothesis which is accepted. A hypothesis which is accepted in the event of \( H_0 \) being rejected, in otherwise a complimentary hypothesis to null hypothesis is called the alternative hypothesis and is generally denoted by \( H_1 \) or \( H_A \).

**(IX) Testing of Hypothesis**

Test of hypothesis or test of significance or rules of decision are a procedure to decide whether to accept or reject the (null) hypothesis. This test determines whether observed sample differ significantly from expected results. Acceptance of hypothesis merely indicates that the data do not give sufficient evidence to refute the hypothesis. Whereas, rejection is a firm conclusion where the sample evidence refutes it.

When N.H. is accepted, result is said to be non-significant i.e. observed differences are due to ‘chance’ caused by process of sampling. When N.H. is rejected (i.e. A.H. is accepted) the result is said to be significant. Thus test of hypothesis decides whether a statement concerning a parameter is true or false instead of estimating the value of the parameter. Since the test is based on
sample observations, the decision of acceptance or rejection of the null hypothesis is always subjected to some error, i.e., some amount of risk.

**(X) Types of Errors in Test of Hypothesis**

When a statistical hypothesis $H_0$ is tested, we have following four possible results:

(i) $H_0$ is true and it is accepted by the test.

(ii) $H_0$ is false and it is rejected by the test.

(iii) $H_0$ is true but it is rejected by the test.

(iv) $H_0$ is false but it is accepted by the test.

The first two are correct decisions but the latter two lead to errors.

**Type I error** involves rejection of null hypothesis when it should be accepted (as true).

**Type II error** involves acceptance of null hypothesis when it is false and should be rejected.

These errors can be represented in a table as follows:

<table>
<thead>
<tr>
<th>Decision</th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>Correct decision</td>
<td>Type I error</td>
</tr>
<tr>
<td>$H_0$ is false</td>
<td>Type II error</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>
(XI) Level of Significance (L.O.S)

The probability level below which leads to the rejection of the hypothesis is known as the significance level.

The level of significance (L.O.S.) usually denoted by ‘$\alpha$’ is the probability of committing type I error. Thus L.O.S. measures the amount of risk or error associated in taking decisions. It is customary to fix $\alpha$ before sample information is collected and to choose (take) generally $\alpha$ as 0.05 or 0.01 i.e. 5% or 1%. These are called significance levels. L.O.S. $\alpha=0.01$ is used for higher precision and $\alpha=0.05$ for moderate precision. We feel confident in rejecting a hypothesis at 1% L.O.S. than at 5% L.O.S. For example, when a decision is taken at 5% L.O.S., then there are 5 chances out of 100 that we would reject the hypothesis when it should be accepted i.e. we are about 95% confident that we have made the right decision. When we test a hypothesis at 1% L.O.S., there is only 1 chance out of 100 that we would reject the hypothesis when it should be accepted i.e. we are about 99% confident that we have made the right decision. L.O.S. is also known as the size of the test. Thus,

$\alpha$ = Probability of committing type I error

$= P(\text{reject } H_0/H_1) = \alpha$

and $\beta$ = Probability of committing type II error

$= P(\text{accept } H_0/H_1) = \beta$

**Power of the test** is computed as $1 - \beta$. 
Note 1: When the size of the sample is increased, the probability of committing both types of errors I and II i.e. $\alpha$ and $\beta$ can be reduced simultaneously.

Note 2: $\alpha$ and $\beta$ are known as producer’s risk and consumer’s risk respectively.

Note 3: When both $\alpha$ and $\beta$ are small, the test procedure is good one giving good chance of making the correct decision.

(XII) Critical Region (C.R.)

It is important to specify, before the sample is taken, which value of a test statistic $S^*$ will lead to a rejection of $H_0$, and which will lead to acceptance of $H_0$. The area of rejection is called the critical region.

Consider the area under the probability curve of the sampling distribution of the test statistic $S^*$ which follows some known (given) distribution. This area under probability curve is divided into dichotomous regions namely (1) the region of rejection (significant region or critical region) where $H_0$ is rejected and (2) the region of acceptance (non-significant region or non-critical region) where $H_0$ is accepted. Thus Critical region is the region of rejection of $H_0$. The area of critical region = The level of significance $\alpha$. Note that C.R. always lies on the tail(s) of the distribution. Depending on the nature of the $H_1$ (alternative hypothesis), C.R. may lie on one side or both sides of the tail(s).

(XIII) One Tailed Test (O.T.T.) and Two Tailed Test (T.T.T.)
The probability curve of the sampling distribution of the test statistic is generally a normal curve. In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

**Right One Tailed Test (R.O.T.T.)**

When the alternative hypothesis (A.H.): $H_1$ is of the greater than type (mean etc), i.e. $H_1: \mu > \mu_0$ or $H_1: \sigma_1^2 > \sigma_2^2$ etc. then the entire critical region of area $\alpha$ lies on the right side tail of the probability density curve as shown in the Fig. 4.1. In such case, the test of hypothesis is known as **right one tailed test**.

![FIGURE. 4.1](image)

**Left One Tailed Test (L.O.T.T.)**

When the alternative hypothesis (A.H.): $H_1$ is of less than type (mean etc) i.e. $H_1: \mu < \mu_0$ or $H_1: \sigma_1^2 < \sigma_2^2$ etc. then the entire critical region of area $\alpha$ lies on the
left side tail of the probability density curve as shown in the Fig. 4.2. In such case, the test of hypothesis is known as **left one tailed test.**

![Figure 4.2](image)

**FIGURE. 4.2**

**Two Tailed Test (T.T.T.)**

If A.H. is of the not equal type i.e., $H_1: \mu_1 \neq \mu_2$ or $H_1: \sigma_1 \neq \sigma_2$ etc. then the C.R. lies on both sides of the right and left tails of the curve such that C.R. of area $\frac{\alpha}{2}$ lies on the right tail and C.R. of area $\frac{\alpha}{2}$ lies on the left tail, as shown in Fig. 4.3.
Thus the test of hypothesis or test of significances or rule of decision consists of the following six steps.

Step 1: Formulate N.H : $H_0$

Step 2: Formulate A.H. : $H_1$

Step 3: Choose L.O.S. : $\alpha$

Step 4: C.R. is determined by the critical value $S\alpha$ and the kind of A.H. (based on which the test is R.O.T.T. or L.O.T.T. or T.T.T.)

Step 5: Compute the test statistic $S^*$ using the sample data.

Step 6: Decision : Accept or reject N.H. depending on the relation between $S^*$ and $S\alpha$.

(XIV) Degrees of Freedom (d.f.)

The number of degrees of freedom can be interpreted as the number of useful item of information generated by a sample of given size with respect to the
estimation of a given population parameter. Thus, it is defined as ‘the total number of observations minus the number of independent constraints imposed on the observations’. It is based on a concept that one could not have exercised his/her freedom to select all the samples.

The concept can be explained by an analogy:

\[ X + Y = 10 \quad \text{............... (4.1)} \]

In the above equation you have freedom to choose a value for X or Y but not both because when you choose one, the other is fixed. If you choose 8 for X, then Y has to be 2. So the degree of freedom here is 1.

\[ X + Y + Z = 15 \quad \text{............... (4.2)} \]

In the equation (2), one can choose values for two variables but not all. You have freedom to choose 8 for X and 2 for Y. If so, then Z is fixed. So the d.f. is 2.

The degrees of freedom is calculated by subtracting 1 from the size of each group. Thus, the d.f. for selecting ‘n’ observations with one restriction is given as n-1, it is n-2 if two restrictions are given. If n is the number of observations and k is the number of independent constraints then n-k is the number of degrees of freedom. Generally, the degrees of freedom is denoted by Grek symbol \( \nu \) (read as nu). Thus d.f. \( \nu = n - k \).

**Chi-Square Test**

The Chi-square test is a non-parametric test of proportions. It is used to test a hypothesis. If the association between two variables is to be tested this test is commonly used.

The test involves the calculation of quantity, called chi-square from the Greek letter ‘chi’(\( \chi \)) and pronounced as ‘kye’. It was developed by Karl Pearson.
In many of the statistical tests we had to assume that the samples came from normal populations. When this assumption cannot be justified it is necessary to use the test procedure that do not require that these condition to be met. $\chi^2$-test belongs to such test procedure. $\chi^2$-test makes no assumption about the population being sampled. The quantity $\chi^2$ describes the magnitude of discrepancy between theory and observation and hence with the help of $\chi^2$-test we can know whether such discrepancy can be attributed to chance or not. If $\chi^2$ is zero, it indicates that the observed values and theoretical values completely coincide.

**Definition of Statistic $\chi^2$**

Chi-square statistics is defined as square of a standard normal variate. If $O_i$ (i=1,2,...,n) are the observed values (or frequencies) and $E_i$ (i=1,2,...,n) are the theoretical values (or frequencies) of a random variable X from a normal population, then the statistic $\chi^2$ is computed as

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

(i=1,2,...,n)

Here $\chi^2$ follows, Chi-square distribution with (n-1) degrees of freedom.

We find $\chi^2 \leq \chi^2_{n-1,0.05}$ from the $\chi^2$-table.

If $\chi^2 \leq \chi^2_{n-1,0.05}$, we accept $H_0$ : the given theoretical values fits the observed (Empirical) values, otherwise we reject $H_0$. 
Note: If we have estimated $k$ parameters in fitting the given theoretical distributions, the d.f. for $\chi^2$ are $n-k-1$.

Conditions for the Application of $\chi^2$ - test

(1) Sample observations must be independent of each other.
(2) The sample data must be drawn at random from the target population.
(3) The sample should contain at least 50 observations.
(4) There should be no less than 5 observations in any one cell.

Application of $\chi^2$ - test

(1) To test the discrepancy between observed values and theoretical values.
(2) To test goodness of fit.
In this we test whether the frequencies are according to certain assumption or not?
The formula is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

(3) To test whether the two attributes are independent or not?
The formula is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$
Where \( O \) = Observed frequency
\[ E = \text{Expected frequency} \]

We use \( \chi^2 \) test with \((r-1)(c-1)\) d.f. where \( r \) is number of rows and \( c \) is number of columns.

**EXAMPLE-1:** Ratio male to female births in universe is expected to be 1 : 1. In one village it was found that male children born were 52 and females 48. Is this difference due to chance?

**SOLUTION:**

<table>
<thead>
<tr>
<th></th>
<th>Observed frequencies (O)</th>
<th>Expected Frequencies (E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>52</td>
<td>50</td>
</tr>
<tr>
<td>Female</td>
<td>48</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ \chi_c^2 = \sum \frac{(O - E)^2}{E} = \frac{(52 - 50)^2}{50} + \frac{(48 - 50)^2}{50} = \frac{8}{50} = 0.16 \]

The degrees of freedom = \( n - 1 = 2 - 1 = 1 \)

At 5% level of significance for 1 d.f. \( \chi^2 = 3.841 \). \( \chi_c^2 < \chi^2 \), and calculated value of \( \chi_c^2 \) i.e. 0.16 is much lower, hence significant. Therefore the observed difference in births of two sexes is due to chance.

**EXAMPLE-2:** In an experiment of pea breeding, the following frequencies were obtained.

<table>
<thead>
<tr>
<th>Round and Yellow</th>
<th>Wrinkled and Yellow</th>
<th>Round and Green</th>
<th>Wrinkled and Green</th>
</tr>
</thead>
</table>
Theory predicts that the frequencies be in proportions 9:3:3:1. Examine the correspondence between the theory and the experiment.

**SOLUTION:**
Let $H_0$: There is no significant difference between observed (experimental) frequencies and theoretical frequencies

Total frequency = $315 + 101 + 108 + 32 = 556$.

On the basis of proportions $9 : 3 : 3 : 1$, the expected frequencies are

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>315</td>
<td>101</td>
<td>108</td>
<td>32</td>
</tr>
</tbody>
</table>

$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75}$

$\chi^2 = 0.016 + 0.101 + 0.136 + 0.135$

$\chi^2 = 0.606$

d.f. = $4 - 1 = 3$

From the $\chi^2$-table we find that $\chi^2_{0.05} = 7.82$

$\therefore \chi^2 < \chi^2_{0.05}$

$\therefore$ We accept $H_0$ at 5% level of significance and conclude that the theory and experiment are in agreement.

**EXAMPLE-3:** The demand for a particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of parts demanded</td>
<td>1124</td>
<td>1125</td>
<td>1110</td>
<td>1120</td>
<td>1120</td>
<td>1115</td>
</tr>
</tbody>
</table>
Test the hypothesis that the number of parts demanded has no association with the days of the week.

SOLUTION:

Let \( H_0 \): The number of spare parts during the six days.

\[
\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(1124 - 1120)^2}{1120} + \frac{(1125 - 1120)^2}{1120} + \frac{(1110 - 1120)^2}{1120}
\]

\[
= \frac{16 + 25 + 100 + 0 + 36 + 25}{1120} = 0.1804,
\]

d.f. = \( n \cdot 1 = 6 - 1 = 5 \)

From the \( \chi^2 \)-table, we find \( \chi^2_{0.05} = 11.07. \)

Here \( \chi^2_c < \chi^2_{0.05} \)

\( \therefore \) We accept \( H_0 \) and conclude that the number of spare parts demanded does not depend on the day of the week.

**EXAMPLE-4:** A die is thrown 150 times and the following results are obtained.

<table>
<thead>
<tr>
<th>Number turned up</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>19</td>
<td>23</td>
<td>28</td>
<td>17</td>
<td>32</td>
<td>31</td>
</tr>
</tbody>
</table>

Test the hypothesis that the die is unbiased at 5% level of significance.

SOLUTION: \( H_0 \): The die is unbiased
If the die is unbiased all the numbers 1, 2, 3, 4, 5, 6 are equiprobable to turn up. The probability is \( \frac{1}{6} \). Hence out of 150 each number will turn up for \( 150 \times \frac{1}{6} = 25 \) times, which is theoretical frequency of each number.

<table>
<thead>
<tr>
<th>Number Turned up</th>
<th>Observed Frequency (O)</th>
<th>Expected Frequency (E)</th>
<th>O-E</th>
<th>(O-E)^2</th>
<th>( \frac{(O-E)^2}{E} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>25</td>
<td>-6</td>
<td>36</td>
<td>1.44</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>25</td>
<td>-2</td>
<td>4</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>25</td>
<td>3</td>
<td>9</td>
<td>0.36</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>25</td>
<td>-8</td>
<td>64</td>
<td>2.56</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>25</td>
<td>7</td>
<td>49</td>
<td>1.96</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>25</td>
<td>6</td>
<td>36</td>
<td>1.44</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>150</td>
<td>-</td>
<td>7.92</td>
<td></td>
</tr>
</tbody>
</table>

Calculated Value of \( \chi^2 \) at 5 d.f at 5% level of significance is 7.92

The degrees of freedom = \( n-1 = 6-1 = 5 \)

At 5% level of significance for 5 d.f. \( \chi^2_{0.05,5} = 11.07 \)

\( \chi^2 < \chi^2_{0.05,5} \)

therefore Ho is accepted and it can be said that the dice is unbiased

**Contingency Table**

If the observed frequencies occupy a single row, the table is called a classification table. Extending the same idea we can arrive at a two way classification table of R-rows and C-columns figuring observed frequencies and such a table is called a *contingency table*. Corresponding to each observed frequency in \( R \times C \) contingency table we can obtain the corresponding expected (theoretical) frequency by the concept
of proportion. The total frequency in each row or column is called *marginal frequency*. The associated degree of freedom = (R-1)(C-1).

**How to make a 2 x 2 Contingency Table?**

In a contingency table first enter raw data, that is the exact number of subject / animals etc.-not percentage, means or fractions. The group (treatment / control, exposure / no exposure) are entered on the left side as rows, with the treatment group in the top row and the control group in the second row. The outcome is entered as columns on the right side with the positive outcome as the first column and the negative or no change outcome as the second column. The columns and rows are also mutually exclusive. A particular subject or patient can be only in one column not in both.

The following is a form of a 2 x 2 table.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example, Patients with symptoms of viral upper respiratory tract infection were divided into two groups and given placebo or Antihistaminic (AH). The result was recorded as improvement or no improvement after 48 hours.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Improvement</th>
<th>No improvement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antihistaminic</td>
<td>58</td>
<td>442</td>
<td>500</td>
</tr>
<tr>
<td>Placebo</td>
<td>67</td>
<td>423</td>
<td>490</td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>865</td>
<td>990</td>
</tr>
</tbody>
</table>

We state the following formula for obtaining the value of \( \chi^2 \) for 2 x 2 contingency table.

\[
\begin{array}{ccc}
 a & b \\
 c & d \\
\end{array}
\]
The 2 x 2 contingency table along with the marginal frequencies is as follows:

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>B</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Totals</td>
<td>a+c</td>
<td>b+d</td>
<td>N = a+b+c+d</td>
</tr>
</tbody>
</table>

Let E(a), E(b), E(c), E(d) respectively denote the expected frequencies corresponding to the observed frequencies a, b, c, d. In respect of the null hypothesis that the classifications are independent we have

\[ E(a) = \frac{(a+c)(a+b)}{N} \]
\[ E(b) = \frac{(b+d)(a+b)}{N} \]
\[ E(c) = \frac{(a+c)(c-d)}{N} \]
\[ E(d) = \frac{(b+d)(c+d)}{N} \]

**EXAMPLE-1:** The following table shows the results of an experiment to investigate the effect of vaccination induced on the animals against a particular disease. Use the Chi-square test to test the hypothesis that there is no difference between the vaccinated and unvaccinated groups i.e. vaccination and this disease are independent.

<table>
<thead>
<tr>
<th></th>
<th>Got disease</th>
<th>Did not get disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccinated</td>
<td>9</td>
<td>42</td>
</tr>
<tr>
<td>Not vaccinated</td>
<td>17</td>
<td>28</td>
</tr>
</tbody>
</table>
(Value of $\chi^2$ for 1 d.f. at 5% level is equal to 3.84)

**SOLUTION:**

**H$_0$:** Vaccination and Medicine are independent.

<table>
<thead>
<tr>
<th></th>
<th>Got disease</th>
<th>Did not get disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vaccinated</strong></td>
<td>9(a)</td>
<td>42(b)</td>
<td>51(a+b)</td>
</tr>
<tr>
<td><strong>Not vaccinated</strong></td>
<td>17(c)</td>
<td>28(d)</td>
<td>45(c+d)</td>
</tr>
</tbody>
</table>

From the given table $a = 9$, $b = 42$, $c = 17$, $d = 28$, $N = a + b + c + d = 96$

We know that,

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

$$= \frac{(96)(-462)^2}{(51)(26)(70)(45)} = 4.906$$

$\chi^2_{cal} = 4.906 > \chi^2_{0.05} = 3.84$

The hypothesis that the vaccination and disease are independent rejected at 5% level of significance.

**EXAMPLE-2:** The following data relate to the sales, in a time of trade depression of a certain article in a wide demand. Do the data suggest that the sales are significantly affected by depression?
SOLUTION:

H₀: Sales and depression are independent

<table>
<thead>
<tr>
<th>Sales</th>
<th>Trade depreciation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not hit by</td>
<td>Hit by</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>140 (120)</td>
<td>60 (30)</td>
</tr>
<tr>
<td>Not satisfactory</td>
<td>40 (60)</td>
<td>60 (40)</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Expected frequency is obtained as follows:

\[ E_{ij} = \frac{R_i \times C_j}{N} \], where \( R_i \) = Total of \( i^{\text{th}} \) row, \( C_j \) = Total of \( j^{\text{th}} \) column, \( N \) = Total frequency

\[ E_{11} = E(140) = \frac{200 \times 180}{300} = 120 \]

The expected frequencies of different sales are indicated in bracket in the respective cells.

\[ \chi^2 = \sum \frac{(O-E)^2}{E} \]

\[ = \frac{(140-120)^2}{120} + \frac{(60-80)^2}{80} + \frac{(40-60)^2}{60} + \frac{(60-40)^2}{40} \]
\[
\begin{align*}
\text{Degrees of freedom} &= (R - 1) (C - 1) \\
&= (2 - 1)(2 - 1) \\
&= 1
\end{align*}
\]

At 5% level of significance for 1 d.f. \( \chi^2 = 3.841 \).

\( \chi^2 > \chi^2 \) \( \Rightarrow \) therefore \( H_0 \) is rejected and we conclude that the sales and trade depreciation are not independent and sales are significantly affected by depression.

**EXAMPLE-3:** In an experiment to study the dependence of hypertension on smoking habits, the following data were taken on 180 individuals:

<table>
<thead>
<tr>
<th></th>
<th>Non Smoker</th>
<th>Moderate Smoker</th>
<th>Heavy Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>21</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>No Hypertension</td>
<td>48</td>
<td>26</td>
<td>19</td>
</tr>
</tbody>
</table>

Test the hypothesis that the presence (or absence) of hypertension is independent of smoking habits.

**SOLUTION:**

\( H_0 \): Hypertension is independent of smoking habit.

**Observed Frequencies(O):**

|                    | Non Smoker | Moderate Smoker | Heavy Smoker | Total |
|--------------------|------------|-----------------|--------------|
| Hypertension       | 21         | 36              | 30           | 87    |
| No Hypertension    | 48         | 26              | 19           | 93    |
### Expected Frequencies (O):
From the above table we have row total and column total,

<table>
<thead>
<tr>
<th></th>
<th>Non Smoker</th>
<th>Moderate Smoker</th>
<th>Heavy Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypertension</strong></td>
<td>87 × 69/180 = 33</td>
<td>62 × 87/180 = 30</td>
<td>49 × 87/180 = 14</td>
</tr>
<tr>
<td><strong>No Hypertension</strong></td>
<td>93 × 69/180 = 36</td>
<td>62 × 93/180 = 32</td>
<td>49 × 93/180 = 25</td>
</tr>
</tbody>
</table>

### $\chi^2$ - calculation:

<table>
<thead>
<tr>
<th></th>
<th>Obs. Freq (O)</th>
<th>Exp. Freq (E)</th>
<th>(O-E)$^2$</th>
<th>$\frac{(O-E)^2}{E}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypertension:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Smoker</td>
<td>21</td>
<td>33</td>
<td>144</td>
<td>4.36</td>
</tr>
<tr>
<td>Moderate Smoker</td>
<td>36</td>
<td>30</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Heavy Smoker</td>
<td>30</td>
<td>24</td>
<td>36</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>No Hypertension:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Smoker</td>
<td>48</td>
<td>36</td>
<td>144</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Moderate Smoker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>32</td>
<td>36</td>
<td>1.125</td>
</tr>
<tr>
<td>Degrees of freedom = (R-1) (C-1)</td>
<td>(2-1) x (3-1) = 2 d.f</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value of $\chi^2$ at 95% at 2 d.f. = 5.991</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\therefore \chi^2_c = 13.625$ and $\chi^2_r = 5.991$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\therefore \chi^2_c &gt; \chi^2_r$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>So computed value of $\chi^2$ is greater than critical value of $\chi^2$. Therefore Null hypothesis is rejected. Hence it may be concluded that hypertension is not independent of smoking.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>