which is the aggregate of the sampled units of each of the stratum, is termed as *stratified sample* and the technique of drawing this sample is known as *stratified sampling*. Such a sample is by far the best and can safely be considered as representative of the population from which it has been drawn.

14·3. PARAMETER AND STATISTIC

In order to avoid verbal confusion with the statistical constants of the population, *viz.*, mean (μ), variance σ^2, etc., which are usually referred to as *parameters*, statistical measures computed from the sample observations alone, *e.g.*, mean ($\bar{x}$), variance (s^2), etc., have been termed by Professor R.A. Fisher as *statistics*.

In practice parameter values are not known and the estimates based on the sample values are generally used. Thus, statistic which may be regarded as an estimate of parameter, obtained from the sample, is a function of the sample values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to sample. The determination or the characterisation of the variation (in the values of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling, is one of the fundamental problems of the sampling theory.

Remarks **1.** Now onwards, μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

2. *Unbiased Estimate.* A statistic $t = t(x_1, x_2, ..., x_n)$, a function of the sample values $x_1, x_2, ..., x_n$ is an unbiased estimate of the population parameter θ, if $E(t) = \theta$. In other words, if:

$$E\,(Statistic) = \text{Parameter}, \qquad \ldots (14{\cdot}1)$$

then statistic is said to be an unbiased estimate of the parameter.

14·3·1. Sampling Distribution of a Statistic. If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is:

$$^{N}C_{n} = \frac{N!}{n!\,(N-n)!} = k, \,(say).$$

For each of these k samples we can compute some statistic $t = t(x_1, x_2, ..., x_n)$, in particular the mean $\bar{x}$, the variance s^2, etc ., as given below.

Sample Number	*Statistic*		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the *sampling distribution* of the statistic. For example, the values $t_1, t_2, t_3, ..., t_k$ determine the sampling distribution of the statistic t. In other words, statistic t may be regarded as a random variable which can take the values $t_1, t_2, t_3, ..., t_k$ and we can compute the various statistical constants like mean variance, skewness, kurtosis,

etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic t are given by :

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + \ldots + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

and $$\text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_k - \bar{t})^2\right] = \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2.$$

14·3·2. Standard Error. The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well-known statistics, *for large samples*, are given below, where n is the sample size, σ^2 the population variance, and P the population proportion, and $Q = 1 - P$; n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population (s).

S. No.	Statistic	Standard Error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample *s.d.* : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1{\cdot}36263\,\sigma/\sqrt{n}$
6.	Sample median	$1{\cdot}25331\,\sigma/\sqrt{n}$
7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment : μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment : μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^3}{10^4}} \simeq \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample *s.d.*'s : $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions : $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

Utility of Standard Error. S.E. plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If t is any statistic, then for large samples :

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1) \qquad (c.f.\ \S\ 14{\cdot}5)$$

$$\Rightarrow \quad Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1), \text{ for large samples.}$$

Thus, if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than z_α (*c f.* § 14·4·5) times its S.E., the null hypothesis is rejected at α level of significance. Similarly, if

$$| t - E(t) | \leq z_\alpha \times \text{S.E. } (t),$$

the deviation is not regarded significant at 5% level of significance. In other words, the deviation, $t - E(t)$, could have arisen due to fluctuations of sampling and the data do not provide us any evidence against the null hypothesis which may, therefore, be accepted at α level of significance. [For details see § *14·4·3.*]

(*i*) The magnitude of the standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the statistic.

$$S.E.(p) = \sqrt{PQ/n} \quad \text{and} \quad S.E.(\bar{x}) = \sigma/\sqrt{n}$$

In other words, the standard errors of p and $\bar{x}$ vary inversely as the square root of the sample size. Thus in order to double the precision, which amounts to reducing the standard error to one half, the sample size has to be increased four times.

(*ii*) S.E. enables us to determine the probable limits within which the population parameter may be expected to lie. For example, the probable limits for population proportion P are given by : $p \pm 3\sqrt{pq/n}$. (*c. f. Remark* § 14·7·1)

Remark. S.E. of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost, labour and time, etc.

14·4. TESTS OF SIGNIFICANCE

A very important aspect of the sampling theory is the study of the *tests of significance*, which enable us to decide on the basis of the sample results, if

(*i*) the deviation between the observed sample statistic and the hypothetical parameter value, or

(*i*) the deviation between two independent sample statistics ; is significant or might be attributed to chance or the fluctuations of sampling.

Since, for large n, almost all the distributions, *e.g.*, Binomial, Poisson, Negative binomial, Hypergeometric (*c.f.* Chapter 8), t, F (Chapter 16), Chi-square (Chapter 15), can be approximated very closely by a normal probability curve, we use the *Normal Test of Significance* (*c. f.* § 14·7) for large samples. Some of the well-known tests of significance for studying such differences for small samples are *t-test*, *F-test* and Fisher's *z-transformation.*

14·4·1. Null and Alternative Hypotheses. The technique of randomisation used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis—a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called ***null hypothesis*** and is usually denoted by H_0. According to *Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Having set up the null hypothesis we compute the probability P that the deviation between the observed sample statistic and the hypothetical parameter value might have occurred due to fluctuations of sampling. If the deviation comes out to be significant (as measured by a test of significance) null hypothesis is refuted or rejected at the particular level of significance adopted (*c.f.* § 14·4·3) and if the deviation is not significant, null hypothesis may be retained at that level.

Any hypothesis which is complementary to the null hypothesis is called an ***alternative hypothesis***, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ_0, (say), *i.e.*, $H_0 : \mu = \mu_0$ then the alternative hypothesis could be :

(*i*) $H_1 : \mu \neq \mu_0$ (*i.e.*, $\mu > \mu_0$ or $\mu < \mu_0$) (*ii*) $H_1 : \mu > \mu_0$, (*iii*) $H_1 : \mu < \mu_0$

The alternative hypothesis in (*i*) is known as a *two-tailed alternative* and the alternatives in (*ii*) and (*iii*) are known as *right-tailed* and *left-tailed alternatives* respectively. The setting of alternative hypothesis is very important since it enables us to decide whether we have to use a single-tailed (right or left) or two-tailed test (*c.f.* § 14·4·4).

14·4·2. Errors in Sampling. The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors :

Type I Error : *Reject H_0 when it is true.*

Type II Error : *Accept H_0 when it is wrong, i.e., accept H_0* when H_1 *is true.*

$$\left.\begin{aligned} &\text{If we write } P\{\text{Reject } H_0 \text{ when it is true}\} = P\{\text{Reject } H_0 \mid H_0\} = \alpha \\ &\text{and } P\{\text{Accept } H_0 \text{ when it is wrong}\} = P\{\text{Accept } H_0 \mid H_1\} = \beta \end{aligned}\right\} \quad \ldots (14\cdot2)$$

then α and β are called the *sizes of type I error and type II error*, respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

$$\left.\begin{aligned} &\text{Thus} \quad P\{\text{Reject a lot when it is good}\} = \alpha \\ &\text{and} \quad P\{\text{Accept a lot when it is bad}\} = \beta \end{aligned}\right\} \quad \ldots (14\cdot2a)$$

where α and β are referred to as *Producer's risk* and *Consumer's risk* respectively.

14·4·3. Critical Region and Level of Significance. A region (corresponding to a statistic t) in the sample space S which amounts to rejection of H_0 is termed as *critical region of rejection*. If ω is the critical region and if $t = t(x_1, x_2, \ldots, x_n)$ is the value of the statistic based on a random sample of size n, then

$$P(t \in \omega \mid H_0) = \alpha, \; P(t \in \bar{\omega} \mid H_1) = \beta \qquad \ldots (14\cdot2b)$$

where $\bar{\omega}$, the complementary set of ω, is called the *acceptance region*.

We have $\omega \cup \bar{\omega} = S$ and $\omega \cap \omega = \phi$

The probability 'α' that a random value of the statistic t belongs to the critical region is known as the *level of significance*. In other words, level of significance is the size of the type I error (or the maximum producer's risk). The levels of significance usually employed in testing of hypothesis are 5% and 1%. The level of significance is always fixed in advance before collecting the sample information.

14·4·4. One-tailed and Two-tailed Tests. 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.

A test of any statistical hypothesis where the alternatives hypothesis is one tailed (right-tailed or left-tailed) is called a *one-tailed test*. For example, a test for testing the mean of a population $H_0 : \mu = \mu_0$ against the alternative hypothesis :

$H_1 : \mu > \mu_0$ (*Right-tailed*) or $H_1 : \mu < \mu_0$ (*Left-tailed*), is a *single tailed test.*

In the right-tailed test ($H_1 : \mu > \mu_0$), the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left-tailed test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the distribution .

A test of statistical hypothesis where the alternative hypothesis is two-tailed such as : $H_0 : \mu \neq \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$), is known as *two tailed test* and in such a case the critical region is given by the portion of the area lying in both tails of the probability curve of the test statistic.

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed, we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.

For example, suppose that there are two population brands of bulbs, one manufactured by standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0 : \mu_1 = \mu_2$ and alternative will be $H_1 : \mu_1 \neq \mu_2$, thus giving us two-tailed test. However, if we want to test if the bulbs produced by new process have higher average life than those produced by standard process, then we have $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$, thus giving us a left-tailed test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have : $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$, thus giving us a right-tailed test. Thus, the decision about applying a two-tailed test or a single-tailed (right or left) test will depend on the problem under study.

14·4·5. Critical Values or Significant Values. The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant value*. It depends upon :

(*i*) The level of significance used, and

(*ii*) The alternative hypothesis, whether it is two-tailed or single-tailed.

As has been pointed out earlier, for large samples, the standardised variable corresponding to the statistic t ,*viz.*,

$$Z = \frac{t - E(t)}{S.E(t)} \sim N(0, 1), \qquad \ldots (*)$$

asymptotically as $n \to \infty$. The value of Z given by (*) under the null hypothesis is known as *test statistic*. The critical value of the test statistic at level of significance α for a two-tailed test is given by z_α, where z_α is determined by the equation :

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots (14{\cdot}2c)$$

i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, from (14·2*c*), we get

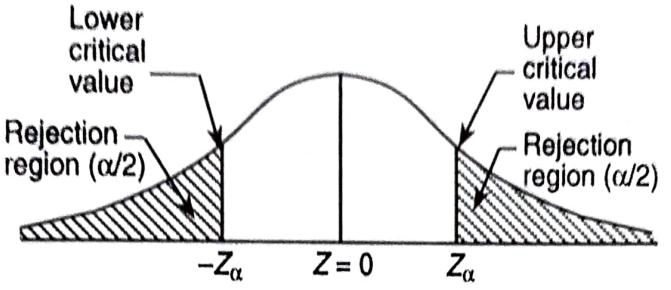
$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \quad \Rightarrow \quad P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

$$\Rightarrow \qquad 2P(Z > z_\alpha) = \alpha, \quad \Rightarrow \quad P(Z > z_\alpha) = \alpha/2$$

In other words, the area of each tail is $\alpha/2$. Thus z_α is the value such that area to the right of z_α is $\alpha/2$ and to the left of $(-z_\alpha)$ is $\alpha/2$, as shown in the following diagram :

TWO-TAILED TEST

(*Level of Significance* 'α')



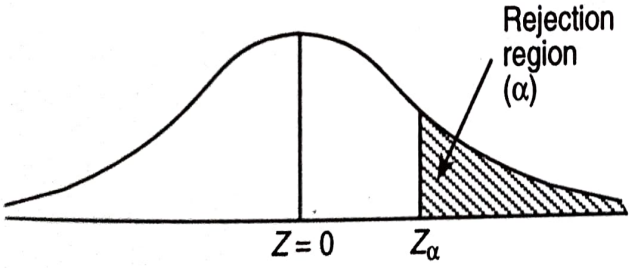
In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $(-z_\alpha)$ is α (See diagrams below), *i.e.*,

For Right-tailed Test : $P(Z > z_\alpha) = \alpha$ ... (14·2*d*)

For Left-tailed Test : $P(Z < -z_\alpha) = \alpha$ ... (14·2*e*)

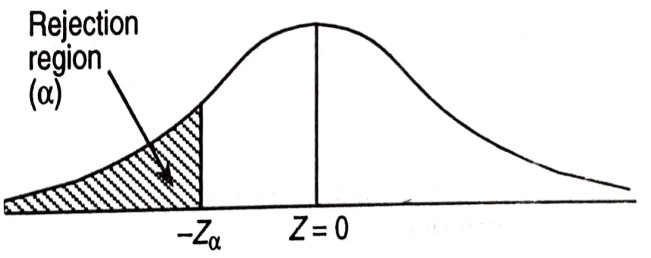
RIGHT-TAILED TEST

(*Level of Significance* 'α')



LEFT-TAILED TEST

(*Level of Significance* 'α')



Thus *the significant* or *critical value of Z for a single-tailed test* (*left or right*) *at level of significance* 'α' *is same as the critical value of Z for a two-tailed test at level of significance* '2α'.

We give below, the critical values of Z at commonly used levels of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (14·2*c*), (14·2*d*), (14·2*e*), on using the Normal Probability Tables as explained in § 14·6.

Critical value (z_α)	*Level of significance* (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2{\cdot}58$	$\lvert Z_\alpha \rvert = 1{\cdot}96$	$\lvert Z_\alpha \rvert = 1{\cdot}645$
Right-tailed test	$Z_\alpha = 2{\cdot}33$	$Z_\alpha = 1{\cdot}645$	$Z_\alpha = 1{\cdot}28$
Left-tailed test	$Z_\alpha = -2{\cdot}33$	$Z_\alpha = -1{\cdot}645$	$Z_\alpha = -1{\cdot}28$

Remark. If n is small, then the sampling distribution of the test statistic Z will not be normal and in the that case we can't use the above significant values which have been obtained from normal probability curves. In this case, *viz.*, n small (usually less than 30), we use the

significant values based on the exact sampling distribution of the statistic Z. [defined in (*), § 14·4·5, which turns out to be t, F, or χ^2 [see Chapters 14 and 16]. These significant values have been tabulated for different values of n and α and are given at the end of Chapters 15 and 16.

14·5. PROCEDURE FOR TESTING OF HYPOTHESIS

We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1. *Null Hypothesis.* Set up the Null Hypothesis H_0.

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis H_1. This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. *Level of Significance.* Choose the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, *i.e.*, α is fixed in advance.

4. *Test Statistic (or Test Criterion).* Compute the test statistic :

$$Z = \frac{t - E(t)}{S.E.(t)}, \text{ under } H_0.$$

5. *Conclusion.* We compare the computed value of Z in *step* 4 with the significant value (tabulated value) z_α at the given level of significance, 'α'.

If $|Z| < z_\alpha$, *i.e.*, if the calculated value of Z (in modulus value) is less than z_α we say it is not significant. By this we mean that the difference $t - E(t)$ is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may, therefore, be accepted.

If $|Z| > z_\alpha$ *i.e.*, if the computed value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance α, *i.e.*, with confidence coefficient $(1 - \alpha)$.

14·6. TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

In this section, we will discuss the tests of significance when samples are large. We have seen that for large values of n, the number of trials, almost all the distributions, *e.g.*, binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test*, which is based upon the following fundamental property (*area property*) of the normal probability curve.

If $X \sim N(\mu, \sigma^2)$, then $$Z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$$

Thus from the normal probability tables, we have

$$P(-3 \le Z \le 3) = 0{\cdot}9973, \quad i.e., \quad P(|Z| \le 3) = 0{\cdot}9973$$

$$\Rightarrow \quad P(|Z| > 3) = 1 - P(|Z| \le 3) = 0{\cdot}0027 \qquad \ldots (14{\cdot}3)$$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability tables, we get

$$P(-1{\cdot}96 \le Z \le 1{\cdot}96) = 0{\cdot}95, \quad i.e., \quad P(|Z| \le 1{\cdot}96) = 0{\cdot}95$$

$$\Rightarrow \quad P(|Z| > 1{\cdot}96) = 1 - 0{\cdot}95 = 0{\cdot}05 \qquad \ldots (14{\cdot}3a)$$

and $$P(|Z| < 2{\cdot}58) = 0{\cdot}99 \Rightarrow P(|Z| > 2{\cdot}58) = 0{\cdot}01 \qquad \ldots (14{\cdot}3b)$$

Thus the significant values of Z at 5% and 1% levels of significance for a two-tailed test are 1·96 and 2·58 respectively.

Thus the steps to be used in the normal test are as follows :

(*i*) Compute the test statistic Z under H_0.

(*ii*) If $|Z| > 3$, H_0 is always rejected.

(*iii*) If $|Z| \leq 3$, we test its significance at certain level of significance, usually at 5% and sometimes at 1% level of significance. Thus, for a two-tailed test if $|Z| > 1{\cdot}96$, H_0 is rejected at 5% level of significance.

Similarly if $|Z| > 2{\cdot}58$, H_0 is contradicted at 1% level of significance and if $|Z| \leq 2{\cdot}58$, H_0 may be accepted at 1% level of significance.

From the normal probability tables, we have :

$$P(Z > 1{\cdot}645) = 0{\cdot}5 - P(0 \leq Z \leq 1{\cdot}645) = 0{\cdot}5 - 0{\cdot}45 = 0{\cdot}5 - 0{\cdot}45 = 0{\cdot}05$$

$$P(Z > 2{\cdot}33) = 0{\cdot}5 - P(0 \leq Z \leq 2{\cdot}33) = 0{\cdot}5 - 0{\cdot}49 = 0{\cdot}01$$

Hence for a single-tail test (Right-tail or Left-tail) we compare the computed value of $|Z|$ with 1·645 (at 5% level) and 2·33 (at 1% level) and accept or reject H_0 accordingly.

Important Remark. In the theoretical discussion that follows in the next sections, the samples under consideration are supposed to be large. For practical purposes, sample may be regarded as large if $n > 30$.

14·7. SAMPLING OF ATTRIBUTES

Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively exhaustive classes—one class possessing a particular attribute, say *A*, and the other class not possessing that attribute, and then note down the number of persons in the sample of size *n*, possessing that attribute. The presence of an attribute in sampled unit may be termed as success and its absence as failure. In this case a sample of *n* observations is identified with that of a series of *n* independent Bernoulli trials with constant probability *P* of success for each trial. Then the probability of *x* successes in *n* trials, as given by the binomial probability distribution is : $p(x) = {}^nC_x P^x Q^{n-x}$; $x = 0, 1, 2, \ldots, n$.

14·7·1. Test of Significance for Single Proportion. If X is the number of successes in *n* independent trials with constant probability *P* of success for each trial, then

$E(X) = nP$ and $V(X) = nPQ$, where $Q = 1 - P$, is the probability of failure.

It has been proved that for large *n*, the binomial distribution tends to normal distribution. Hence for large *n*, $X \sim N(nP, nPQ)$, *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1) \qquad \ldots (14{\cdot}4)$$

and we can apply the normal test.

Remarks 1. In a sample of size *n*, let X be the number of persons possessing the given attribute. Then

Observed proportion of successes = $X/n = p$, (say).

$$\therefore \quad E(p) = E(X/n) = \frac{1}{n} E(X) = \frac{1}{n} nP = P \qquad \ldots (14{\cdot}4a)$$

Thus the sample proportion '*p*' gives an unbiased estimate of the population proportion *P*.

Also $V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n} \quad \Rightarrow \quad \text{S.E.}(p) = \sqrt{\frac{PQ}{n}}$ ... (14·4b)

Since X and consequently X/n is asymptotically normal for large n, the normal test for the proportion of successes becomes :

$$Z = \frac{p - E(p)}{\text{S. E. }(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \qquad \text{... (14·4c)}$$

2. If we have sampling from a finite population of size N, then

$$\text{S.E.}(p) = \sqrt{\left(\frac{N-n}{N-1}\right) \cdot \frac{PQ}{n}} \qquad \text{... (14·4d)}$$

3. Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are :

$$E(p) \pm 3\,\text{S.E.}(p), \quad \textit{i.e.,} \quad P \pm 3\sqrt{PQ/n}.$$

If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are : $p \pm 3\sqrt{pq/n}$. ... (14·4e)

However, the limits for P at level of significance α are given by : $p \pm z_\alpha \sqrt{pq/n}$, ... (14·4f)

where z_α is the significant value of Z at level of significance α.

In particular : 95% confidence limits for P are given by : $p \pm 1{\cdot}96\sqrt{pq/n}$, ... (14·4g)

and 99% confidence limits for P are given by : $p \pm 2{\cdot}58\sqrt{pq/n}$. ... (14·4h)

Example 14·1. *A die is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the die cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.*

Solution. If the coming of 3 or 4 is called a success, then in usual notations :

$$n = 9{,}000\ ; \quad X = \text{Number of successes} = 3{,}240$$

Under the null hypothesis (H_0) *that the die is an unbiased one,* we get

$$P = \text{Probability of success} = \text{Probability of getting a 3 or 4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Alternative hypothesis, $H_1 : p \neq \frac{1}{3}$, (*i.e.,* die is biased).

We have $Z = \frac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$, since n is large.

Now $$Z = \frac{3240 - 9000 \times (1/3)}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44{\cdot}73} = 5{\cdot}36$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the die is almost certainly biased.

Since die is not unbiased, $P \neq \frac{1}{3}$. The probable limits for 'P' are given by :

$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n}$, where $\hat{P} = p = \frac{3{,}240}{9{,}000} 0{\cdot}36$ and $\hat{Q} = q = 1 - p = 0{\cdot}64$.

Probable limits for population proportion of successes may be taken as :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}36 \pm 3\sqrt{\frac{0{\cdot}36 \times 0{\cdot}64}{9000}} = 0{\cdot}36 \pm 3 \times \frac{0{\cdot}6 \times 0{\cdot}8}{30\sqrt{10}} = 0{\cdot}345 \text{ and } 0{\cdot}375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0·345 and 0·375.

Example 14·2. *A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the proportion of bad ones in a sample of this size is 0·015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8·5 and 17·5.*

Solution. Here we are given : $n = 500$

X = Number of bad pineapples in the sample = 65

p = Proportion of bad pineapples in the sample = $\frac{65}{500} = 0{\cdot}13 \Rightarrow q = 1 - p = 0{\cdot}87$

Since P, the proportion of bad pineapples in the consignment is not known, we may take (as in the last example) : $\hat{P} = p = 0{\cdot}13$, $\hat{Q} = q = 0{\cdot}87$.

S.E. of proportion = $\sqrt{\hat{P}\hat{Q}/n} = \sqrt{0{\cdot}13 \times 0{\cdot}87/500} = 0{\cdot}015$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}130 \pm 3 \times 0{\cdot}015 = 0{\cdot}130 \pm 0{\cdot}045 = (0{\cdot}085, 0{\cdot}175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8·5 and 17·5.

Example 14·3. *A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage of bad apples in the consignment.*

Solution. We have :

p = Proportion of bad apples in the sample = $\frac{60}{500} = 0{\cdot}12$

Since significant value of Z at 98% confidence coefficient (level of significance 2%) is 2·33, [from Normal Tables], 98% confidence limits for population proportion are :

$$p \pm 2{\cdot}33\sqrt{pq/n} = 0{\cdot}12 \pm 2{\cdot}33\sqrt{0{\cdot}12 \times 0{\cdot}88/500} = 0{\cdot}12 \pm 2{\cdot}33 \times \sqrt{0{\cdot}0002112}$$
$$= 0{\cdot}12 \pm 2{\cdot}33 \times 0{\cdot}01453 = (0{\cdot}08615, 0{\cdot}15385)$$

Hence 98% confidence limits for percentage of bad apples in the consignment are (8·61, 15·38).

Example 14·4. *In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance ?*

Solution. In the usual notations, we are given : $n = 1{,}000$

X = Number of rice eaters = 540

$\therefore$ p = Sample proportion of rice eaters = $\frac{X}{n} = \frac{540}{1000} = 0{\cdot}54$

Null Hypothesis, H_0 : Both rice and wheat are equally popular in the State so that

P = Population proportion of rice eaters in Maharashtra = 0·5 $\Rightarrow$ $Q = 1 - P = 0{\cdot}5$.

Alternative Hypothesis, $H_1 : P \neq 0{\cdot}5$ (two-tailed alternative)

14·8·2. Standard Error of Sample Mean. *The variance of the sample mean is σ^2/n, where σ is the population standard deviation and n is the size of the random sample. The S.E. of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.*

Proof. Let x_i, $(i = 1, 2, \ldots, n)$ be a random sample of size n from a population with variance σ^2, then the sample mean $\bar{x}$ is : $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$

$$\therefore \quad V(\bar{x}) = V\left\{\frac{1}{n}(x_1 + x_2 + \ldots + x_n)\right\} = \frac{1}{n^2} V(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n^2}\left\{V(x_1) + V(x_2) + \ldots + V(x_n)\right\},$$

the covariance terms vanish since the sample observations are independent.

But $V(x_i) = \sigma^2$, $(i = 1, 2, \ldots, n)$ [From (3) of § 14·8·1]

$$\therefore \quad V(\bar{x}) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \quad \Rightarrow \quad \text{S.E.}(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad \ldots(14\cdot9)$$

14·8·3. Test of Significance for Single Mean. We have proved that if x_i, $(i = 1, 2, \ldots, n)$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$. However, this result holds, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$, even in random sampling from non-normal population provided the sample size n is large [*c.f.* Central Limit Theorem]. Thus for large samples, the *standard normal variate* corresponding to $\bar{x}$ is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Under the *null hypothesis* H_0, that the sample has been drawn from a population with mean μ and variance σ^2, *i.e.*, there is no significant difference between the sample mean ($\bar{x}$) and population mean (μ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(14\cdot9a)$$

Remarks **1.** If the population *s.d.* σ is unknown then we use its estimate provided by the sample variance given by [See (14·8*b*)]. $\hat{\sigma}^2 = s^2 \Rightarrow \hat{\sigma} = s$ (for large samples).

2. *Confidence limits for μ.* 95% confidence interval for μ is given by :

$$|Z| \le 1\cdot96, \text{ i.e., } \left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \le 1\cdot96 \Rightarrow \bar{x} - 1\cdot96\,(\sigma/\sqrt{n}) \le \mu \le \bar{x} + 1\cdot96\,(\sigma/\sqrt{n}) \qquad \ldots(14\cdot10)$$

and $\bar{x} \pm 1\cdot96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ are $\bar{x} \pm 2\cdot58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2\cdot33\sigma/\sqrt{n}$.

However, in sampling from a finite population of size N, the corresponding 95% and 99% confidence limits for μ are respectively

$$\bar{x} \pm 1\cdot96\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{and} \quad \bar{x} \pm 2\cdot58\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \ldots(14\cdot10a)$$

3. The confidence limits for any parameter (P, μ, etc.) are also known as its *fiducial limits.*

$$\Rightarrow \qquad P\left[\,|\,\bar{x}-\mu\,|\,<1{\cdot}96\times(\sigma/\sqrt{n})\right]=0{\cdot}95 \qquad \ldots(**)$$

From (*) and (**), we get $\dfrac{1{\cdot}96\times\sigma}{\sqrt{n}}=10{,}000 \quad\Rightarrow\quad \dfrac{1{\cdot}96\times 30{,}000}{\sqrt{n}}=10{,}000$

$$\therefore \qquad n=(1{\cdot}96\times 3)^2=(5{\cdot}88)^2=34{\cdot}56\simeq 35$$

Aliter. Using Remark to Example 14·19,

$$n=\left(\frac{z_\alpha\cdot\sigma}{E}\right)^2=\left(\frac{1{\cdot}96\times 30{,}000}{10{,}000}\right)^2\simeq 35.$$

14·8·4. Test of Significance for Difference of Means. Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then, since sample sizes are large,

$$\bar{x}_1\sim N(\mu_1,\,\sigma_1^2/n_1) \quad \text{and} \quad \bar{x}_2\sim N(\mu_2,\,\sigma_2^2/n_2)$$

Also $\bar{x}_1-\bar{x}_2$, being the difference of two independent normal variates is also a normal variate. The value of Z (*S.N.V.*) corresponding to $\bar{x}_1-\bar{x}_2$ is given by :

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-E(\bar{x}_1-\bar{x}_2)}{S.E.\,(\bar{x}_1-\bar{x}_2)}\sim N(0,1)$$

Under the null hypothesis, $H_0:\mu_1=\mu_2$, *i.e.*, there is no significant difference between the sample means, we get

$$E(\bar{x}_1-\bar{x}_2)=E(\bar{x}_1)-E(\bar{x}_2)=\mu_1-\mu_2=0;\; V(\bar{x}_1-\bar{x}_2)=V(\bar{x}_1)+V(\bar{x}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2},$$

the covariance term vanishes, since the sample means $\bar{x}_1$ and $\bar{x}_2$ are independent.

Thus under $H_0:\mu_1=\mu_2$, the test statistic becomes (for large samples),

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}}\sim N(0,1) \qquad \ldots(14{\cdot}11)$$

Remarks **1.** If $\sigma_1^2=\sigma_2^2=\sigma^2$, *i.e.*, if the samples have been drawn from the populations with common S.D. σ, then under $H_0:\mu_1=\mu_2$,

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{\{(1/n_1)+(1/n_2)\}}}\sim N(0,1) \qquad \ldots[14{\cdot}11a]$$

2. If in (14·11*a*), σ is not known, then its estimate based on the sample variances is used. If the sample sizes are not sufficiently large, then an unbiased estimate of σ^2 is given by :

$$\hat{\sigma}^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)},\ \text{since}$$

$$E(\hat{\sigma}^2)=\frac{1}{n_1+n_2-2}\{(n_1-1)\,E(S_1^2)+(n_2-1)\,E(S_2^2)\}=\frac{1}{n_1+n_2-2}\left[(n_1-1)\,\sigma^2+(n_2-1)\,\sigma^2\right]=\sigma^2$$

But since sample sizes are large, $S_1^2\simeq s_1^2$, $S_2^2\simeq s_2^2$, $n_1-1\simeq n_1$, $n_2-1\simeq n_2$. Therefore in practice, for large samples, the following estimate of σ^2 without any serious error is used :

$$\hat{\sigma}^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2} \qquad \ldots[14{\cdot}11b]$$

However, if sample sizes are small, then an exact sample test, t-test for difference of means (*c.f.* Chapter 16) is to be used.

Hence the sample size of each of the two groups should be increased by at leas
78 – 50 = 28, in order that the difference between the mean heights of the two group
is significant.

14·8·5. Test of Significance for the Difference of Standard Deviations. I
s_1 and s_2 are the standard deviations of two independent samples, then under nu
hypothesis, $H_0 : \sigma_1 = \sigma_2$, *i.e.*, *i.e.*, sample standard deviations don't differ significantly
the statistic :

$$Z = \frac{s_1 - s_2}{S.E.\ (s_1 - s_2)} \sim N(0, 1), \text{ for large samples.}$$

But in case of large samples, the S.E. of the difference of the sample standard
deviations is given by : $\quad S.E.\ (s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$

$$\therefore \quad Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)}} \sim N(0, 1), \qquad \ldots(14\cdot12$$

σ_1^2 and σ_2^2 are usually unknown and for large samples, we use their estimates
given by the corresponding sample variances. Hence the test statistic reduces to

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)}} \sim N(0, 1) \text{ (for large samples)} \qquad \ldots(14\cdot13$$

Example 14·30. *Random samples drawn from two countries gave the following data relating to the heights of adult males :*

	Country A	*Country B*
Mean height (in inches)	*67·42*	*67·25*
Standard deviation (in inches)	*2·58*	*2·50*
Number in samples	*1,000*	*1,200*

(i) Is the difference between the means significant ?

(ii) Is the difference between the standard deviations significant ?

CHAPTER FIFTEEN

Exact Sampling Distributions-I [Chi-square (χ^2) Distribution]

LEARNING OBJECTIVES. Upon completion of this chapter, you should be able to :

1. Derive chi-square distribution.
2. Explain various concepts like *m.g.f.*, characteristic function, etc. related to chi-square distribution.
3. Discuss various theorems and properties of chi-square distribution.
4. Understand how the chi-square distribution is used to make inferences about a population variance.
5. Demonstrate the use of the chi-square distribution to conduct tests of (*i*) Goodness of fit, and (*ii*) Independence of attributes.
6. Emphasise the need for various other applications of the chi-square distribution.

CHAPTER OUTLINE

CHAPTER CONCEPTS QUIZ / DISCUSSION & REVIEW QUESTIONS / ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

15·1. INTRODUCTION

The square of a standard normal variate is known as a chi-square variate (*pronounced as Ki-Sky without S*) with 1 degree of freedom (*d.f.*).

Thus if $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim N(0, 1)$ and

$$Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2 \text{ is a chi-square variate with 1 } d.f. \qquad \ldots (15\cdot1)$$

In general if X_i, $(i = 1, 2, \ldots, n)$ are n independent normal variates with means μ_i and variances σ_i^2, $(i = 1, 2, \ldots, n)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2, \text{ is a chi-square variate with } n \text{ } d.f. \qquad \ldots (15\cdot1a)$$

15·2. DERIVATION OF THE CHI-SQUARE (χ^2) DISTRIBUTION

First Method—*Method of Moment Generating Function*

If X_i, $(i = 1, 2, \ldots, n)$ are independent $N(\mu_i, \sigma_i^2)$, we want the distribution of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} U_i^2, \text{ where } U_i = \frac{X_i-\mu_i}{\sigma_i} \sim N(0, 1)$$

Since X_i's are independent, U_i's are also independent. Therefore,

$$M_{\chi^2}(t) = M_{\Sigma U_i^2}(t) = \prod_{i=1}^{n} M_{U_i^2}(t) = [M_{U_i^2}(t)]^n, \qquad [\because U_i\text{'s are } i.i.d.\ N(0, 1)] \qquad \ldots (*)$$

$$M_{U_i^2}(t) = E[\exp(tU_i^2)] = \int_{-\infty}^{\infty} \exp(tu_i^2)\, f(x_i)\, dx_i$$

$$= \int_{-\infty}^{\infty} \exp(tu_i^2)\, \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i-\mu)^2/2\sigma^2\} dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tu_i^2) \exp(-u_i^2/2)\, du_i, \qquad \left[u_i = \frac{x_i-\mu}{\sigma}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{1-2t}{2}\right) u_i^2\right\} du_i = \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{\pi}}{\left(\frac{1-2t}{2}\right)^{1/2}} = (1-2t)^{-1/2}$$

$$\left[\because \int_{-\infty}^{\infty} e^{-a^2x^2}\, dx = \frac{\sqrt{\pi}}{a}\right]$$

$$\therefore \qquad M_{\chi^2}(t) = (1-2t)^{-n/2}, \qquad \text{[From (*)]}$$

which is the *m.g.f.* of a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$.

Hence, by uniqueness theorem of *m.g.f.*'s,

$$\chi^2 = \sum_{i}^{n} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2, \text{ is a Gamma variate with parameters } \frac{1}{2} \text{ and } \frac{1}{2}n.$$

$$\therefore \quad dP(\chi^2) = \frac{(\frac{1}{2})^{n/2}}{\Gamma(n/2)} \cdot [\exp(-\tfrac{1}{2}\chi^2)]\,(\chi^2)^{(n/2)-1}\,d\chi^2$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} [\exp(-\chi^2/2)]\,(\chi^2)^{(n/2)-1}\,d\chi^2,\ 0 \le \chi^2 < \infty \qquad \ldots(15\cdot2)$$

which is the required *p.d.f.* of chi-square distribution with n degrees of freedom.

Remarks 1. If a *r.v.* X has a chi-square distribution with n *d.f.*, we write $X \sim \chi^2_{(n)}$ and its *p.d.f.* is :

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-x/2}\, x^{(n/2)-1};\ 0 \le x < \infty \qquad \ldots(15\cdot2a)$$

2. If $X \sim \chi^2_{(n)}$, then $\frac{1}{2}X \sim \gamma\left(\frac{1}{2}n\right)$.

Proof. The *p.d.f.* of $Y = \frac{1}{2}X$, is given by :

$$g(y) = f(x).\left|\frac{dx}{dy}\right| = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-y} \cdot (2y)^{(n/2)-1} \cdot 2 = \frac{1}{\Gamma(n/2)} e^{-y}\, y^{(n/2)-1};\ 0 \le y < \infty$$

$$\therefore \qquad Y = \tfrac{1}{2}X \sim \gamma(\tfrac{1}{2}n).$$

Second Method—*Method of Induction*

If $X_i \sim N(0, 1)$, then $\frac{1}{2}\ X_i^2$ is a $\gamma\left(\frac{1}{2}\right)$ so that X_i^2 is a χ^2 variate with *d.f.* 1.

If X_1 and X_2 are independent standard normal variates then $X_1^2 + X_2^2$ is a chi-square variate with 2 *d.f.* which may be proved as follows :

The joint probability differential of X_1 and X_2 is given by :

$$dP(x_1, x_2) = f(x_1, x_2)\,dx_1\,dx_2 = f_1(x_1)\,f_2(x_2)\,dx_1\,dx_2$$

$$= \frac{1}{2\pi}\exp\{-(x_1^2 + x_2^2)/2\}\,dx_1\,dx_2,\ -\infty < (x_1, x_2) < \infty$$

Let us transform to polar co-ordinates by substitution $x_1 = r\cos\theta$, $x_2 = r\sin\theta$. Jacobian of transformation J is given by :

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \theta} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{vmatrix} = r$$

Also we have $r^2 = x_1^2 + x_2^2$ and $\tan\theta = x_2/x_1$. As x_1 and x_2 range from $-\infty$ to $+\infty$, r varies from 0 to ∞ and θ from 0 to 2π. The joint probability differential of r and θ now becomes $$dG(r, \theta) = \frac{1}{2\pi}\exp(-r^2/2)\,r\,dr\,d\theta;\ 0 \le r \le \infty,\ 0 \le \theta \le 2\pi$$

Integrating over θ, the marginal distribution of r is given by :

$$dG_1(r) = \int_0^{2\pi} dG(r,\theta) = r\exp(-r^2/2)\,dr\left|\frac{\theta}{2\pi}\right|_0^{2\pi} = \exp(-r^2/2)\,r\,dr$$

$$\Rightarrow \quad dG_1(r^2) = \frac{1}{2}\exp(-r^2/2)\,dr^2 = \frac{1}{\Gamma(1)}\exp(-r^2/2)\,(r^2/2)^{1-1}\,d(r^2/2)$$

Thus $\frac{r^2}{2} = \frac{X_1^2 + X_2^2}{2}$ is a $\gamma(1)$ variate and hence $r^2 = X_1^2 + X_2^2$ is a χ^2-variate with 2 d.f.

For n variables X_i, $(i = 1, 2, \ldots, n)$, we transform $(X_1, X_2, \ldots, X_n)$ to $(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1})$; (1 – 1 transformation) by :

$$\left.\begin{aligned} x_1 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-1} \\ x_2 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-2} \sin\theta_{n-1} \\ x_3 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-3} \sin\theta_{n-2} \\ &\vdots \\ x_j &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-j} \sin\theta_{n-j+1} \\ &\vdots \\ x_n &= \chi \sin\theta_1 \end{aligned}\right\} \quad \ldots(15\cdot3)$$

where $\chi > 0$, $-\pi < \theta_1 < \pi$ and $-\frac{1}{2}\pi < \theta_i < \frac{1}{2}\pi$; for $i = 2, 3, \ldots \frac{1}{2}(n-1)$.

Then $x_1^2 + x_2^2 + \ldots + x_n^2 = \chi^2$ and $|J| = \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}$

(*c.f.* Advanced Theory of Statistics Vol. 1, by Kendall and Stuart)

The joint distribution of $X_1, X_2, \ldots, X_n$, *viz.*,

$$dF(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\Sigma x_i^2/2) \prod_{i=1}^{n} dx_i, \text{ transforms to}$$

$$dG(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1}) = \exp\left(-\frac{1}{2}\chi^2\right) \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}\, d\chi\, d\theta_1\, d\theta_2 \ldots d\theta_{n-1}$$

Integrating over $\theta_1, \theta_2, \ldots, \theta_{n-1}$, we get the distribution of χ^2 as :

$$dP(\chi^2) = k \exp(-\chi^2/2)\, (\chi^2)^{(n/2)-1}\, d\chi^2,\ 0 \le \chi^2 < \infty$$

The constant k is determined from the fact that total probability is unity, *i.e.*,

$$\int_0^\infty dP(\chi^2) = 1 \quad \Rightarrow \quad k \int_0^\infty \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1}\, d\chi^2 = 1 \quad \Rightarrow \quad k = \frac{1}{2^{n/2}\,\Gamma(n/2)}$$

$$\therefore \quad dP(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1},\ 0 \le \chi^2 < \infty$$

Hence $\frac{1}{2}\chi^2 = \frac{1}{2}\sum_{i=1}^{n} X_i^2$ is a $\gamma(n/2)$ variate $\Rightarrow$ $\chi^2 = \sum_{i=1}^{n} X_i^2$ is a chi-square variate with n degrees of freedom (*d.f.*) and (15·2) gives *p.d.f.* of chi-square distribution with n *d.f.*

Remarks **1.** If X_i ; $i = 1, 2, \ldots, n$ are n independent normal variates with mean μ_i and S.D. σ_i then $\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ is a χ^2-variate with n *d.f.*

2. In random sampling from a normal population with mean μ and S.D. σ, $\bar{x}$ is distributed normally about the mean μ with S.D. $\sigma/\sqrt{n}$.

$$\therefore \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \Rightarrow \quad \left[\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right]^2 \text{ is a } \chi^2\text{-variate with 1 } d.f.$$

3. Normal distribution is a particular case of χ^2-distribution when $n = 1$, since for $n = 1$,

$$p(\chi^2) = \frac{1}{\sqrt{2}\,\Gamma(1/2)} \exp(-\chi^2/2)\,(\chi^2)^{\frac{1}{2}-1}\,d\chi^2,\ 0 \le \chi^2 < \infty$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-\chi^2/2)\,d\chi,\ -\infty \le \chi < \infty$$

Thus χ is a standard normal variate.

4. For $n = 2$,

$p(\chi^2) = \frac{1}{2}\exp\left(-\frac{1}{2}\chi^2\right),\ \chi^2 \ge 0 \quad \Rightarrow \quad p(x) = \frac{1}{2}\exp\left(\frac{-x}{2}\right),\ x \ge 0$ which is the *p.d.f.* of exponential distribution with mean 2.

15·3. M.G.F. OF CHI-SQUARE DISTRIBUTION

Let $X \sim \chi^2_{(n)}$, then

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)\,dx = \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty e^{tx} \cdot e^{-x/2}\, x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right] \cdot x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \cdot \frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} \qquad \text{[Using Gamma Integral]}$$

$$= (1-2t)^{-n/2},\ |\,2t\,| < 1 \qquad \ldots(15\cdot4)$$

which is the required *m.g.f.* of a χ^2-variate with n *d.f.*

Remarks 1. Using Binomial expansion for negative index, we get from (15·4)

$$M(t) = 1 + \frac{n}{2}(2t) + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)}{2!}(2t)^2 + \ldots + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)}{r!}(2t)^r + \ldots$$

$\therefore$ μ_r' = Coefficient of $\frac{t^r}{r!}$ in the expansion of $M(t)$

$$= 2^r \frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)$$

$$= n(n+2)(n+4)\ldots(n+2r-2) \qquad \ldots(15\cdot4a)$$

2. If n is even so that $n/2$ is a positive integer, then

$$\mu_r' = 2^r\,\Gamma[(n/2)+r]/\Gamma(n/2) \qquad \ldots(15\cdot4b)$$

15·3·1. Cumulant Generating Function of χ^2-Distribution. If $X \sim \chi^2_{(n)}$, then

$$K_X(t) = \log M_X(t) = -\frac{n}{2}\log(1-2t) = \frac{n}{2}\left[2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \ldots\right]$$

$\therefore$ κ_1 = Coefficient of t in $K(t) = n$, κ_2 = Coefficient of $\frac{t^2}{2!}$ in $K(t) = 2n$,

κ_3 = Coefficient of $\frac{t^3}{3!}$ in $K(t) = 8n$, and κ_4 = Coefficient of $\frac{t^4}{4!}$ in $K(t) = 48n$

In general, κ_r = Coefficient of $\frac{t^r}{r!}$ in $K(t) = n\,2^{r-1}(r-1)!$ $\qquad \ldots(15\cdot4c)$

Hence

$$\left.\begin{aligned} \text{Mean} &= \kappa_1 = n, & \text{Variance} &= \mu_2 = \kappa_2 = 2n \\ \mu_3 &= \kappa_3 = 8n, & \mu_4 &= \kappa_4 + 3\kappa_2^2 = 48n + 12n^2 \\ \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n} \quad \text{and} & \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3 \end{aligned}\right\} \quad \ldots(15\cdot4)$$

15·3·2. Limiting Form of χ^2 Distribution for Large Degrees of Freedom. If $X \sim \chi^2_{(n)}$, then $M_X(t) = (1-2t)^{-n/2}$, $|t| < \frac{1}{2}$.

The *m.g.f.* of standard χ^2-variate Z is : $M_{X-\mu/\sigma}(t) = e^{-\mu t/\sigma} M_X(t/\sigma)$

$$\Rightarrow \quad M_Z(t) = e^{-\mu t/\sigma}(1-2t/\sigma)^{-n/2} = e^{-nt/\sqrt{2n}}\left(1-\frac{2t}{\sqrt{2n}}\right)^{-n/2} \qquad (\because \mu = n, \sigma^2 = 2n)$$

$$\therefore \quad K_Z(t) = \log M_Z(t) = -t\sqrt{\frac{n}{2}} - \frac{n}{2}\log\left(1 - t\sqrt{\frac{2}{n}}\right)$$

$$= -t\sqrt{\frac{n}{2}} + \frac{n}{2}\left[t\cdot\sqrt{\frac{2}{n}} + \frac{t^2}{2}\cdot\frac{2}{n} + \frac{t^3}{3}\left(\frac{2}{n}\right)^{3/2} + \ldots\right]$$

$$= -t\sqrt{\frac{n}{2}} + t.\sqrt{\frac{n}{2}} + \frac{t^2}{2} + O(n^{-1/2}) = \frac{t^2}{2} + O(n^{-1/2}),$$

where $O(n^{-1/2})$ are terms containing $n^{1/2}$ and higher powers of n in the denominator.

$$\therefore \lim_{n\to\infty} K_Z(t) = \frac{t^2}{2} \quad \Rightarrow \quad M_Z(t) = e^{t^2/2} \text{ as } n \to \infty,$$

which is the *m.g.f.* of a standard normal variate. Hence, by uniqueness theorem of *m.g.f.* Z is asymptotically normal. In other words, standard χ^2 variate tends to standard normal variate as $n \to \infty$. Thus, χ^2. distribution tends to normal distribution for large *d.f.*

In practice for $n \geq 30$, the χ^2-approximation to normal distribution is fairly good. So whenever $n \geq 30$, we use the normal probability tables for testing the significance of the value of χ^2. That is why in the tables (given on *page 15·56*), the significant values of χ^2 have been tabulated till $n = 30$ only.

Remark. For the distribution of χ^2-variate for large values of n, see Example 15·7 and also Remark 2 to § 15·6·1.

15·3·3. Characteristic Function of χ^2-Distribution. If $X \sim \chi^2_{(n)}$, then

$$\phi_X(t) = E\{\exp(itX)\} = \int_0^\infty \exp(itx)\, f(x)\, dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left\{-\left(\frac{1-2it}{2}\right)x\right\}(x)^{\frac{n}{2}-1}\, dx = (1-2it)^{-n/2} \qquad \ldots(15\cdot4a)$$

15·3·4. Mode and Skewness of χ^2-Distribution. Let $X \sim \chi^2_{(n)}$, so that

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-x/2}\, x^{(n/2)-1},\ 0 \leq x < \infty \qquad \ldots(*)$$

Mode of the distribution is the solution of $f'(x) = 0$ and $f''(x) < 0$.

Logarithmic differentiation *w.r.to* x in (*) gives :

$$\frac{f'(x)}{f(x)} = 0 - \frac{1}{2} + \left(\frac{n}{2} - 1\right)\cdot\frac{1}{x} = \frac{n-2-x}{2x} \qquad \ldots(15\cdot5)$$

Since $f(x) \neq 0$, $f'(x) = 0 \Rightarrow x = n - 2$.

It can be easily seen that at the point, $x = (n - 2)$, $f''(x) < 0$.

Hence mode of the chi-square distribution with n *d.f.* is $(n - 2)$.

Also Karl Pearson's coefficient of skewness is given by :

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{n - (n-2)}{\sqrt{2n}} = \sqrt{\frac{2}{n}} \qquad \ldots(15\cdot6)$$

Since Pearson's coefficient of skewness is greater than zero for $n \geq 1$, the χ^2-distribution is positively skewed. Further since skewness is inversely proportional to the square root of *d.f.*, it rapidly tends to symmetry as the *d.f.* increases.

15·3·5. Additive Property of χ^2-variates. *The sum of independent chi-square variates is also a χ^2-variate. More precisely, if X_i, $(i = 1, 2, \ldots, k)$ are independent χ^2-variates with n_i d.f. respectively, then the sum $\sum_{i=1}^{k} X_i$ is also a chi-square variate with $\sum_{i=1}^{k} n_i$ d.f.*

Proof. We have $M_{X_i}(t) = (1 - 2t)^{-n_i/2}$; $i = 1, 2, \ldots, k$.

The *m.g.f.* of the sum $\sum_{i=1}^{k} X_i$ is given by :

$$M_{\Sigma X_i}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_k}(t) \qquad [\because X_i\text{'s are independent}]$$

$$= (1 - 2t)^{-n_1/2} (1 - 2t)^{-n_2/2} \ldots (1 - 2t)^{-n_k/2} = (1 - 2t)^{-(n_1 + n_2 + \ldots + n_k)/2}$$

which is the *m.g.f.* of a χ^2-variate with $(n_1 + n_2 + \ldots + n_k)$ *d.f.* Hence by uniqueness theorem of *m.g.f.*'s, $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*

Remarks **1.** Converse is also true, *i.e.*, if X_i ; $i = 1, 2, \ldots, k$ are χ^2-variates with n_i ; $i = 1, 2, \ldots, k$ *d.f.* respectively and if $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*, then X_i's are independent.

2. Another useful version of the converse is as follows :

If X and Y are independent non-negative variates such that $X + Y$ follows chi-square distribution with $n_1 + n_2$ *d.f.* and if one of them say X is a χ^2-variate with n_1 *d.f.* then the other, viz., Y, is a χ^2-variate with n_2 *d.f.*

Proof. Since X and Y are independent variates, $M_{X+Y}(t) = M_X(t)\, M_Y(t)$

$$\Rightarrow \quad (1 - 2t)^{-(n_1 + n_2)/2} = (1 - 2t)^{-n_1/2} \cdot M_Y(t) \qquad [\because X + Y \sim \chi^2_{(n_1 + n_2)} \text{ and } X \sim \chi^2_{(n_1)}]$$

$$\therefore \quad M_Y(t) = (1 - 2t)^{-n_2/2},$$

which is the *m.g.f.* of χ^2-variate with n_2 *d.f.* Hence by uniqueness theorem of *m.g.f.*'s, $Y \sim \chi^2_{(n_2)}$

3. Still another form of the above theorem is *"Cochran theorem"* which is as follows :

Let $X_1, X_2, \ldots, X_n$ be independently distributed as standard normal variates, *i.e.*, $N(0, 1)$.

Let $\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k$, where each Q_i is a sum of squares of linear combinations of $X_1, X_2, \ldots, X_n$ with n_i degrees of freedom. Then if $n_1 + n_2 + \ldots + n_k = n$, the quantities $Q_1, Q_2, \ldots, Q_k$ are independent χ^2-variates with $n_1, n_2, \ldots, n_k$ *d.f.* respectively.

15·3·6. Chi-square Probability Curve. We get from (15·5),

$$f'(x) = \left[\frac{n-2-x}{2x}\right] f(x) \qquad \ldots(15\cdot7)$$

Since $x > 0$ and $f(x)$ being *p.d.f.* is always non-negative, we get from (15·7) :

$$f'(x) < 0 \quad \text{if} \quad (n-2) \leq 0,$$

for all values of x. Thus the χ^2-probability curve for 1 and 2 degrees of freedom is monotonically decreasing. When $n > 2$,

$$f'(x) = \begin{cases} > 0, \text{ if } x < (n-2) \\ = 0, \text{ if } x = n-2 \\ < 0, \text{ if } x > (n-2) \end{cases}$$

This implies that for $n > 2$, $f(x)$ is monotonically increasing for $0 < x < (n-2)$ and monotonically decreasing for $(n-2) < x < \infty$, while at $x = n-2$, it attains the maximum value.

For $n \geq 1$, as x increases, $f(x)$ decreases rapidly and finally tends to zero as $x \to \infty$. Thus for $n > 1$, the χ^2-probability curve is positively skewed [*c.f.* (15·6)] towards higher values of x. Moreover, x-axis is an asymptote to the curve. The shape of the curve for $n = 1, 2, 3, \ldots, 6$ is given in *Fig. 15·1*. For $n = 2$, the curve will meet $y = f(x)$ axis at $x = 0$, *i.e.*, at $f(x) = 0{\cdot}5$. For $n = 1$, it will be an inverted J-shaped curve.

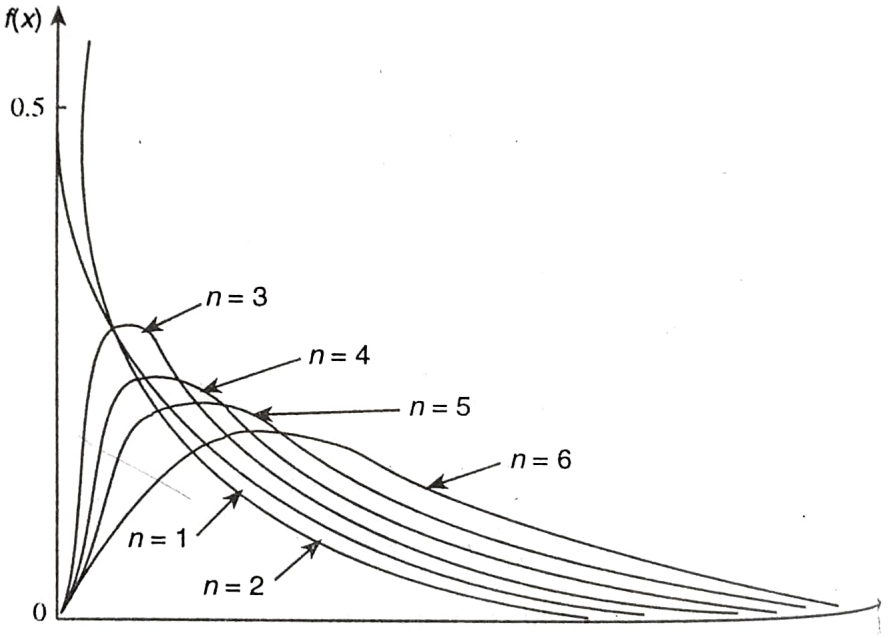


Fig. 15·1: Probability Curve of Chi-square Distribution

15·4. SOME THEOREMS ON CHI-SQUARE DISTRIBUTION

Theorem 15·1. *If X_1 and X_2 are two independent χ^2-variates with n_1 and n_2 d.f. respectively, then* $\dfrac{X_1}{X_2}$ *is a* $\beta_2\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ *variate.*

Proof. Since X_1 and X_2 are independent χ^2 variates with n_1 and n_2 *d.f.* respectively their joint probability differential is given by the compound probability theorem as :

$$dP(x_1, x_2) = dP_1(x_1)\, dP_2(x_2)$$

$$= \left[\frac{1}{2^{n_1/2}\Gamma(n_1/2)} \exp(-x_1/2)\,(x_1)^{(n_1/2)-1} dx_1\right]$$

$$\times \left[\frac{1}{2^{n_2/2}\Gamma(n_2/2)} \exp(-x_2/2)\,(x_2)^{(n_2/2)-1} dx_2\right]$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(x_1+x_2)/2\}$$

$$\times (x_1)^{\frac{n_1}{2}-1}(x_2)^{\frac{n_2}{2}-1}\,dx_1dx_2,\ 0 \le (x_1, x_2) < \infty$$

Let us make the transformation :

$u = x_1/x_2$ and $v = x_2$ so that $x_1 = uv$ and $x_2 = v$.

Jacobian of transformation J is given by : $J = \dfrac{\partial(x_1, x_2)}{\partial(u, v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$

Thus the joint distribution of random variables U and V becomes :

$$dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(1+u)v/2\} \times (uv)^{\frac{n_1}{2}-1} v^{\frac{n_2}{2}-1}\,du\,dv,$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(1+u)v/2\} \times u^{\frac{n_1}{2}-1} v^{\frac{n_1+n_2}{2}-1}\,du\,dv,$$

$$0 \le (u, v) < \infty$$

Integrating *w.r.to* v over the range 0 to ∞, we get marginal distribution of U as :

$$dG_1(u) = \int_0^\infty dG(u, v)$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} u^{(n_1/2)-1}\,du \times \int_0^\infty \exp\left\{-\left(\frac{1+u}{2}\right)v\right\} v^{[(n_1+n_2)/2]-1}dv$$

$$= \frac{u^{(n_1/2)-1}}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma\{(n_1+n_2)/2\}}{[(1+u)/2]^{(n_1+n_2)/2}}\,du$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{u^{(n_1/2)-1}}{[1+u]^{(n_1+n_2)/2}}\,du,\ 0 \le u < \infty$$

Hence $U = \dfrac{X_1}{X_2} \sim \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate.

Theorem 15·2. *If X_1 and X_2 are independent χ^2-variates with n_1 and n_2 d.f. respectively, then $U = \dfrac{X_1}{X_1+X_2}$ and $V = X_1 + X_2$ are independently distributed, U as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and V as a χ^2-variate with $(n_1 + n_2)$ d.f.*

Proof. From *Theorem 15·1*, we have

$$dP(x_1, x_2) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\left\{-(x_1+x_2)/2\right\}$$

$$\times (x_1)^{(n_1/2)-1}(x_2)^{(n_2/2)-1}dx_1\,dx_2,\ 0 \le (x_1, x_2) < \infty$$

Let us transform to u and v defined as follows :

$u = \dfrac{x_1}{x_1+x_2}$ and $v = x_1 + x_2$ so that $x_1 = uv$ and $x_2 = v - x_1 = (1-u)v$

As x_1 and x_2 both range from 0 to ∞, u ranges from 0 to 1 and v from 0 to ∞.

Jacobian of transformation J is : $J = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$

$$\therefore \quad dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp(-v/2) \times (uv)^{(n_1/2)-1} \times [(1-u)v]^{(n_2/2)-1} \left|\begin{matrix} v & u \\ -v & 1-u \end{matrix}\right| du\,dv$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} u^{(n_1/2)-1}(1-u)^{(n_2/2)-1} \times \exp(-v/2) \times v^{(n_1+n_2/2)-1}\,du\,dv$$

$$= \left[\frac{\Gamma\{(n_1+n_2)/2\}}{\Gamma(n_1/2)\Gamma(n_2/2)} u^{(n_1/2)-1}(1-u)^{(n_2/2)-1}du\right] \times \left[\frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp(-v/2)\, v^{\{(n_1+n_2)/2\}-1}\,dv\right]$$

Since the joint probability differential of U and V is the product of their respecti probability differentials, U and V are independently distributed, with

$$dG_1(u) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} u^{(n_1/2)-1}(1-u)^{(n_2/2)-1}\,du, \quad 0 \le u \le 1$$

and

$$dG_2(v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp(-v/2)\, v^{\{(n_1+n_2)/2\}-1}\,dv, \quad 0 \le v < \infty$$

i.e., U as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and V as a χ^2-variate with $(n_1 + n_2)$ *d.f.*

Remark. The results in Theorems 15·1 and 15·2 can be summarised as follows :

If $X \sim \chi^2_{(n_1)}$ and $Y \sim \chi^2_{(n_2)}$ are independent chi-square variates, then

(*i*) $X + Y \sim \chi^2_{(n_1+n_2)}$ *i.e.,* the sum of two independent chi-square variates is also a c square variate.

(*ii*) $\frac{X}{Y} \sim \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$, *i.e.,* the ratio of two independent chi-square variates is a β_2-variat

(*iii*) $\frac{X}{X+Y} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$

Theorem 15·3. *In a random and large sample,* $\chi^2 = \sum_{i=1}^{k} \left[\frac{(n_i - np_i)^2}{np_i}\right]$, ...(15

follows chi-square distribution approximately with (k–1) degrees of freedom, where n_i is observed frequency and np_i is the corresponding expected frequency of the ith class, (i = 1, ..., k), $\sum_{i=1}^{k} n_i = n$.

Proof. Let us consider a random sample of size n, whose members are distribut at random in k classes or cells. Let p_i be the probability that sample observation fall in the ith cell, $(i = 1, 2, \ldots, k)$. Then the probability P of there being n_i member the ith cell, $(i = 1, 2, \ldots, k)$ respectively is given by the multinomial probabil law, by the expression :

$$P = \frac{n!}{n_1!\; n_2!\; \ldots\; n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \text{ where } \sum_{i=1}^{k} n_i = n \text{ and } \sum_{i=1}^{k} p_i = 1.$$

Hence $\sum_{i=1}^{k} \xi_i^2 = \sum_{i=1}^{k} \left[\frac{(n_i - \lambda_i)^2}{\lambda_i}\right]$, being the sum of the squares of k independ standard normal variates is a χ^2-variate with $(k-1)$ *d.f.*, one *d.f.* being lost becau the linear constraint

$$\sum_{i=1}^{k} \xi_i \sqrt{\lambda_i} = \Sigma(n_i - \lambda_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \lambda_i$$

Remarks 1. If O_i and E_i $(i = 1, 2, \ldots, k)$, be a set of observed and expected frequencies, th

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i}\right], \left(\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i\right)$$

follows chi-square distribution with $(k-1)$ *d.f.*

Another convenient form of this formula is as follows :

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i}\right) = \sum_{i=1}^{k} \left(\frac{O_i^2}{E_i} + E_i - 2O_i\right)$$

$$= \sum_{i=1}^{k} (O^2{}_i/E_i) + \sum_{i=1}^{k} E_i - 2\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} (O_i^2/E_i) - N,$$

where $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = N$ (say), is the total frequency.

2. *Conditions for the Validity of χ^2-test.* χ^2-test is an approximate test for large value n. For the validity of chi-square test of '*goodness of fit*' between theory and experiment following conditions must be satisfied :

(*i*) The sample observations should be independent.

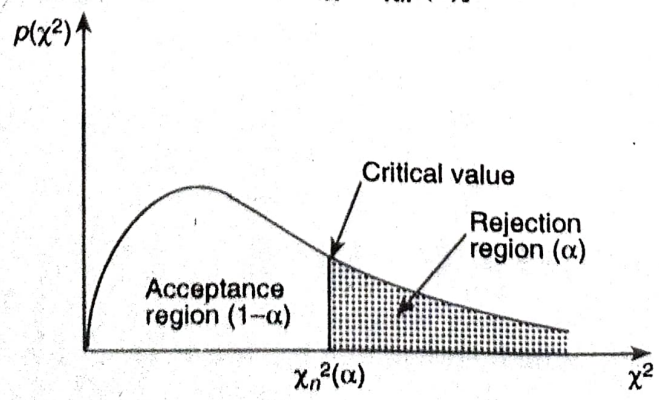
(*ii*) Constraints on the cell frequencies, if any, should be linear, *e.g.*, $\Sigma n_i = \Sigma\lambda_i$ or $\Sigma O_i = \Sigma$

(*iii*) N, the total frequency should be reasonably large, say, greater than 50.

(*iv*) No theoretical cell frequency should be less than 5. (The chi square distributio essentially a continuous distribution but it cannot maintain its character of continuity if frequency is less than 5.) If any theoretical cell frequency is less than 5, then for the applica of χ^2-test, it is pooled with the preceding or succeeding frequency so that the pooled freque is more than 5 and finally adjust for the *d.f.* lost in pooling.

3. It may be noted that the χ^2-test depends only on the set of observed and expec frequencies and on degrees of freedom (*d.f.*). It does not make any assumptions regarding parent population from which the observations are taken. Since χ^2 defined in (15·8) does involve any population parameters, it is termed as a statistic and the test is known as *Parametric Test* or *Distribution-Free Test.*

4. *Critical Values.* Let $\chi_n^2(\alpha)$ denote the value of chi-square for n *d.f.* such that the are the right of this point is α, *i.e.*, $P[\chi^2 > \chi_n^2(\alpha)] = \alpha$



Remark. This approximation is often used for the value of n larger than 30. This result does not reflect anything as to how good the approximation is, for moderate values of n. R. A. Fisher has proved that the approximation is improved by taking $\sqrt{(2n-1)}$ instead of $\sqrt{2n}$.

A still better approximation is $\left(\frac{\chi^2}{n}\right)^{1/3} \sim N\left(1-\frac{2}{9n}, \frac{2}{9n}\right)$.

Example 15·8. *For a chi-square distribution with n d.f. establish the following recurrence relation between the moments : $\mu_{r+1} = 2r(\mu_r + n\mu_{r-1})$, $r \geq 1$. Hence find β_1 and β_2.*

Solution. If $X \sim \chi^2_{(n)}$, then its *m.g.f.* about origin is :

$$M_X(t) = E(e^{tX}) = (1-2t)^{-n/2}\ ;\ t < \frac{1}{2}$$

Also $E(X) = n = \mu$ (say).

Hence *m.g.f.* about mean, say, $M(t)$ is :

$$M(t) = M_{X-\mu}(t) = E[e^{t(X-\mu)}] = e^{-\mu t} \cdot E(e^{tX}) = e^{-nt}(1-2t)^{-n/2}$$

Taking logarithms of both sides, $\log M(t) = -nt - \frac{n}{2}\log(1-2t)$ [Using

Differentiating *w.r. to* t, we have

$$\frac{M'(t)}{M(t)} = -n + \frac{n}{2} \cdot \frac{2}{(1-2t)} = \frac{2nt}{(1-2t)} \quad \Rightarrow \quad (1-2t)\,M'(t) = 2nt\,M(t)$$

Differentiating r times *w.r.to* t by Leibnitz theorem, we get

$$(1-2t)\,M^{r+1}(t) + r(-2)\,M^r(t) = 2nt\,M^r(t) + 2nr\,M^{r-1}(t)$$

Putting $t = 0$ and using the relation, $\mu_r = \left[\frac{d^r}{dt^r}M(t)\right]_{t=0} = M^r(0)$, we get

$$\mu_{r+1} - 2r\mu_r = 2nr\mu_{r-1} \quad \Rightarrow \quad \mu_{r+1} = 2r(\mu_r + n\mu_{r-1}),\ r \geq 1. \qquad \ldots(**)$$

Taking $r = 1, 2, 3$ in (**), we get

$$\mu_2 = 2n\mu_0 = 2n, \qquad \mu_3 = 4(\mu_2 + n\mu_1) = 8n \qquad [\because \mu_1 = 0 \text{ and } \mu_0 =$$

$$\mu_4 = 6(\mu_3 + n\mu_2) = 48n + 12n^2$$

$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{12}{n}.$$

15·6. APPLICATIONS OF CHI-SQUARE DISTRIBUTION

χ^2-distribution has a large number of applications in Statistics, some of which are enumerated below :

(*i*) To test if the hypothetical value of the population variance is $\sigma^2 = \sigma_0^2$ (say).

(*ii*) To test the 'goodness of fit'.

(*iii*) To test the independence of attributes.

(*iv*) To test the homogeneity of independent estimates of the population variance.

(*v*) To combine various probabilities obtained from independent experiments to give a single test of significance.

(*vi*) To test the homogeneity of independent estimates of the population correlation coefficient.

In this section we will introduce various hypothesis -testing procedures based on the use of the chi-square distribution. As with other hypothesis-testing procedures, these tests compare the sample results with those that are expected when the null hypothesis is true. The acceptance or rejection of the null hypothesis is based upon how 'close' the sample or observed results are to the expected results. For detailed discussion on Testing of Hypothesis, see Chapter 17.

15·6·1. Inferences About a Population Variance. Suppose we want to test if a random sample x_i, $(i = 1, 2, \ldots, n)$ has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$ (say).

Under the null hypothesis that the population variance is $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})^2}{\sigma_0^2}\right] = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] = \frac{ns^2}{\sigma_0^2} \qquad \ldots(15\cdot14)$$

follows chi-square distribution with $(n-1)$ *d.f.*

By comparing the calculated value with the tabulated value of χ^2 for $(n-1)$ *d.f.* at certain level of significance (usually 5%), we may retain or reject the null hypothesis.

Remarks **1.** The above test (15·14) can be applied only if the population from which the sample is drawn is normal.

2. If the sample size n is large (>30), then we can use Fisher's approximation

$$\sqrt{2\chi^2} \sim N(\sqrt{2n-1},\ 1), \quad i.e., Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1) \qquad \ldots(15\cdot14a)$$

and apply Normal Test.

Example 15·9. *It is believed that the precision (as measured by the variance) of an instrument is no more than 0·16. Write down the null and alternative hypothesis for testing this belief. Carry out the test at 1% level given 11 measurements of the same subject on the instrument :*

2·5, 2·3, 2·4, 2·3, 2·5, 2·7, 2·5, 2·6, 2·6, 2·7, 2·5.

Solution.

Null Hypothesis,
$H_0 : \sigma^2 = 0{\cdot}16$
Alternative Hypothesis,
$H_1 : \sigma^2 > 0{\cdot}16$

COMPUTATION OF SAMPLE VARIANCE

X	$X - \bar{X}$	$(X - \bar{X})^2$
2·5	− 0·01	0·0001
2·3	− 0·21	0·0441
2·4	− 0·11	0·0121
2·3	− 0·21	0·0441
2·5	− 0·01	0·0001
2·7	+ 0·19	0·0361
2·5	− 0·01	0·0001
2·6	+ 0·09	0·0081
2·6	+ 0·09	0·0081
2·7	+ 0·19	0·0361
2·5	− 0·01	0·0001
$\bar{X} = \frac{27{\cdot}6}{11} = 2{\cdot}51$		$\Sigma(X - \bar{X})^2 = 0{\cdot}1891$

Under the null hypothesis, H_0 : $\sigma^2 = 0{\cdot}16$, the test statistic is :

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma(X-\bar{X})^2}{\sigma^2} = \frac{0{\cdot}1891}{0{\cdot}16} = 1{\cdot}182,$$

which follows χ^2-distribution with *d.f.* $n-1 = (11-1) = 10$.

Since the calculated value of χ^2 is less than the tabulated value 23·2 of χ^2 for 10 d.f. at 1% level of significance, it is not significant. Hence H_0 may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0·16.

Example 15·10. *Test the hypothesis that $\sigma = 10$, given that $s = 15$ for a random sample of size 50 from a normal population.*

Solution. *Null Hypothesis,* $H_0 : \sigma = 10$.

We are given $n = 50$, $s = 15$. Now $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{50 \times 225}{100} = 112{\cdot}5$

Since n is large, using (15·14*a*), the test statistic is : $Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$

$\therefore \quad Z = \sqrt{225} - \sqrt{99} = 15 - 9{\cdot}95 = 5{\cdot}05$

Since $|Z| > 3$, it is significant at all levels of significance and hence H_0 is rejected and we conclude that $\sigma \neq 10$.

15·6·2. Goodness of Fit Test. A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as *"Chi-square test of goodness of fit"*. It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If f_i $(i = 1, 2, \ldots, n)$ is a set of observed (experimental) frequencies and e_i $(i = 1, 2, \ldots, n)$ is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by :

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - e_i)^2}{e_i}\right], \quad \left(\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i\right) \qquad \ldots(15{\cdot}15)$$

follows chi-square distribution with $(n-1)$ *d.f.*

Remark. This is an approximate test for large values of n. Conditions for the validity of the χ^2-test of goodness of fit have already been given in § 15·4 Remark 2 on page 15·12.

The goodness of fit test uses the chi-square distribution to determine if a hypothesized probability distribution for a population provides a good fit. Acceptance or rejection of the hypothesized population distribution is based upon differences between observed frequencies (f_i's) in a sample and the expected frequencies (e_i's) obtained under null hypothesis H_0.

Decision rule :Accept H_0 if $\chi^2 \leq \chi^2_\alpha (n-1)$ and reject H_0 if $\chi^2 > \chi^2_\alpha (n-1)$, where χ^2 is the calculated value of chi-square obtained on using (15·15) and $\chi^2_\alpha (n-1)$ is the tabulated value of chi-square for $(n-1)$ *d.f.* and level of significance α.

Example 15·11. *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 :*

Days	:	*Mon.*	*Tues.*	*Wed.*	*Thurs.*	*Fri.*	*Sat.*
No. of parts demanded	:	*1124*	*1125*	*1110*	*1120*	*1126*	*1115*

Test the hypothesis that the number of parts demanded does not depend on the day of the week. (Given : the values of chi-square significance at 5, 6, 7, d.f. are respectively 11·07 12·59, 14·07 at the 5% level of significance.)

Solution. Here we set up the *null hypothesis*, H_0 that the number of parts demanded does not depend on the day of week.

Under the null hypothesis, the expected frequencies of the spare part demanded on each of the six days would be :

$$\frac{1}{6}(1124 + 1125 + 1110 + 1120 + 1126 + 1115) = \frac{6720}{6} = 1120$$

TABLE 15·2 : CALCULATIONS FOR χ^2

Days	Frequency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
Mon.	1124	1120	16	0·014
Tues.	1125	1120	25	0·022
Wed.	1110	1120	100	0·089
Thurs.	1120	1120	0	0
Fri.	1126	1120	36	0·032
Sat.	1115	1120	25	0·022
Total	6720	6720		0·179

$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 0·179$

The number of degrees of freedom = 6 – 1 = 5 (since we are given 6 frequencies subjected to only one linear constraint : $\sum f_i = \sum e_i = 6720$)

The tabulated $\chi^2_{0·05}$ for 5 *d.f.* = 11·07.

Since calculated value of χ^2 is less than the tabulated value, it is not significant and the null hypothesis may accepted at 5% level of significance. Hence we conclude that the number of parts demanded are same over the 6-day period.

Example 15·12. *The following figures show the distribution of digits in numbers chosen at random from a telephone directory :*

Digits :	0	1	2	3	4	5	6	7	8	9	Total
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory.

Solution. Here we set up the *null hypothesis* that the digits occur equally frequently in the directory.

Under the null hypothesis, the expected frequency for each of the digits 0, 1, 2, ..., 9 is 10,000/10 = 1000. The value of χ^2 is computed as follows :

TABLE 15·3 : CALCULATIONS FOR χ^2

Digits	Frequency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
0	1026	1000	676	0·676
1	1107	1000	11449	11·449
2	997	1000	9	0·009
3	966	1000	1156	1·156
4	1075	1000	5625	5·625
5	933	1000	4489	4·489
6	1107	1000	11149	11·449
7	972	1000	784	0·784
8	964	1000	1296	1·296
9	853	1000	21609	21·609
Total	10,000	10,000		58·542

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$
$$= 58·542$$

The number of degrees of freedom
= Number of observations − Number of independent constraints
= 10 − 1 = 9

Tabulated $\chi^2_{0·05}$ for 9 *d.f.* = 16·919

Since the calculated value of χ^2 is much greater than the tabulated value, it is highly significant and we reject the null hypothesis. Thus we conclude that the digits are not uniformly distributed in the directory.

√ ***Example 15·13.*** *A sample analysis of examination results of 200 MBA's was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in first division. Are these figures commensurate with the general examination result which is in the ratio of 4 : 3 : 2 : 1 for various categories respectively ?*

Solution. Set up the *null hypothesis* that the observed figures do not differ significantly from the hypothetical frequencies which are in the ratio of 4 : 3 : 2 : 1. In other words *the given data are commensurate with the general examination result* which is in the ratio of 4 : 3 : 2 : 1 for the various categories.

Under the null hypothesis, the expected frequencies can be computed as shown in the adjoining table :

Category	Frequency	
	Observed (f_i)	Expected (e_i)
Failed	46	$\frac{4}{10} \times 200 = 80$
III Division	68	$\frac{3}{10} \times 200 = 60$
II Division	62	$\frac{2}{10} \times 200 = 40$
I Division	24	$\frac{1}{10} \times 200 = 20$
Total	200	200

TABLE 15·4 : CALCULATIONS FOR χ^2

Category	Frequency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
Failed	46	80	1156	14·450
III Division	68	60	64	1·067
II Division	62	40	484	12·100
I Division	24	20	16	0·800
Total	200	200		28·417

$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 28·417$

d.f. = 4 − 1 = 3, tabulated $\chi^2_{0·05}$ for 3 *d.f.* = 7·815

Since the calculated value of χ^2 is greater than the tabulated value, it is significant and the null hypothesis is rejected at 5% level of significance. Hence we may conclude that data are not commensurate with the general examination result.

Example 15·14. *A survey of 800 families with four children each revealed the following distribution :*

No. of boys	:	*0*	*1*	*2*	*3*	*4*
No. of girls	:	*4*	*3*	*2*	*1*	*0*
No. of families	:	*32*	*178*	*290*	*236*	*64*

Is this result consistent with the hypothesis that male and female births are equally probable ?

Solution. Let us set up the *null hypothesis* that *the data are consistent with the hypothesis of equal probability for male and female births.* Then under the null hypothesis :

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

$p(r)$= Probability of 'r' male births in a family of 4 = ${}^4C_r \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-r} = {}^4C_r \left(\frac{1}{2}\right)^4$

The frequency of r male births is given by :

$$f(r) = N.\ p(r) = 800 \times {}^4C_r \left(\frac{1}{2}\right)^4 = 50 \times {}^4C_r\ ;\ r = 0, 1, 2, 3, 4. \qquad \ldots (*)$$

Substituting r = 0, 1, 2, 3, 4 successively in (*), we get the expected frequencies as follows :

$f(0) = 50 \times 1 = 50$, $f(1) = 50 \times {}^4C_1 = 200$, $f(2) = 50 \times {}^4C_2 = 300$,

$f(3) = 50 \times {}^4C_3 = 200$, $f(4) = 50 \times {}^4C_4 = 50$.

TABLE 15·5 : CALCULATIONS FOR χ^2

No. of male births	Frequency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
0	32	50	324	6·48
1	178	200	484	2·42
2	290	300	100	0·33
3	236	200	1296	6·48
4	64	50	196	3·92
Total	800	800		19·63

$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$

$= 19·63$

Tabulated $\chi^2_{0·05}$ for $5 - 1$ $= 4$ *d.f.* is 9·488.

Since calculated value of χ^2 is greater than tabulated value, it is significant at 5% level of significance. Hence we reject the null hypothesis and conclude that male and female births are not equally probable.

Example 15·15. *When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows :*

No. of mistakes in a page (x) :	0	1	2	3	4	5	6
No. of pages (f) :	275	72	30	7	5	2	1

Fit a Poisson distribution to the above data and test the of goodness of fit.

Solution. Mean of the given distribution is : $\bar{X} = \frac{1}{N}\Sigma fx = \frac{189}{392} = 0·482$

In order to fit a Poisson distribution to the given data, we take the mean (parameter) m of the Poisson distribution equal to the mean of the given distribution, *i.e.*, we take $m = \bar{X} = 0·482$.

The frequency of r mistakes per page is given by the Poisson law as follows :

$$f(r) = Np(r) = 392 \times \frac{e^{-0·482}\,(0·482)^r}{r\,!}\,;\ r = 0, 1, 2, \ldots, 6$$

Now $f(0) = 392 \times e^{-0·482} = 392 \times \text{Antilog}\,(-0·482 \log_{10} e)$

$= 392 \times \text{Antilog}\,(-0·482 \times \log_{10} 2·7183)$ $(\because e = 2·7183)$

$= 392 \times \text{Antilog}\,(-0·482 \times 0·4343) = 392 \times \text{Antilog}\,(-0·2093)$

$= 392 \times \text{Antilog}\,(\bar{1}·7907) = 392 \times 0·6176 = 242·1$

$f(1) = m \times f(0) = 0·482 \times 242·1 = 116·69$, $f(2) = \frac{m}{2} \times f(1) = 0·241 \times 116·69 = 28·12$

$f(3) = \frac{m}{3} \times f(2) = \frac{0·482}{3} \times 28·12 = 4·518$, $f(4) = \frac{m}{4} \times f(3) = \frac{0·482}{4} \times 4·51 = 0·544$

$f(5) = \frac{m}{5} \times f(4) = \frac{0·482}{5} \times 0·544 = 0·052$, $f(6) = \frac{m}{6} \times f(5) = \frac{0·482}{6} \times 0·052 = 0·004$

Hence the theoretical Poisson frequencies correct to one decimal place are as given below :

X :	0	1	2	3	4	5	6	Total
Expected Frequency :	242·1	116·7	28·1	4·5	0·5	0·1	0	392

TABLE 15·6 : CALCULATIONS FOR χ^2

Mistakes per page (X)	Frequency: Observed (f_i)	Frequency: Expected (e_i)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
0	275	242·1	1082·41	4·471
1	72	116·7	1998·09	17·121
2	30	28·1	3·61	0·128
3	7	4·5		
4	5	0·5		
	15	5·1	98·01	19·217
5	2	0·1		
6	1	0		
Total	392	392		40·937

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 40·937$$

$$d.f. = 7 - 1 - 1 - 3 = 2$$

(One *d.f.* being lost because of the linear constraint $\sum f_i = \sum e_i$; 1 *d.f.* is lost because the parameter m has been estimated from the given data and is then used for computing the expected frequ-encies; 3 *d.f.* are lost because of pooling the last four expected cell frequencies which are less than five.)

Tabulated value of χ^2 for 2 *d.f.* at 5% level of significance is 5·99.

Conclusion. Since calculated value of χ^2 (40·937) is much greater than 5·99, it is highly significant. Hence we conclude that Poisson distribution is not a good fit to the given data.

15·6·3. Test of Independence of Attributes—Contingency Tables. Let us consider two attributes A and B, A divided into r classes $A_1, A_2, \ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Such a classification in which attributes are divided into more than two classes is known as *manifold classification*. The various cell frequencies can be expressed in the following table known as $r \times s$ *manifold contingency table* where (A_i) is the number of persons possessing the attribute A_i, $(i = 1, 2, \ldots, r)$, (B_j) is the number of persons possessing the attribute B_j $(j = 1, 2, \ldots, s)$ and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j, $(i = 1, 2, \ldots, r; j = 1, 2, \ldots, s)$.

Also $\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N$, where N is the total frequency.

TABLE 15·7 : $r \times s$ CONTINGENCY TABLE

A / B	A_1	A_2	...	A_i	...	A_r	Total
B_1	(A_1B_1)	(A_2B_1)	...	(A_iB_1)	...	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	...	(A_iB_2)	...	(A_rB_2)	(B_2)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_j	(A_1B_j)	(A_2B_j)	...	(A_iB_j)	...	(A_rB_j)	(B_j)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_s	(A_1B_s)	(A_2B_s)	...	(A_iB_s)	...	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	...	(A_i)	...	(A_r)	N

TABLE 15.

SINGNIFICANT VALUES $\chi^2(\alpha)$ OF CHI-SQUARE DISTRIBUTION (RIGHT TAIL AREAS) FOR GIVEN PROBABILITY α,

where $P = P_r[\chi^2 > \chi_v^2(\alpha)] = \alpha$

AND v IS DEGREES OF FREEDOM (*d.f.*)

* χ^2-DISTRIBUTION VALUES OF $\chi_v^2(\alpha)$

Degrees of freedom (v)	Probability (α)							
	0·995	0·99	0·995	0·95	0·05	0·025	0·01	0·005
1	0·000	0·000	0·001	0·004	3·841	5·024	6·635	7·879
2	0·010	0·020	0·051	0·103	5·991	7·378	9·210	10·597
3	0·072	0·115	0·216	0·352	7·815	9·348	11·345	12·838
4	0·207	0·297	0·484	0·711	9·488	11·143	13·277	14·860
5	0·412	0·554	0·831	1·145	11·070	12·832	15·086	16·750
6	0·676	0·872	1·237	1·634	12·592	14·449	16·812	18·548
7	0·989	1·239	1·690	2·167	14·067	16·013	18·475	20·278
8	1·344	1·646	2·180	2·733	15·507	17·535	20·090	21·955
9	1·735	2·088	2·700	3·325	16·919	19·023	21·666	23·589
10	2·156	2·558	3·247	3·940	18·307	20·483	23·209	25·188
11	2·603	3·053	3·816	4·575	19·675	21·920	24·725	26·757
12	3·074	3·571	4·404	5·226	21·026	23·337	26·217	28·300
13	3·565	4·107	5·009	5·892	22·362	24·736	24·888	29·819
14	4·075	4·660	5·629	6·571	23·685	26·119	29·141	31·319
15	4·601	5·229	6·262	7·261	24·996	27·488	30·578	32·801
16	5·142	5·812	6·908	7·962	26·296	28·845	32·000	34·267
17	5·697	6·408	7·564	8·672	27·587	30·191	33·409	35·718
18	6·265	7·015	8·231	9·390	28·869	31·526	34·805	37·156
19	6·844	7·633	8·907	10·117	30·144	32·852	36·191	38·582
20	7·434	8·260	9·591	10·851	31·410	34·170	37·566	39·997
21	8·034	8·897	10·283	11·591	32·671	35·479	38·932	41·401
22	8·643	9·542	10·982	12·338	33·924	36·781	40·289	42·796
23	9·260	10·196	11·688	13·091	35·172	38·076	41·638	44·181
24	9·886	10·856	12·401	13·848	36·415	39·364	42·980	45·558
25	10·520	11·524	13·120	14·611	37·652	40·646	44·314	46·928
26	11·160	12·198	13·844	15·379	38·885	41·923	45·642	48·290
27	11·808	12·879	14·573	16·151	40·113	43·194	46·963	49·645
28	12·461	13·565	15·308	16·928	41·337	44·461	48·278	50·993
29	13·121	14·256	16·047	17·708	42·557	45·722	49·588	52·336
30	13·787	14·953	16·791	18·493	43·773	46·979	50·892	53·672
40	20·706	22·164	24·433	26·509	55·759	59·342	63·691	66·766
50	27·991	29·707	32·357	34·764	67·505	71·420	76·154	79·490
60	35·535	37·485	40·482	43·188	79·082	83·298	88·379	91·952
70	43·275	45·442	48·758	51·739	90·531	95·023	100·425	104·215
80	51·172	53·540	57·153	60·391	101·879	106·629	112·329	116·321
90	59·196	61·754	65·647	69·126	113·145	118·136	124·116	128·299
100	67·328	70·065	74·222	77·929	124·342	129·561	135·807	140·169

For larger values of v, quantity $\sqrt{2\chi^2} - \sqrt{2v-1}$ may be used as a standard normal variable.

* *Abridged from Table 8 of Biometrika Tables for Statisticians, Vol. I.*