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 :

$$E\,(Statistic) = \text{Parameter}, \qquad \ldots (14\cdot1)$$

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 :

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

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

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

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

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 : $Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1)$ (*c.f.* § 14·5)

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

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

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

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

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

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 : $H_0 : \mu = \mu_0$, 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

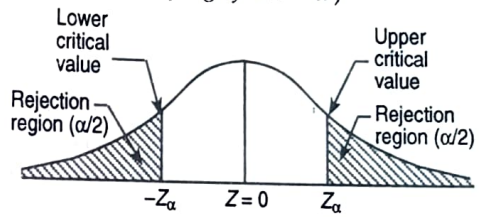
$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \Rightarrow P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha \quad \text{[By symmetry]}$$

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

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')



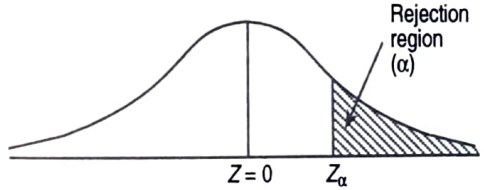
In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $(-z_\alpha)$ is α (See diagrams below), *i.e.*,

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

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

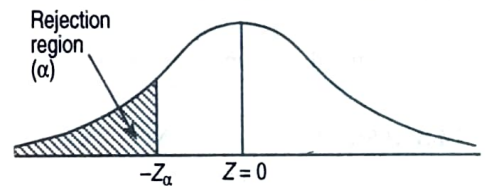
RIGHT-TAILED TEST

(*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 $|Z| \le 3$, 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 $|Z| \le 2{\cdot}58$, H_0 may be accepted at 1% level of significance.

From the normal probability tables, we have :

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

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

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

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

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 :

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

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 :

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

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

P = Probability of success = 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

$\therefore \quad p$ = 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)

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

Now $$Z = \frac{0{\cdot}54 - 0{\cdot}50}{\sqrt{0{\cdot}5 \times 0{\cdot}5/1000}} = \frac{0{\cdot}04}{0{\cdot}0138} = 2{\cdot}532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two-tailed test is 2·58. Since computed $Z = 2{\cdot}532$ is less than 2·58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

Example 14·5. *Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more, at 5% level. (Use Large Sample Test.)*

Solution. In the usual notations, we are given : $n = 20$.

X = Number of persons who survived after attack by a disease = 18

p = Proportion of persons survived in the sample $= \frac{18}{20} = 0{\cdot}90$

Null Hypothesis, $H_0 : P = 0{\cdot}85$, *i.e.,* the proportion of persons survived after attack by a disease in the lot is 85%.

Alternative Hypothesis, $H_1 : P > 0{\cdot}85$ (Right-tailed alternative).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1), \text{ (since sample is large).}$$

Now $$Z = \frac{0{\cdot}90 - 0{\cdot}85}{\sqrt{0{\cdot}85 \times 0{\cdot}15/20}} = \frac{0{\cdot}05}{0{\cdot}079} = 0{\cdot}633$$

Conclusion. Since the alternative hypothesis is one-sided (right-tailed), we shall apply right-tailed test for testing significance of Z. The significant value of Z at 5% level of significance for right-tailed test is + 1·645. Since computed value of $Z = 0{\cdot}633$ is less than 1·645, it is not significant and we may accept the null hypothesis at 5% level of significance.

Example 14·6. *Work sampling studies are conducted to find the utilization of a machine. Out of 200 observations made, only 40 observations indicated the machine to be idle. Find the number of observations to be made in order to satisfy 95% confidence to state the utilization of machine with expected accuracy of ± 5%.*

Solution. The sample size for n estimating the population proportion P with confidence coefficient $(1 - \alpha) = 0{\cdot}95$ is given by the equation :

$$P_r[\,|\,p - P\,|\,\leq 1{\cdot}96\sqrt{PQ/n}\,] = 0{\cdot}95 \qquad \ldots (*)$$

We want $$P_r\,[\,|\,p - P\,|\,< 0{\cdot}05] = 0{\cdot}95. \qquad \ldots (**)$$

From (*) and (**), we obtain $1{\cdot}96\sqrt{PQ/n} = 0{\cdot}05 \Rightarrow n = \frac{PQ\,(1{\cdot}96)^2}{(0{\cdot}05)^2}$

Since P is not known, we may use its sample estimate $\hat{P} = p = \frac{40}{200} = 0{\cdot}2$.

$$\therefore \quad n = \frac{0{\cdot}2 \times (1 - 0{\cdot}2) \times 3{\cdot}8416}{0{\cdot}0025} = \frac{0{\cdot}6147}{0{\cdot}0025} = 245{\cdot}88 \simeq 246.$$

14·7·2. Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by : $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$.

If P_1 and P_2 are population proportions, then

$$E(p_1) = P_1,\ E(p_2) = P_2 \qquad [c.\,f.\ \text{Equation (14·4}a\text{)}]$$

and $$V(p_1) = \frac{P_1 Q_1}{n_1} \quad \text{and} \quad V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are independently and asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by :

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1) \qquad \ldots (*)$$

Under the *null hypothesis,* $H_0 : P_1 = P_2$, *i.e.,* there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0 \qquad (\text{Under } H_0)$$

Also $$V(p_1 - p_2) = V(p_1) + V(p_2),$$

the covariance term Cov (p_1, p_2) vanishes, since sample proportions are independent.

$$\Rightarrow \quad V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

[$\because$ under $H_0 : P_1 = P_2 = P$ (say), and $Q_1 = Q_2 = Q$].

Hence, under $H_0 : P_1 = P_2$, the test statistic for the difference of proportions becomes :

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots (14{\cdot}5)$$

In general, we do not have any information as to the proportion of A's in the populations from which the samples have been taken. Under $H_0 : P_1 = P_2 = P$ (say), an unbiased estimate of the population proportion P, based on both the samples is given by :

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \qquad \ldots (14{\cdot}5a)$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E\left[n_1 p_1 + n_2 p_2\right] = \frac{1}{n_1 + n_2}\left[n_1 E(p_1) + n_2 E(p_2)\right]$$

$$= \frac{1}{n_1 + n_2}(n_1 P_1 + n_2 P_2) = P \qquad [\because P_1 = P_2 = P, \text{ under } H_0]$$

Thus (14·5) along with (14·5a) gives the required test statistic.

Remarks **1.** Suppose we want to test the significance of the difference between p_1 and p, where $p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$ gives a pooled estimate of the population proportion on the basis of both the samples. We have $V(p_1 - p) = V(p_1) + V(p) - 2\,\text{Cov}\,(p_1, p)$.

Since p_1 and p are not independent, Cov $(p_1, p) \neq 0$. $\ldots (*)$

$$\text{Cov}(p_1, p) = E[\{p_1 - E(p_1)\}\{p - E(p)\}]$$

$$= E\left[\{p_1 - E(p_1)\}\left\{\frac{1}{n_1+n_2}\{n_1p_1 + n_2p_2 - E(n_1p_1 + n_2p_2)\}\right\}\right]$$

$$= \frac{1}{n_1+n_2} E\left[\{p_1 - E(p_1)\}\{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\}\right]$$

$$= \frac{1}{n_1+n_2}\left[n_1 E\{p_1 - E(p_1)\}^2 + n_2 E\{(p_1 - E(p_1))(p_2 - E(p_2))\}\right]$$

$$= \frac{1}{n_1+n_2}\left[n_1 V(p_1) + n_2 \text{Cov}(p_1, p_2)\right] = \frac{1}{n_1+n_2} n_1 V(p_1), \quad [\because \text{Cov}(p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1+n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1+n_2}.$$

$$\text{Var}(p) = \text{Var}\left[\frac{n_1p_1 + n_2p_2}{n_1+n_2}\right] = \frac{1}{(n_1+n_2)^2}\text{Var}(n_1p_1 + n_2p_2)$$

$$= \frac{1}{(n_1+n_2)^2}\left[n_1^2 \text{Var}(p_1) + n_2^2 \text{Var}(p_2)\right],$$

covariance term vanishes since p_1 and p_2 are independent.

$$\therefore \quad \text{Var}(p) = \frac{1}{(n_1+n_2)^2}\left(n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2}\right) = \frac{pq}{n_1+n_2}$$

Substituting (*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1+n_2} - 2\frac{pq}{n_1+n_2} = pq\left[\frac{n_2}{n_1(n_1+n_2)}\right]$$

Also $E(P_1 - P) = E(P_1) - E(P) = P - P = 0$

Thus, the test statistic in this case becomes :

$$Z = \frac{(p_1 - p) - E(p_1 - p)}{S.E.(p_1 - p)} = \frac{p_1 - p}{\sqrt{\left\{\frac{n_2}{(n_1+n_2)} \cdot \frac{pq}{n_1}\right\}}} \sim N(0, 1) \qquad \ldots (14{\cdot}5b)$$

2. Suppose the population proportions P_1 and P_2 are given to be distinctly different, *i.e.*, $P_1 \neq P_2$ and we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be hidden in simple samples of sizes n_1 and n_2 from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the *null hypothesis* $H_0 : p_1 = p_2$, *i.e.*, the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}} \sim N(0, 1) \qquad \ldots (14{\cdot}5c)$$

Example 14·7. *Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level.*

Solution. *Null Hypothesis* $H_0 : P_1 = P_2 = P$ (say), *i.e.*, there is no significant difference between the opinions of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given :

$n_1 = 400$, X_1 = Number of men favouring the proposal = 200

$n_2 = 600$, X_2 = Number of women favouring the proposal = 325

$\therefore$ p_1 = Proportion of men favouring the proposal in the sample $= \frac{X_1}{n_1} = \frac{200}{400} = 0{\cdot}5$

p_2 = Proportion of women favouring the proposal in the sample $= \frac{X_2}{n_2} = \frac{325}{600} = 0{\cdot}541$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis, H_0 is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1), \text{ where}$$

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1+n_2} = \frac{X_1 + X_2}{n_1+n_2} = \frac{200+325}{400+600} = 0{\cdot}525 \quad \Rightarrow \quad \hat{Q} = 1 - \hat{P} = 1 - 0{\cdot}525 = 0{\cdot}475$$

$$\therefore \quad Z = \frac{0{\cdot}500 - 0{\cdot}541}{\sqrt{0{\cdot}525 \times 0{\cdot}475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = \frac{-0{\cdot}041}{\sqrt{0{\cdot}001039}} = \frac{-0{\cdot}041}{0{\cdot}0323} = -1{\cdot}269$$

Conclusion. Since $|Z| = 1{\cdot}269$ which is less than 1·96, it is not significant at 5% level of significance. Hence H_0 may be accepted at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

Example 14·8. *In a large city A, 20 per cent of a random sample of 900 school children had defective eye-sight. In other large city B, 15 per cent of random sample of 1,600 children had the same defect. Is this difference between the two proportions significant ? Obtain 95% confidence limits for the difference in the population proportions.*

Solution. In usual notations : $n_1 = 900$, $p_1 = 20\% = 0{\cdot}20$, $n_2 = 1600$, $p_2 = 15\% = 0{\cdot}15$.

Null hypothesis, $H_0 : P_1 = P_2$.

Alternative hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}(p_1 - p_2)} \sim N(0, 1), \text{ (since the samples are large.)}$$

where $\hat{P} = \frac{900 \times 0{\cdot}20 + 1600 \times 0{\cdot}15}{900 + 1600} = 0{\cdot}168 \quad \Rightarrow \quad \hat{Q} = 1 - \hat{P} = 0{\cdot}832$

$$\text{S.E.}(p_1 - p_2) = \sqrt{\left\{0{\cdot}168 \times 0{\cdot}832\left(\frac{1}{900} + \frac{1}{1600}\right)\right\}} = \sqrt{0{\cdot}0002427} = 0{\cdot}0156.$$

$$\therefore \quad Z = \frac{0{\cdot}20 - 0{\cdot}15}{0{\cdot}0156} = 3{\cdot}21$$

Conclusion. Since the calculated value of Z is greater than 1·96, it is significant at 5% level. We, therefore, reject the null hypothesis H_0 and conclude that the difference between the two proportions is significant.

The 95% confidence limits for the difference $P_1 - P_2$ are :

$(p_1 - p_2) \pm 1{\cdot}96$ S.E. of $(p_1 - p_2)$,

where $$\text{S.E. of } (p_1 - p_2) = \sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)} \simeq \sqrt{\left(\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}\right)}$$

$$\simeq \sqrt{\left(\frac{0{\cdot}20 \times 0{\cdot}80}{900} + \frac{0{\cdot}15 \times 0{\cdot}85}{1600}\right)} = 0{\cdot}016.$$

Hence, the 95% confidence limits for $P_1 - P_2$ are :

$(0{\cdot}20 - 0{\cdot}15) \pm 1{\cdot}96\ (0{\cdot}016) = 0{\cdot}05 \pm 0{\cdot}031 = (0{\cdot}019$ and $0{\cdot}081)$.

Example 14·9. *A company has the head office at Kolkata and a branch at Mumbai. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Kolkata, 62% favoured the new plan. At Mumbai out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?*

Solution. In the usual notations, we are given :

$$n_1 = 500,\ p_1 = 0{\cdot}62 \quad \text{and} \quad n_2 = 400,\ p_2 = 1 - 0{\cdot}41 = 0{\cdot}59$$

Null hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference between the two groups in their attitude towards the new plan.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E. } (p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1), \text{ where}$$

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{500 \times 0{\cdot}62 + 400 \times 0{\cdot}59}{500 + 400} = 0{\cdot}607 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}393$$

$$\therefore\ Z = \frac{0{\cdot}62 - 0{\cdot}59}{\sqrt{\left\{0{\cdot}607 \times 0{\cdot}393 \times \left(\frac{1}{500} + \frac{1}{400}\right)\right\}}} = \frac{0{\cdot}03}{\sqrt{0{\cdot}00107}} = \frac{0{\cdot}03}{0{\cdot}0327} = 0{\cdot}917.$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1·96. Thus the critical region consists of all values of $Z \geq 1{\cdot}96$ or $Z \leq -1{\cdot}96$.

Conclusion. Since the calculated value of $|Z| = 0{\cdot}917$ is less than the critical value of Z (1·96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we may conclude that there is no significant difference between the two groups in their attitude towards the new plan.

Example 14·10. *Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty ?*

Solution. In the usual notations, we have $n_1 = 1{,}000$; $n_2 = 1{,}200$.

p_1 = Sample proportion of tea drinkers before increase in excise duty $= \frac{800}{1000} = 0{\cdot}80$

p_2 = Sample proportion of tea drinkers after increase in excise duty $= \frac{800}{1200} = 0{\cdot}67$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the consumption of tea before and after the increase in excise duty.

Alternative Hypothesis, $H_1 : P_1 > P_2$ (Right-tailed alternative).

Test Statistic. Under the null hypothesis, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1) \qquad \text{(Since samples are large)}$$

where $$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \quad \text{and} \quad \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore\ Z = \frac{0{\cdot}80 - 0{\cdot}67}{\sqrt{\left\{\frac{16}{22} \times \frac{6}{22} \times \left(\frac{1}{1000} + \frac{1}{1200}\right)\right\}}} = \frac{0{\cdot}13}{0{\cdot}019} = 6{\cdot}842$$

Conclusion. Since Z is much greater than 1·645 as well as 2·33 (since test is one-tailed), it is highly significant at both 5% and 1% levels of significance. Hence, we reject the null hypothesis H_0 and conclude that there is a significant decrease in the consumption of tea after increase in the excise duty.

Example 14·11. *A cigarette manufacturing firm claims that its brand A of the cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim. (Use 5% level of significance.)*

Solution. In usual notations, we are given :

$$n_1 = 200,\ X_1 = 42 \Rightarrow p_1 = \frac{X_1}{n_1} = \frac{42}{200} = 0{\cdot}21$$

$$n_2 = 100,\ X_2 = 18 \Rightarrow p_2 = \frac{X_2}{n_2} = \frac{18}{100} = 0{\cdot}18$$

We set up the *Null Hypothesis* that 8% difference in the sale of two brands of cigarettes is a valid claim, *i.e.*, $H_0 : P_1 - P_2 = 0{\cdot}08$.

Alternative Hypothesis : $H_1 : P_1 - P_2 \neq 0{\cdot}08$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\left\{\hat{P}\,\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N\,(0, 1) \qquad \text{[since samples are large]}$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{42 + 18}{200 + 100} = \frac{60}{300} = 0{\cdot}20 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}80$

$$\therefore \quad Z = \frac{(0{\cdot}21 - 0{\cdot}18) - 0{\cdot}08}{\sqrt{\left\{0{\cdot}2 \times 0{\cdot}8\left(\frac{1}{200} + \frac{1}{100}\right)\right\}}} = \frac{-0{\cdot}05}{\sqrt{0{\cdot}16 \times 0{\cdot}015}} = \frac{-0{\cdot}05}{0{\cdot}04899} = -1{\cdot}02$$

Since $|Z| = 1{\cdot}02 < 1{\cdot}96$, it is not significant at 5% level of significance. Hence null hypothesis may be retained at 5% level of significance and we may conclude that a difference of 8% in the sale of two brands of cigarettes is a valid claim by the firm.

Example 14·12. *On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30 per cent and the remaining 70 per cent. Consider the first question of this examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is no good at discriminating ability of the type being examined here ?*

Solution. Here, we have

n = Total number of candidates = 200

n_1 = The number of candidates in the upper 30% group = $\frac{30}{100} \times 200 = 60$

n_2 = The number of candidates in the remaining 70% group = $\frac{70}{100} \times 200 = 140$.

X_1 = The number of candidates, with correct answer in the first group = 40

X_2 = The number of candidates, with correct answer in the second group = 80

$$\therefore \quad p_1 = \frac{X_1}{n_1} = \frac{40}{60} = 0{\cdot}6666 \quad \text{and} \quad p_2 = \frac{X_2}{n_2} = \frac{80}{140} = 0{\cdot}5714$$

Null Hypothesis, H_0 : There is no significance difference in the sample proportions, *i.e.*, $P_1 = P_2$. In other words, the first question is not good at discriminating the ability of the type being examined here.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$.

Test Statistic. Under H_0 the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

where $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{40 + 80}{60 + 140} = 0{\cdot}6, \quad \hat{Q} = 1 - \hat{P} = 0{\cdot}4$

$$\therefore \quad Z = \frac{0{\cdot}6666 - 0{\cdot}5714}{\sqrt{\left\{0{\cdot}6 \times 0{\cdot}4\left(\frac{1}{60} + \frac{1}{140}\right)\right\}}} = \frac{0{\cdot}0953}{0{\cdot}0756} = 1{\cdot}258$$

Conclusion. Since $|Z| < 1{\cdot}96$, the data are consistent with the null hypothesis at 5% level of significance. Hence we conclude that the first question is not good enough to distinguish between the ability of the two groups of candidates.

Example 14·13. *In a year there are 956 births in a town A, of which 52·5% were males, while in towns A and B combined, this proportion in a total of 1,406 births was 0·496. Is there any significant difference in the proportion of male births in the two towns ?*

Solution. In usual notations, we are given :

$n_1 = 956$, $n_1 + n_2 = 1{,}406$ or $n_2 = 1{,}406 - 956 = 450$

p_1 = Proportion of males in the sample of town $A = 0{\cdot}525$.

Let p_2 be the proportion of males in the sample (of size n_2) of town B. Then

$\hat{P}$ = Proportion of males in both the samples combined = $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0{\cdot}496$ (Given)

$$\therefore \quad \frac{956 \times 0{\cdot}525 + 450 \times p_2}{1{,}406} = 0{\cdot}496 \quad \Rightarrow \quad p_2 = 0{\cdot}434 \qquad \text{(On simplification).}$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the proportion of male births in the two towns A and B.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{p_1 - p_2}{\sqrt{\left\{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

where $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0{\cdot}496 \Rightarrow \hat{Q} = 1 - \hat{P} = 0{\cdot}504$

$$\therefore \quad Z = \frac{0{\cdot}525 - 0{\cdot}434}{\sqrt{\left\{0{\cdot}496 \times 0{\cdot}504\left(\frac{1}{956} + \frac{1}{450}\right)\right\}}} = \frac{0{\cdot}091}{0{\cdot}027} = 3{\cdot}368$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected, *i.e.*, the data are inconsistent with the hypothesis $P_1 = P_2$ and we conclude that there is significant difference in the proportion of male births in the towns A and B.

Example 14·14. *In two large populations, there are 30 and 25 per cent respectively of blue-eyed people. Is this difference likely to be hidden in samples of 1,200 and 900 respectively from the two populations ?*

Solution. Here, we are given : $n_1 = 1{,}200$, $n_2 = 900$.

P_1 = Proportion of blue-eyed people in the first population = 30% = 0·30

P_2 = Proportion of blue-eyed people in the second population = 25% = 0·25

$$\therefore \quad Q_1 = 1 - P_1 = 0{\cdot}70 \quad \text{and} \quad Q_2 = 1 - P_2 = 0{\cdot}75$$

We set up the *null hypothesis*, H_0 that $p_1 = p_2$, *i.e.*, the sample proportions are equal. In other words, the difference in population proportions is likely to be hidden in sampling.

Test Statistic. Under $H_0 : p_1 = p_2$, the test statistic is :

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\left(\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large.)}$$

$$\therefore \quad |Z| = \frac{0{\cdot}30 - 0{\cdot}25}{\sqrt{\left(\frac{0{\cdot}3 \times 0{\cdot}7}{1{,}200} + \frac{0{\cdot}25 \times 0{\cdot}75}{900}\right)}} = \frac{0{\cdot}05}{0{\cdot}0195} = 2{\cdot}56$$

Conclusion. Since $|Z| > 1{\cdot}96$, the null hypothesis $(p_1 = p_2)$ is refuted at 5% level of significance and we conclude that the difference in population proportions is unlikely to be hidden in sampling. In other words, the samples will reveal the difference in the population proportions.

Example 14·15. *In a random sample of 400 students of the university teaching departments, it was found that 300 students failed in the examination. In another random sample of 500 students of the affiliated colleges, the number of failures in the same examination was found to be 300. Find out whether the proportion of failures in the university teaching departments is significantly greater than the proportion of failures in the university teaching departments and affiliated colleges taken together.*

Solution. Here we are given : $n_1 = 400$, $n_2 = 500$, $p_1 = \frac{300}{400} = 0{\cdot}75$, $p_2 = \frac{300}{500} = 0{\cdot}60$

$\therefore$ $q_1 = 1 - p_1 = 1 - 0{\cdot}75 = 0{\cdot}25$ and $q_2 = 1 - p_2 = 1 - 0{\cdot}60 = 0{\cdot}40$

Here we set up the *null hypothesis*, H_0 that p_1 and $p = \hat{P}$, where $\hat{P}$ is the pooled estimate, *i.e.*, proportion of failures in the university teaching departments and affiliated colleges taken together, do not differ singificantly.

$$\text{S.E. of } (p - p_1) = \sqrt{\left(\frac{pq}{n_1 + n_2} \times \frac{n_2}{n_1}\right)} \qquad [c.f.\ (14{\cdot}5b)] \qquad \ldots (*)$$

where $p = \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 \times 0{\cdot}75 + 500 \times 0{\cdot}60}{400 + 500} = 0{\cdot}67 \Rightarrow q = 1 - 0{\cdot}67 = 0{\cdot}33$

$\therefore$ S.E. of $(p - p_1) = \sqrt{\left(\frac{0{\cdot}67 \times 0{\cdot}33}{400 + 500} \times \frac{500}{400}\right)} = 0{\cdot}018$ [Using (*)]

Test Statistic. Under the null hypothesis, H_0, the test statistic is :

$$Z = \frac{p - p_1}{\text{S.E. of } (p - p_1)} \sim N(0, 1), \qquad \text{(Since samples are large.)}$$

$\therefore$ $Z = \frac{0{\cdot}67 - 0{\cdot}75}{0{\cdot}018} = \frac{-0{\cdot}08}{0{\cdot}018} = 4{\cdot}08$

Conclusion. Since the calculated value of $|Z|$ is greater than 3, it is significant. Hence the null hypothesis H_0 is rejected and we conclude that there is significant difference between p_1 and $p = \hat{P}$.

Example 14·16. *If for one-half of n events, the chance of success is p and the chance of failure is q, while for the other half the chance of success is q and the chance of failure is p, show that the standard deviation of the number of successes is the same as if the chance of successes were p in all the cases, i.e., $\sqrt{npq}$ but that the mean of the number of successes is n/2 and not np.*

Solution. Let X_1 and X_2 denote the number of successes in the first half and the second half of n events respectively. Then according to the given conditions, we have

$$\left.\begin{aligned} E(X_1) &= \tfrac{n}{2}p \\ V(X_1) &= \tfrac{n}{2}pq \end{aligned}\right\} \quad \text{and} \quad \left.\begin{aligned} E(X_2) &= \tfrac{n}{2}q \\ V(X_2) &= \tfrac{n}{2}qp \end{aligned}\right\}$$

The mean and variance of the number of successes in all the n events are given by :

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \tfrac{n}{2}p + \tfrac{n}{2}q = \tfrac{n}{2} \qquad (\because p + q = 1)$$

and $V(X_1 + X_2) = V(X_1) + (X_2) = \tfrac{n}{2}pq + \tfrac{n}{2}qp = npq,$

since the first and second half of events are independent.

Hence the variance is the same as if the probability of success in all the n events is p.

14·8. SAMPLING OF VARIABLES

In the present section we will discuss in detail the sampling of variables such as height, weight, age, income, etc. In the case of sampling of variables each member of the population provides the value of the variable and the aggregate of these values forms the frequency distribution of the population. From the population, a random sample of size n can be drawn by any of the sampling methods discussed earlier, which is same as choosing n values of the given variable from the distribution.

14·8·1. Unbiased Estimate for Population Mean (μ) and Variance (σ^2). Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a large population $X_1, X_2, \ldots, X_N$ (of size N) with mean μ and variance σ^2. Then the sample mean ($\bar{x}$) and variance (s^2) are given by : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Now $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i)$

Since x_i is a sample observation from the population X_i, $(i = 1, 2, \ldots, N)$ it can take any one of the values $X_1, X_2, \ldots, X_N$ each with equal probability $1/N$.

$\therefore$ $E(x_i) = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \ldots + \frac{1}{N}X_N = \frac{1}{N}(X_1 + X_2 + \ldots + X_N) = \mu$ ...(1)

Hence $E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mu) = \frac{1}{n}n\mu \Rightarrow E(\bar{x}) = \mu$ ...(14·6)

Thus, *the sample mean ($\bar{x}$) is an unbiased estimate of the population mean (μ).*

Now $E(s^2) = E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) - E(\bar{x}^2)$ ...(2)

We have $V(x_i) = E[x_i - E(x_i)]^2 = E(x_i - \mu)^2$, [From (1)]

$= \frac{1}{N}\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2\right] = \sigma^2$ ...(3)

Also $V(x) = E(x^2) - [E(x)]^2 \Rightarrow E(x^2) = V(x) + [E(x)]^2$ ...(4)

In particular $E(x_i^2) = V(x_i) + [E(x_i)]^2 = \sigma^2 + \mu^2$ ...(5)

Also from (4), we obtain $E(\bar{x}^2) = V(\bar{x}) + [E(\bar{x})]^2$

But $V(\bar{x}) = \frac{\sigma^2}{n}$, where σ^2 is the population variance. [*c.f.* § 14·8·2]

$\therefore$ $E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$ [Using (14·6)] ...(5*a*)

Substituting from (5) and (5a) in (2) we get

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{1}{n}n(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \left(1-\frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2 \qquad \ldots(14\cdot7)$$

Since $E(s^2) \neq \sigma^2$, *sample variance is not an unbiased estimate of population variance.*

From (14·7), we get $\quad \frac{n}{n-1}E(s^2) = \sigma^2 \cdot \Rightarrow \quad E\left(\frac{ns^2}{n-1}\right) = \sigma^2$

$$\Rightarrow \quad E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} = \sigma^2, \quad i.e., \quad E(S^2) = \sigma^2 \qquad \ldots(14\cdot8)$$

where $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \ldots(14\cdot8a)$

$\therefore$ *S^2 is an unbiased estimate of the population variance σ^2.*

Aliter for $E(s^2)$.

$$s^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n}\left[\sum_{i=1}^{n}\{(x_i-\mu)-(\bar{x}-\mu)\}^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + n(\bar{x}-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)\right]$$

But $\quad \sum_i (x_i-\mu) = \sum_i x_i - n\mu = n\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\therefore \quad s^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}V(x_i) - V(\bar{x}) = \left(1-\frac{1}{n}\right)\sigma^2$$

***Remarks* 1.** Here we see that although sample mean is an unbiased estimate of population mean, sample variance is not an unbiased estimate of population variance. However, an unbiased estimate of σ^2 is given by S^2, defined in equation (14·8a).

S^2 plays a very important role in sampling theory, particularly in small sampling theory. Whenever σ^2 is not known, its estimate S^2 given by (14·8a) is used for practical purposes.

2. We have $\quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

$$\therefore \quad ns^2 = (n-1)S^2 \quad \Rightarrow \quad s^2 = \left(1-\frac{1}{n}\right)S^2$$

Hence for large samples, *i.e.,* for $n \to \infty$, we have $s^2 \to S^2$.

In other words, for large samples (*i.e.*, $n \to \infty$), we may take $\hat{\sigma}^2 = s^2$ $\qquad \ldots(14\cdot8b)$

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 : $\quad \bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$

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

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

the covariance terms vanish since the sample observations are independent.

But $\quad V(x_i) = \sigma^2, (i = 1, 2, \ldots, n)$ $\qquad$ [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 (14·8b)]. $\hat{\sigma}^2 = s^2 \quad \Rightarrow \quad \hat{\sigma} = s$ (for large samples).

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

$$|Z| \leq 1{\cdot}96, \quad i.e., \quad \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96 \Rightarrow \bar{x} - 1{\cdot}96\,(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + 1{\cdot}96\,(\sigma/\sqrt{n}) \qquad \ldots(14\cdot10)$$

and $\bar{x} \pm 1{\cdot}96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ are $\bar{x} \pm 2{\cdot}58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2{\cdot}33\sigma/\sqrt{n}$.

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

$$\bar{x} \pm 1{\cdot}96\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{and} \quad \bar{x} \pm 2{\cdot}58\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \ldots(14\cdot10a)$$

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

Example 14·17. *A sample of 900 members has a mean 3·4 cms. and s.d. 261 cms. Is the sample from a large population of mean 3·25 cms. and s.d. 2·61 cms. ?*

If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of true mean.

Solution. *Null Hypothesis,* (H_0) : The sample has been drawn from the population with mean $\mu = 3{\cdot}25$ cms. and S.D. $\sigma = 2{\cdot}61$ cms.

Alternative Hypothesis, $H_1 : \mu \neq 3{\cdot}25$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is : $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, (Since n is large.)

Here, we are given : $\bar{x} = 3{\cdot}4$ cms., $n = 900$ cms., $\mu = 3{\cdot}25$ cms. and $\sigma = 2{\cdot}61$ cms.

$$\therefore \qquad Z = \frac{3{\cdot}40 - 3{\cdot}25}{2{\cdot}61/\sqrt{900}} = \frac{0{\cdot}15 \times 30}{2{\cdot}61} = 1{\cdot}73$$

Since $|Z| < 1{\cdot}96$, we conclude that the data don't provide us any evidence against the null hypothesis (H_0) which may, therefore, be accepted at 5% level of significance.

95% *fiducial limits* for the population mean μ are :

$$\bar{x} \pm 1{\cdot}96\,(\sigma/\sqrt{n}) = 3{\cdot}40 \pm 1{\cdot}96\,(2{\cdot}61/\sqrt{900}) = 3{\cdot}40 \pm 0{\cdot}1705, \textit{ i.e.,}\ 3{\cdot}5705 \text{ and } 3{\cdot}2295$$

98% fiducial limits for μ are given by :

$$\bar{x} \pm 2{\cdot}33\frac{\sigma}{\sqrt{n}} = 3{\cdot}40 \pm 2{\cdot}33 \times \frac{2{\cdot}61}{30} = 3{\cdot}40 \pm 0{\cdot}2027, \textit{ i.e.,}\ 3{\cdot}6027 \text{ and } 3{\cdot}1973$$

Remark. 2·33 is the value z_1 of Z from standard normal probability integrals, such that

$$P(|Z| > z_1) = 0{\cdot}98 \quad \Rightarrow \quad P(Z > z_1) = 0{\cdot}49$$

Example 14·18. *An insurance agent has claimed that the average age of policy-holders who insure through him is less than the average for all agents, which is 30·5 years.*

A random sample of 100 policy-holders who had insured through him gave the following age distribution :

Age last birthday	:	*16—20*	*21—25*	*26—30*	*31—35*	*36—40*
No. of persons	:	12	22	20	30	16

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance. You are given that $Z(1{\cdot}645) = 0{\cdot}95$.

Solution. *Null Hypothesis,* $H_0 : \mu = 30{\cdot}5$ years, *i.e.,* the sample mean $(\bar{x})$ and population mean (μ) do not differ significantly.

Alternative Hypothesis, $H_1 : \mu < 30{\cdot}5$ years (Left-tailed alternative).

CALCULATIONS FOR SAMPLE MEAN AND S.D.

Age last birthday	*No. of persons (f)*	*Mid-point x*	$d = \dfrac{x - 28}{5}$	fd	fd^2
16—20	12	18	−2	−24	48
21—25	22	23	−1	−22	22
26—30	20	28	0	0	0
31—35	30	33	1	30	30
36—40	16	38	2	32	64
Total	$N = 100$			$\Sigma fd = 16$	$\Sigma fd^2 = 164$

$$\bar{x} = 28 + \frac{5 \times 16}{100} = 28{\cdot}8 \text{ years}; \quad s = 5 \times \sqrt{\left\{\frac{164}{100} - \left(\frac{16}{100}\right)^2\right\}} = 6{\cdot}35 \text{ years}$$

Since the sample is large, $\hat{\sigma} \simeq s = 6{\cdot}35$ years.

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim N(0, 1), \qquad \text{(Since sample is large.)}$$

Now
$$Z = \frac{28{\cdot}8 - 30{\cdot}5}{\sqrt{(6{\cdot}35)^2/100}} = \frac{-1{\cdot}7}{0{\cdot}635} = -2{\cdot}681.$$

Conclusion. Since computed value of $Z = -2{\cdot}681 < -1{\cdot}645$ or $|Z| = 2{\cdot}681 > 1{\cdot}645$, it is significant at 5% level of significance. Hence, we reject the null hypothesis, H_0 (Accept H_1) at 5% level of significance and conclude that the insurance agent's claim, that the average age of policy-holders who insure through him is less than the average for all agents, is valid.

Example 14·19. *As an application of Central Limit Theorem, show that if E is such that* $P(|\bar{X} - \mu| < E) > 0{\cdot}95$, *then the minimum sample size n is given by* $n = \{(1{\cdot}96)^2 \sigma^2/E^2\}$ *where* μ *and* σ^2 *are the mean and variance respectively of the population and* $\bar{X}$ *is the mean of the random sample.*

Solution. By Central Limit Theorem, we know that $\bar{x} \sim N(\mu, \sigma^2/n)$ asymptotically, *i.e.,* for large n.

$$\therefore \qquad Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ asymptotically, } \textit{i.e.,} \text{ for large } n.$$

From normal probability tables, we have $P(|Z| \leq 1{\cdot}96) = 0{\cdot}95$

$$\Rightarrow \quad P\left\{\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96\right\} = 0{\cdot}95 \quad \text{or} \quad P\left[|\bar{x} - \mu| \leq 1{\cdot}96\frac{\sigma}{\sqrt{n}}\right] = 0{\cdot}95 \qquad \ldots(*)$$

We are given that $P[|\bar{X} - \mu| < E] > 0{\cdot}95$ $\qquad \ldots(**)$

From (*) and (**), we have $\quad E > \dfrac{1{\cdot}96\sigma}{\sqrt{n}} \quad \Rightarrow \quad n > \dfrac{(1{\cdot}96)^2\sigma^2}{E^2} = \dfrac{3{\cdot}84\sigma^2}{E^2}$.

Hence minimum sample size n for estimating μ with 95% confidence coefficient is given by $n = 3{\cdot}84\,\sigma^2/E^2$, where E is the permissible error.

Remark. The minimum sample size for estimating μ with confidence coefficient $(1-\alpha)$ is given by $\sigma^2 z_\alpha{}^2/E^2$, where z_α is the significant value of Z at level of significance α and E is the permissible error in the estimate.

Arguing similarly, the minimum sample size for estimating population proportion P with confidence coefficient $(1-\alpha)$ is given by $n = PQ\, z_\alpha{}^2/E^2$, where z_α is the significant value of Z at 'α' level of significance and E is the permissible error in the estimate. If P is unknown, we may use $\hat{P} = p$.

Example 14·20. *The mean muscular endurance score of a random sample of 60 subjects was found to be 145 with a s.d. of 40. Construct a 95% confidence interval for the true mean. Assume the sample size to be large enough for normal approximation. What size of sample is required to estimate the mean within 5 of the true mean with a 95% confidence ?*

Solution. In usual notations, we are given : $n = 60$, $\bar{x} = 145$ and $s = 40$.

95% confidence limits for true mean (μ) are :

$\bar{x} \pm 1{\cdot}96\, s/\sqrt{n}$ ($\sigma^2 = s^2$, since sample is large)

$$= 145 \pm \frac{1{\cdot}96 \times 40}{\sqrt{60}} = 145 \pm \frac{78{\cdot}4}{7{\cdot}75} = 145 \pm 10{\cdot}12 = 134{\cdot}88 \quad \text{and} \quad 155{\cdot}12$$

Hence 95% confidence interval for μ is (134·88, 155·12). In the notations of Example 14·19, we have

$$n = \left(\frac{z_\alpha \cdot \sigma}{E}\right)^2 = \left(\frac{1{\cdot}96 \times 40}{5}\right)^2 = (15{\cdot}68)^2 = 245{\cdot}86 \simeq 246.$$

$[\because z_{0{\cdot}05} = 1{\cdot}96,\ \hat{\sigma} = s = 40$ and $|\bar{x} - \mu| < 5 = E]$

Example 14·21. *The standard deviation of a population is 2·70 cms. Find the probability that in a random sample of size 66 (i) the sample mean will differ from the population mean by 0·75 cm. or more, and (ii) the sample mean will exceed the population mean by 0·75 cm. or more (given that the value of the standard normal probability integral from 0 to 2·25 is 0·4877).*

Solution. Here we are given $n = 66$, $\sigma = 2{\cdot}70$ cms. Since n is large, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$.

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

We want ...(*)

(*i*) $P[|\bar{x} - \mu| \geq 0{\cdot}75] = 1 - P[|\bar{x} - \mu| < 0{\cdot}75]$

$$= 1 - P\left(\left|\frac{\sigma Z}{\sqrt{n}}\right| < 0{\cdot}75\right) \qquad \text{[From (*)]}$$

$$= 1 - P\left(|Z| < 0{\cdot}75\,\frac{\sqrt{n}}{\sigma}\right) = 1 - 2\,P\left(0 < Z < 0{\cdot}75\,\frac{\sqrt{n}}{\sigma}\right)$$

$$= 1 - 2\,P\left(0 < Z < 0{\cdot}75\,\frac{\sqrt{66}}{2{\cdot}70}\right) = 1 - 2\,P\left(0 < Z < \frac{0{\cdot}75 \times 8{\cdot}124}{2{\cdot}70}\right)$$

$$= 1 - 2\,P(0 < Z < 2{\cdot}25) = 1 - 2 \times 0{\cdot}4877 = 0{\cdot}0246$$

(*ii*) $P(\bar{x} - \mu > 0{\cdot}75) = P(Z > 0{\cdot}75\sqrt{n}/\sigma) = P(Z > 2{\cdot}25)$

$$= 0{\cdot}5 - P(0 < Z < 2{\cdot}25) = 0{\cdot}5 - 0{\cdot}4877 = 0{\cdot}0123.$$

Example 14·22. *A normal population has a mean of 0·1 and standard deviation of 2·1. Find the probability that mean of a sample of size 900 will be negative.*

Solution. Here we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 0{\cdot}1$ and $\sigma = 2{\cdot}1$ and $n = 900$. Since $X \sim N(\mu, \sigma^2)$, the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. The standard normal variate corresponding to $\bar{x}$ is given by :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0{\cdot}1}{2{\cdot}1/30} = \frac{\bar{x} - 0{\cdot}1}{0{\cdot}07} \quad \Rightarrow \quad \bar{x} = 0{\cdot}1 + 0{\cdot}07Z, \text{ where } Z \sim N(0, 1)$$

The required probability p, that the sample mean is negative is given by :

$$p = P(\bar{x} < 0) = P(0{\cdot}1 + 0{\cdot}07Z < 0) = P\left(Z < \frac{-0{\cdot}10}{0{\cdot}07}\right)$$

$$= P(Z < -1{\cdot}43) = P(Z \geq 1{\cdot}43) = 0{\cdot}5 - P(0 < Z < 1{\cdot}43) = 0{\cdot}5 - 0{\cdot}4236 = 0{\cdot}0764.$$

(From Normal Probability Tables)

Example 14·23. *The guaranteed average life of a certain type of electric light bulbs is 1,000 hours with a standard deviation of 125 hours. It is decided to sample the output so as to ensure that 90 per cent of the bulbs do not fall short of the guaranteed average by more than 2·5 per cent. What must be the minimum size of the sample ?*

Solution. Here $\mu = 1{,}000$ hours, $\sigma = 125$ hours.

Since we do not want the sample mean to be less than the guaranteed average mean ($\mu = 1{,}000$) by more than 2·5%, we should have

$$\bar{x} > 1{,}000 - 2{\cdot}5\% \text{ of } 1{,}000 \ \Rightarrow \ \bar{x} > 1{,}000 - 25 = 975$$

Let n be the given sample size. Then

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \text{ since sample is large.}$$

We want $\quad Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \dfrac{975 - 1{,}000}{125/\sqrt{n}} = -\dfrac{\sqrt{n}}{5} \qquad (\because \bar{x} > 975)$

According to the given condition : $P(Z > -\sqrt{n}/5) = 0{\cdot}90 \ \Rightarrow \ P(0 < Z < \sqrt{n}/5) = 0{\cdot}40$

$$\Rightarrow \qquad \frac{\sqrt{n}}{5} = 1{\cdot}28 \qquad \text{(From Normal Probability Tables)}$$

$$\therefore \qquad n = 25 \times (1{\cdot}28)^2 = 41 \text{ (approx.)}$$

Example 14·24. *A survey is proposed to be conducted to know the annual earnings of the old Statistics graduates of Delhi University. How large should the sample be taken in order to estimate the mean monthly earnings within plus and minus Rs. 10,000 at 95% confidence level ? The standard deviation of the annual earnings of the entire population is known to be Rs. 30,000.*

Solution. We are given : σ = Rs. 30,000.

We want : $\quad P[|\bar{x} - \mu| < 10{,}000] = 0{\cdot}95 \qquad ...(*)$

We know that, in sampling from normal population or for large samples from any population, $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence from Normal Probability Tables, we have

$$P[|Z| \leq 1{\cdot}96] = 0{\cdot}95 \ \Rightarrow \ P\left[\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| \leq 1{\cdot}96\right] = 0{\cdot}95$$

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

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.

3. If $\sigma_1^2\neq\sigma_2^2$ and σ_1 and σ_2 are not known, then they are estimated from sample values. This results in some error, which is practically immaterial, if samples are large. These estimates for large samples are given by : $\hat{\sigma}_1^2=S_1^2\simeq s_1^2$ and $\hat{\sigma}_2^2=S_2^2\simeq s_2^2$. (Since samples are large.)

In this case, (14·11) gives $\quad Z=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}}\sim N(0,\,1) \qquad \ldots[14{\cdot}11(c)]$

Example 14·25. *The means of two single large samples of 1,000 and 2,000 members are 67·5 inches and 68·0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches ? (Test at 5% level of significance.)*

Solution. In usual notations, we are given :

$$n_1=1{,}000,\ n_2=2{,}000,\ \bar{x}_1=67{\cdot}5\text{ inches},\ \bar{x}_2=68{\cdot}0\text{ inches}.$$

Null hypothesis, $H_0:\mu_1=\mu_2$ and $\sigma=2{\cdot}5$ inches, *i.e.*, the samples have been drawn from the same population of standard deviation 2·5 inches.

Alternative Hypothesis, $H_1:\mu_1\neq\mu_2$ (Two-tailed)

Test Statistic. Under H_0, the test statistic is :

$$Z=\frac{\bar{x}-\bar{x}_2}{\sqrt{\left\{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}}}\sim N(0,\,1) \qquad \text{(since samples are large)}$$

Now $\quad Z=\dfrac{67{\cdot}5-68{\cdot}0}{2{\cdot}5\times\sqrt{\left(\frac{1}{1000}+\frac{1}{2000}\right)}}=\dfrac{-0{\cdot}5}{2{\cdot}5\times 0{\cdot}0387}=-5{\cdot}1.$

Conclusion. Since $|\,Z\,|>3$, the value is highly significant and we reject the null hypothesis and conclude that samples are certainly not from the same population with standard deviation 2·5.

Example 14·26. *In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.*

Solution. In the usual notations, we are given that

$$n_1=400, \qquad \bar{x}_1=\text{Rs. }250, \qquad s_1=\text{Rs. }40$$

$$n_2=400, \qquad \bar{x}_2=\text{Rs. }220, \qquad s_2=\text{Rs. }55$$

Null hypothesis, $H_0:\mu_1=\mu_2$, *i.e.*, the average weekly food expenditures of the two populations of shoppers are equal.

Alternative Hypothesis, $H_1:\mu_1\neq\mu_2$. (Two-tailed)

Test Statistic. Since samples are large, under H_0, the test statistic is :

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}\sim N(0,\,1)$$

Since σ_1 and σ_2, the population standard deviations are not known, we can take for large samples $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$ and then Z is given by :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{250 - 220}{\sqrt{\left\{\frac{(40)^2}{400} + \frac{(55)^2}{400}\right\}}} = 8{\cdot}82 \text{ (approx.)}$$

Conclusion. Since $|Z|$ is much greater than 2·58, the null hypothesis ($\mu_1 = \mu_2$) is rejected at 1% level of significance and we conclude that the average weekly expenditures of two populations of shoppers in markets *A* and *B* differ significantly.

Example 14·27. *The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2·56 with a standard deviation of Rs. 1·08. The average hourly wage of a sample of 200 workers in plant 'B' was Rs. 2·87 with a standard deviation of Rs. 1·28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A' ?*

Solution. Let X_1 and X_2 denote the hourly wages (in Rs.) of workers in plant *A* and plant *B* respectively. Then, in usual notations we are given :

$$\left.\begin{aligned} n_1 &= 150, \quad \bar{x}_1 = 2{\cdot}56, \quad s_1 = 1{\cdot}08 = \hat{\sigma}_1 \\ n_2 &= 200, \quad \bar{x}_2 = 2{\cdot}87, \quad s_2 = 1{\cdot}28 = \hat{\sigma}_2 \end{aligned}\right\} \text{(Since samples are large.)}$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* there is no significant difference between the mean level of wages of workers in plant *A* and plant *B*.

Alternative Hypothesis, $H_1 : \mu_2 > \mu_1$ *i.e.,* $\mu_1 < \mu_2$ (Left-tailed test)

Test Statistic. Under H_0, the test statistic (for large samples) is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \quad Z = \frac{2{\cdot}56 - 2{\cdot}87}{\sqrt{\left\{\frac{(1{\cdot}08)^2}{150} + \frac{(1{\cdot}28)^2}{200}\right\}}} = \frac{-0{\cdot}31}{\sqrt{0{\cdot}016}} = \frac{-0{\cdot}31}{0{\cdot}126} = -2{\cdot}46.$$

Critical region. For a one-tailed test, the critical value of Z at 5% level of significance is 1·645. The critical region for left-tailed test thus consists of all values of $Z \leq -1{\cdot}645$.

Conclusion. Since calculated value of Z (−2·46) is less than critical value (−1·645), it is significant at 5% level of significance. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the average hourly wages paid by plant *'B'* are certainly higher than those paid by plant *'A'*.

Example 14·28. *In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 ozs. with a standard deviation of 12 ozs. while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Obtain the standard error of difference between the two sample means. Is this difference significant ? Also find the 99% confidence limits for the difference in the average weights of items produced by the two processes respectively.*

Solution. In usual notations, we are given :

$$\left.\begin{aligned} n_1 &= 250, \quad \bar{x}_1 = 120 \text{ oz.}, \quad s_1 = 12 \text{ oz.} = \hat{\sigma}_1 \\ n_2 &= 400, \quad \bar{x}_2 = 124 \text{ oz.}, \quad s_2 = 14 \text{ oz.} = \hat{\sigma}_1 \end{aligned}\right\} \text{, (Since samples are large.)}$$

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \simeq \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} \quad \text{(Since samles are large.)}$$

$$= \sqrt{\left(\frac{144}{250} + \frac{196}{400}\right)} \simeq \sqrt{0{\cdot}576 + 0{\cdot}490} = 1{\cdot}034$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.,* the sample means do not differ significantly.

Alternative Hypothesis, $H_1 : \mu_1 \neq \mu_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{S.E.(\bar{x}_1 - \bar{x}_2)} = \frac{120 - 124}{1{\cdot}034} \sim N(0, 1)$$

$$\therefore \quad |Z| = \frac{4}{1{\cdot}034} = 3{\cdot}87$$

Conclusion. Since $|Z| > 3$, the null hypothesis is rejected and we conclude that there is significant difference between the sample means.

99% confidence limits for $|\mu_1 - \mu_2|$, *i.e.,* for the difference in the average weights of items produced by two processes, are :

$|\bar{x}_1 - \bar{x}_2| \pm 2{\cdot}58\ S.E.(\bar{x}_1 - \bar{x}_2) = 4 \pm 2{\cdot}58 \times 1{\cdot}034 = 4 \pm 2{\cdot}67$ (approx.) = 6·67 and 1·33

$$\therefore \quad 1{\cdot}33 < |\mu_1 - \mu_2| < 6{\cdot}67.$$

Example 14·29. *The mean height of 50 male students who showed above average participation in college athletics was 68·2 inches with a standard deviation of 2·5 inches; while 50 male students who showed no interest in such participation had a mean height of 67·5 inches with a standard deviation of 2·8 inches.*

(i) Test the hypothesis that male students who participate in college athletics are taller than other male students.

(ii) By how much should the sample size of each of the two groups be increased in order that the observed difference of 0·7 inches in the mean heights be significant at the 5% level of significance.

Solution. Let X_1 and X_2 denote the height (in inches) of athletic participants and non-athletic participants respectively. In the usual notations, we are given :

$$n_1 = 50, \ \bar{x}_1 = 68{\cdot}2, \ s_1 = 2{\cdot}5; \ n_2 = 50, \ \bar{x}_2 = 67{\cdot}5, \ s_2 = 2{\cdot}8$$

Null Hypothesis, $H_0 : \mu_1 = \mu_2$.

Alternative Hypothesis, $H_1 : \mu_1 > \mu_2$ (Right-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$\therefore \quad Z = \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{50} + \frac{(2{\cdot}8)^2}{50}\right\}}} = \frac{0{\cdot}7}{\sqrt{0{\cdot}282}} = \frac{0{\cdot}7}{0{\cdot}53} = 1{\cdot}32$$

For a right-tailed test, the critical (significant) value of Z at 5% level of significance is 1·645.

(*i*) Since the calculated value of $Z(1{\cdot}32)$ is less than the critical value (1·645), it is not significant at 5% level of significance. Hence the null hypothesis is accepted and we conclude that the college athletes are not taller than other male students.

(*ii*) The difference between the mean heights of two groups, each of size n will be significant at 5% level of significance if $Z \geq 1{\cdot}645$

$$\Rightarrow \frac{68{\cdot}2 - 67{\cdot}5}{\sqrt{\left\{\frac{(2{\cdot}5)^2}{n} + \frac{(2{\cdot}8)^2}{n}\right\}}} \geq 1{\cdot}645 \quad \text{or} \quad \frac{0{\cdot}7}{\sqrt{14{\cdot}09/n}} \geq 1{\cdot}645, \text{ i.e., } \frac{0{\cdot}7}{3{\cdot}754/\sqrt{n}} \geq 1{\cdot}645$$

$$\therefore \quad n \geq \left(\frac{1{\cdot}645 \times 3{\cdot}754}{0{\cdot}7}\right)^2 = (8{\cdot}8219)^2 = 77{\cdot}83 \simeq 78$$

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

14·8·5. Test of Significance for the Difference of Standard Deviations. If s_1 and s_2 are the standard deviations of two independent samples, then under null 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 : $\quad S.E.\,(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$

$$\therefore \quad Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)}} \sim N(0, 1), \qquad \ldots(14{\cdot}12)$$

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

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}\right)}} \sim N(0, 1) \text{ (for large samples)} \qquad \ldots(14{\cdot}13)$$

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 ?

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,

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

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

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

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

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

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

$$\left.\begin{aligned} &\text{Mean} = \kappa_1 = n, \quad \text{Variance} = \mu_2 = \kappa_2 = 2n \\ &\mu_3 = \kappa_3 = 8n, \quad \mu_4 = \kappa_4 + 3\kappa_2^2 = 48n + 12n^2 \\ &\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3 \end{aligned}\right\} \quad \ldots(15\cdot4d)$$

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

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

$$= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty \exp\left\{-\left(\frac{1-2it}{2}\right)x\right\} (x)^{\frac{n}{2}-1}\, dx = (1-2it)^{-n/2} \qquad \ldots(15\cdot4e)$$

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 :

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

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, $M_{X+Y}(t) = M_X(t)\, M_Y(t)$

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

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

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),

$$f'(x) = \left[\frac{n-2-x}{2x}\right] f(x) \qquad \ldots(15·7)$$

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 $f(x) = 0·5$.For $n = 1$, it will be an inverted J-shaped curve.

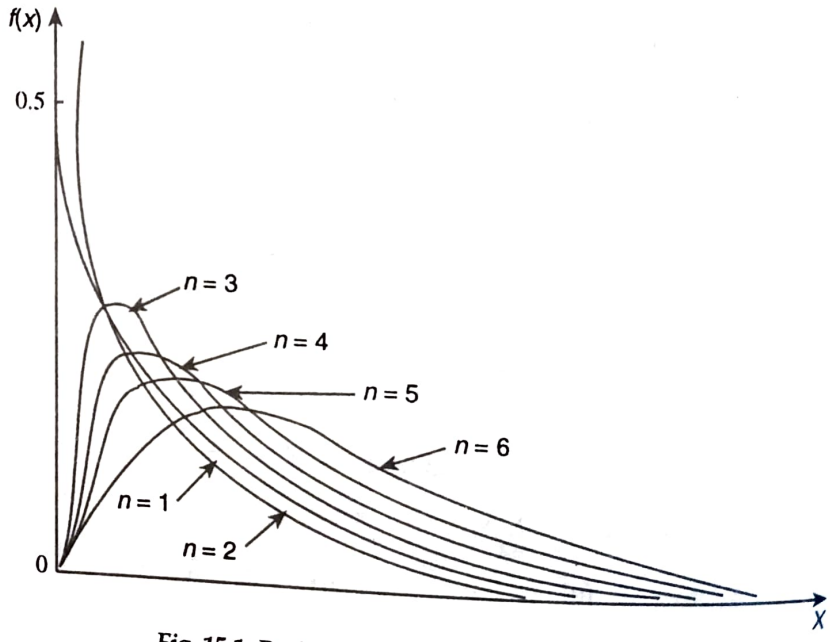


Fig. 15·1: Probability Curve of Chi-square Distribution

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* $\frac{X_1}{X_2}$ *is a* $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ *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 :

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

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

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

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

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

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

The tabulated $\chi^2_{0{\cdot}05}$ 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 Observed (f_i)	Frequency Expected (e_i)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
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		
5	2	0·1	98·01	19·217
6	1	0		
Total	392	392		40·937

(15 and 5·1 group the last four expected cell frequencies)

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

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

(One *d.f.* being lost because of the linear constraint $\Sigma f_i = \Sigma e_i$; 1 *d.f.* is lost because the parameter m has been estimated from the given data and is then used for computing the expected frequencies; 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

The problem is to test if the two attributes A and B under consideration are independent or not.

Under the *null hypothesis that the attributes are independent*, the theoretical cell frequencies are calculated as follows :

$P[A_i]$ = Probability that a person possesses the attribute $A_i = \frac{(A_i)}{N}$; $i = 1, 2, \ldots, r$

$P[B_j]$ = Probability that a person possesses the attribute $B_j = \frac{(B_j)}{N}$; $j = 1, 2, \ldots, s$

$P[A_iB_j]$ = Probability that a person possesses the attributes A_i and $B_j = P(A_i)P(B_j)$

(By compound probability theorem, since the attributes A_i and B_j are independent, under the null hypothesis.)

$$\therefore \qquad P[A_iB_j] = \frac{(A_i)}{N} \cdot \frac{(B_j)}{N}; i = 1, 2, \ldots, r; j = 1, 2, \ldots, s \text{ and}$$

$(A_iB_j)_0$ = Expected number of persons possessing both the attributes A_i and B_j

$$= N.P[A_iB_j] = \frac{(A_i)(B_j)}{N}$$

$$\Rightarrow \qquad (A_iB_j)_0 = \frac{(A_i)(B_j)}{N}, (i = 1, 2, \ldots, r; j = 1, 2, \ldots, s) \qquad \ldots (15\cdot16)$$

By using this formula, we can find out expected frequencies for each of the cell-frequencies (A_iB_j) $(i = 1, 2, \ldots, r; j = 1, 2, \ldots, s)$, under the null hypothesis of independence of attributes.

The exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples, (large N), by the χ^2-test of goodness of fit, *viz.*,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\{(A_iB_j) - (A_iB_j)_0\}^2}{(A_iB_j)_0}\right] = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \qquad \ldots(15\cdot16a)$$

where f_{ij} = observed frequency for contingency table category in column i and row j,

e_{ij} = expected frequency for contingency table category in column i and row j,

which is distributed as a χ^2-variate with $(r-1)(s-1)$ *d.f.* [*c.f.* Note below on degrees of freedom].

Remarks 1. $\phi^2 = \chi^2/N$ is known as *mean-square contingency*.

Since the limits for χ^2 and ϕ^2 vary in different cases, they cannot be used for establishing the closeness of the relationship between qualitative characters under study. Prof. Karl Pearson suggested another measure, known as *"coefficient of mean square contingency"* which is denoted by C and is given by :

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\phi^2}{1 + \phi^2}} \qquad \ldots(15\cdot17)$$

Obviously C is always less than unity. The maximum value of C depends on r and s, the number of classes into which A and B are divided. In a $r \times r$ contingency table, the maximum value of $C = \sqrt{(r-1/r)}$. Since the maximum value of C differs for different classification, *viz.*, $r \times r$ $(r = 2, 3, 4, \ldots)$, strictly speaking, the values of C obtained from different types of classifications are not comparable.

2. Note on Degrees of Freedom (d.f.). The number of independent variates which make up the statistic (*e.g.*, χ^2) is known as the *degrees of freedom* (*d.f.*) and is usually denoted by υ (the letter 'Nu' of the Greek alphabet).

The number of degrees of freedom, in general, is the total number of observations less the number of independent constraints imposed on the observations. For example, if k is the number of independent constraints in a set of data of n observations then $v = (n - k)$.

Thus in a set of n observations usually, the degrees of freedom for χ^2 are $(n-1)$, one *d.f.* being lost because of the linear constraint $\sum_i f_i = \sum_i e_i = N$, on the frequencies (*c.f.* Theorem 15·3). If 'r' independent linear constraints are imposed on the cell frequencies, then the *d.f.* are reduced by 'r'.

In addition, if any of the population parameter(s) is (are) calculated from the given data and used for computing the expected frequencies then in applying χ^2-test of goodness of fit, we have to subtract one *d.f.* for each parameter calculated. Thus if 's' is the number of population parameters estimated from the sample observations (n in number), then the required number of degrees of freedom for χ^2-test is $(n - s - 1)$.

If any one or more of the theoretical frequencies is less than 5, then in applying χ^2-test we have also to subtract the degrees of freedom lost in pooling these frequencies with the preceding or succeeding frequency (or frequencies).

In a $r \times s$ contingency table, in calculating the expected frequencies, the row totals, the column totals and the grand totals remain fixed. The fixation of 'r' column totals and 's' row totals imposes $(r + s)$ constraints on the cell frequencies. But since $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$, the total number of independent constraints is only $(r + s - 1)$. Further, since the total number of cell-frequencies is $r \times s$, the required number of *d.f.* is : $v = rs - (r + s - 1) = (r - 1)(s - 1)$ ...(5·17*a*)

***Example* 15·16.** *Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the adjoining table. Examine whether the nature of the area is related to voting preference in this election.*

Area	*Votes for*		*Total*
	A	*B*	
Rural	*620*	*380*	*1000*
Urban	*550*	*450*	*1000*
Total	*1170*	*830*	*2000*

Solution. Under the *null hypothesis* that the nature of the area is independent of the voting preference in the election, we get the expected frequencies as follows :

$$E(620) = \frac{1170 \times 1000}{2000} = 585, \qquad E(380) = \frac{830 \times 1000}{2000} = 415,$$

$$E(550) = \frac{1170 \times 1000}{2000} = 585, \quad \text{and} \quad E(450) = \frac{830 \times 1000}{2000} = 415$$

Aliter. In a 2×2 contingency table, since *d.f.* $= (2-1)(2-1) = 1$, only one of the cell frequencies can be filled up independently and the remaining will follow immediately, since the observed and theoretical marginal totals are fixed. Thus having obtained any one of the theoretical frequencies (say) $E(620) = 585$, the remaining theoretical frequencies can be easily obtained as follows :

$E(380) = 1000 - 585 = 415$, $E(550) = 1170 - 585 = 585$, and $E(450) = 1000 - 585 = 415$.

$$\therefore \qquad \chi^2 = \sum_i \left[\frac{(f_i - e_i)^2}{e_i}\right] = \frac{(620-585)^2}{585} + \frac{(380-415)^2}{415} + \frac{(550-585)^2}{585} + \frac{(450-415)^2}{415}$$

$$= (35)^2\left(\frac{1}{585}+\frac{1}{415}+\frac{1}{585}+\frac{1}{415}\right) = (1225)[2\times 0{\cdot}002409 + 2\times 0{\cdot}001709] = 10{\cdot}0891$$

Tabulated $\chi^2_{0{\cdot}05}$ for $(2-1)(2-1) = 1$ *d.f.* is 3·841. Since calculated χ^2 is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

Example 15·17. (2 × 2 CONTINGENCY TABLE). *For the 2 × 2 table,*

a	b
c	d

, prove that chi-square test of independence gives

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)},\ N = a+b+c+d \qquad \ldots(15{\cdot}18)$$

Solution. Under the hypothesis of independence of attributes,

$$E(a) = \frac{(a+b)(a+c)}{N}$$

$$E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N}$$

and $$E(d) = \frac{(b+d)(c+d)}{N}$$

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

$$\therefore\quad \chi^2 = \frac{[a-E(a)]^2}{E(a)} + \frac{[b-E(b)]^2}{E(b)} + \frac{[c-E(c)]^2}{E(c)} + \frac{[d-E(d)]^2}{E(d)} \qquad \ldots(*)$$

$$a - E(a) = a - \frac{(a+b)(a+c)}{N} = \frac{a(a+b+c+d) - (a^2+ac+ab+bc)}{N} = \frac{ad-bc}{N}$$

Similarly, we will get : $\quad b - E(b) = -\frac{ad-bc}{N} = c - E(c)\ ; \quad d - E(d) = \frac{ad-bc}{N}$

Substituting in (*), we get

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)}+\frac{1}{E(b)}+\frac{1}{E(c)}+\frac{1}{E(d)}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\left\{\frac{1}{(a+b)(a+c)}+\frac{1}{(a+b)(b+d)}\right\}+\left\{\frac{1}{(a+c)(c+d)}+\frac{1}{(b+d)(c+d)}\right\}\right]$$

$$= \frac{(ad-bc)^2}{N}\left[\frac{b+d+a+c}{(a+b)(a+c)(b+d)}+\frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right]$$

$$= (ad-bc)^2\left[\frac{c+d+a+b}{(a+b)(a+c)(b+d)(c+d)}\right] = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Remark. We can calculate the value of χ^2 for 2 × 2 contingency table by using (15·18) directly. The reader is advised to obtain the value of χ^2 in *Example* 15·16 by using (15·18).

Example 15·18. *Out of 8,000 graduates in a town 800 are females, out of 1,600 graduate employees 120 are females. Use χ^2 to determine if any distinction is made in appointment on the basis of sex. Value of χ^2 at 5% level for one degree of freedom is 3·84.*

Solution. We set up the *Null hypothesis* that no distinction is made in appointment on the basis of sex, and test it against the *Alternative hypothesis* that distinction is made in appointment on the basis of sex.

The observed and expected frequencies are shown in the following table :

TABLE NO. OBSERVED FREQUENCIES EXPECTED FREQUENCIES

	Employed	Not employed	Total	Employed	Not employed	Total
Male	1480	5720	7200	$\frac{7200\times 1600}{8000}$ = 1440	7200 − 1440 = 5760	7200
Female	120	680	800	1600 − 1440 = 160	6400 − 5760 = 640	800
Total	1600	6400	8000	1600	6400	8000

TABLE 15·8 : CALCULATIONS FOR χ^2

Class	Frequency		$(f_i - e_i)$	$\frac{(f_i-e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
Male employed	1480	1440	40	$\frac{1600}{1440}$ = 1·11
Male unemployed	5720	5760	− 40	$\frac{1600}{5760}$ = 0·28
Female employed	120	160	− 40	$\frac{1600}{160}$ = 10·00
Female unemployed	680	640	40	$\frac{1600}{640}$ = 2·50

$\chi^2 = \sum\frac{(f_i-e_i)^2}{e_i}$ = 13·89

d.f. = (2 − 1) (2 − 2) = 1

Tabulated $\chi^2_{0{\cdot}05}$ for 1 *d.f.* = 3·841.

Conclusion. Since the calculated value of χ^2 (13·89) is much greater than the tabulated value of χ^2 (3·841), the value of χ^2 is highly significant and null hypothesis is rejected. Hence we conclude that distinction is made in appointment on the basis of sex.

Example 15·19. *A random sample of students of XYZ University was selected and asked their opinions about 'autonomous colleges'. The results are given below. The same number of each sex was included within each class-group. Test the hypothesis at 5% level that opinions are independent of the class groupings :*

Class	Numbers		Total
	Favouring 'autonomous colleges'	Opposed to 'autonomous colleges'	
B.A./B.Sc./B.Com. Part I	120	80	200
B.A./B.Sc./B.Com. Part II	130	70	200
B.A./B.Sc./B.Com. Part III	70	30	100
M.A./M.Sc./M.Com.	80	20	100
Total	400	200	600

Solution. We set up the null hypothesis that the opinions about autonomous colleges are independent of the class-groupings.

Here the frequencies are arranged in the form of a 4×2 contingency table. Hence the *d.f.* are $(4-1) \times (2-1) = 3 \times 1 = 3$. Hence we need to compute independently only three expected frequencies and the remaining expected frequencies can be obtained by subtraction from the row and column totals.

Under the null hypothesis of independence :

$$E(120) = \frac{400 \times 200}{600} = 133{\cdot}33 \; ; \; E(130) = \frac{400 \times 200}{600} = 133{\cdot}33 : E(70) = \frac{400 \times 100}{600} = 66{\cdot}67$$

Now the table of expected frequencies can be completed as shown below :

Class	Numbers			Total
	Favouring 'autonomous colleges'	Opposed to 'autonomous colleges'		
B.A./B.Sc./B.Com. Part I	133·33	200 – 133·33	= 66·67	200
B.A./B.Sc./B.Com. Part II	133·33	200 – 133·33	= 66·67	200
B.A./B.Sc./B.Com. Part III	66·67	100 – 66·67	= 33·33	100
M.A./M.Sc./M.Com.	66·67	100 – 66·67	= 33·33	100
Total	400		200	600

TABLE 15·9 : CALCULATIONS FOR CHI-SQUARE

f_i	e_i	$f_i - e_i$	$(f_i - e_i)^2$	$(f_i - e_i)^2 / e_i$
120	113·33	– 13·33	177·6889	1·3327
130	133·33	– 3·33	11·0889	0·0832
70	66·67	3·33	11·0889	0·1663
80	66·67	13·33	177·6889	2·6652
80	66·67	13·33	177·6889	2·6652
70	66·67	3·33	11·0889	0·1663
30	33·33	–3·33	11·0889	0·3327
20	33·33	– 13·33	177·6889	5·3312
Total 400	400			12·7428

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 12{\cdot}7428$$

Tabulated value of χ^2 for $(4-1) \times (2-1) = 3$ *d.f.* at 5% level of significance is 7·815.

Conclusion. Since calculated value of χ^2 is greater than the tabulated value, it is significant at 5% level of significance and we reject the null hypothesis. Hence, we conclude that the opinions about autonomous colleges are dependent on the class-groupings.

***Example* 15·20.** *Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows :*

Researcher	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different ? (Given 5% value of χ^2 for 2 d.f. and 3 d.f. are 5·991 and 7·82 respectively.)

Solution. We set up the *null hypothesis* that the data obtained are independent of the sampling techniques adopted by the two researchers. In other words, the null *hypothesis* is that there is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

Here we have a 4×2 contingency table and *d.f.* $= (4-1) \times (2-1) = 3 \times 1 = 3$. Hence we need to compute only 3 independent expected frequencies and the remaining expected frequencies can be obtained by subtraction from the marginal row and column totals.

Under the null hypothesis of independence, we have

$$E(86) = \frac{126 \times 200}{300} = 84; \quad E(60) = \frac{93 \times 200}{300} = 62 \; ; \quad E(44) = \frac{69 \times 200}{300} = 46.$$

The table of expected frequencies can now be completed as shown below :

Researchers	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	84	62	46	200 –192 = 8	200
Y	126 – 84 = 42	93 – 62 = 31	69 – 46 = 23	12 – 8 = 4	100
Total	126	93	69	12	300

Since we cannot apply the χ^2-test straightway here as the last expected frequency is less than 5, we should use the technique of pooling in this case as given below :

TABLE 15·10 : COMPUTATION OF THE VALUE OF χ^2

Researchers	Type of Students	f_i	e_i	$f_i - e_i$	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
X	Below average	86	84	2	4	0·048
	Average	60	62	– 2	4	0·064
	Above average	44	46	– 2	4	0·087
	Genius	10	8	2	4	0·500
Y	Below average	40	42	– 2	4	0·095
	Average	33	31	2	4	0·129
	Above average	25 } 27	23 } 27	0	0	0
	Genius	2 }	4 }			
Total		300	300			0·923

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = 0{\cdot}923$$

and the *d.f.* $= (4-1) \times (2-1) - 1 = 3 - 1 = 2$, since 1 *d.f.* is lost in the method of pooling. Tabulated value of χ^2 for 2 *d.f.* at 5% level of significance is 5·991.

Conclusion. Since calculated value is less than the tabulated value, null hypothesis may be accepted at 5% level of significance and we may conclude that there is no significant difference in the sampling techniques used by the two researchers.

15·6·4. Yate's Correction. In a 2×2 contingency table, the number of *d.f.* is $(2-1)(2-1) = 1$. If any one of the theoretical cell frequencies is less than 5, then use of pooling method for χ^2-test results in χ^2 with 0 *d.f.* (since 1 *d.f.* is lost in pooling) which is meaningless. In this case we apply a correction due to F. Yates (1934), which is usually known as *"Yate's Correction for Continuity"* [As already pointed out, χ^2 is a continuous distribution and it fails to maintain its character of continuity if any of the expected frequency is less than 5; hence the name 'Correction for Continuity'.] This consists in adding 0·5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. The χ^2-test of goodness of fit is then applied without pooling method.

For a 2×2 contingency table,

a	b
c	d

, we have $\chi^2 = \dfrac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$

According to Yate's correction, as explained above, we subtract (or add) $\frac{1}{2}$ from a and d and add (subtract) $\frac{1}{2}$ to b and c so that the marginal totals are not disturbed at all. Thus, corrected value of χ^2 is given as :

$$\chi^2 = \frac{N\left[(a \mp \frac{1}{2})(d \mp \frac{1}{2}) - (b \pm \frac{1}{2})(c \pm \frac{1}{2})\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$\text{Numerator} = N\left[(ad-bc) \mp \frac{1}{2}(a+b+c+d)\right]^2 = N\left[\,|\,ad-bc\,|\, - \frac{N}{2}\right]^2$$

$$\therefore \quad \chi^2 = \frac{N\,[\,|\,ad-bc\,|\, - N/2]^2}{(a+c)(b+d)(a+b)(c+d)} \qquad \ldots (15\cdot18a)$$

Remarks **1.** If N is large, the use of Yate's correction will make very little difference in the value of χ^2. If, however, N is small, the application of Yate's correction may overstate the probability.

2. It is recommended by many authors and it seems quite logical in the light of the above discussion that Yate's correction be applied to every 2×2 table, even if no theoretical cell frequency is less than 5.

15·6·5. Brandt and Snedecor Formula for $2 \times k$ Contingency Table. Let the observations a_{ij}, $(i = 1, 2;\ j = 1, 2, \ldots, k)$ be arranged in a $2 \times k$ contingency table as follows :

A / B	A_1	A_2	.....	A_i	.....	A_k	Total
B_1	a_{11}	a_{12}	.....	a_{1i}	.....	a_{1k}	m_1
B_2	a_{21}	a_{22}	.....	a_{2i}	.....	a_{2k}	m_2
Total	n_1	n_2	.....	n_i	.....	n_k	N

Under the hypothesis of independence of attributes, we have