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 ($\overline{x}$), 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 $\overline{x}$ 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 :

$$E\,(\textit{Statistic}) = \text{Parameter}, \qquad \text{... (14·1)}$$

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

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

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

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

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

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

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

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

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	$1·36263\ \sigma/\sqrt{n}$
6.	Sample median	$1·25331\ \sigma/\sqrt{n}$
7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment : μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment : μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^3}{10^4}} \simeq \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample $s.d.$'s : $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions : $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

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

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

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

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.

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

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.*

If we write P {Reject H_0 when it is true} = P {Reject $H_0 \mid H_0$} = α

and P { Accept H_0 when it is wrong} = P {Accept $H_0 \mid H_1$} = β ... (14·2)

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.

$$\left.\begin{array}{lll}\text{Thus} & P\,\{\text{Reject a lot when it is good}\} & = \alpha \\ \text{and} & P\,\{\text{Accept a lot when it is bad}\} & = \beta\end{array}\right\} \quad \text{... (14·2}a\text{)}$$

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 $t = t\,(x_1, x_2, ..., x_n)$ is the value of the statistic based on a random sample of size n, then

$$P\,(t \in \omega \mid H_0) = \alpha, \quad P\,(t \in \bar{\omega} \mid H_1) = \beta \quad \text{...(14·2}b\text{)}$$

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

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

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

14·4·4. One-tailed and Two-tailed Tests. In any test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

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

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

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

A test of statistical hypothesis where the alternative hypothesis is two-tailed such as : $H_0 : \mu = \mu_0$, 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 :

$$P(|Z| > z_\alpha) = \alpha \qquad \ldots (14\cdot2c)$$

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 (14·2c), we get

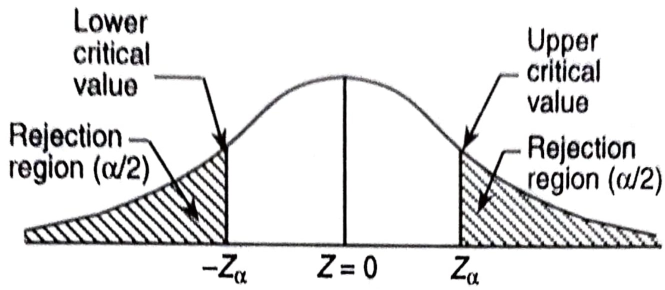
$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \Rightarrow P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha \qquad \text{[By symmetry]}$$

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

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')



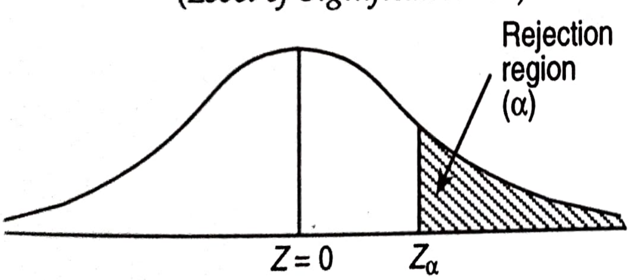
In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $(-z_\alpha)$ is α (See diagrams below), *i.e.*,

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

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

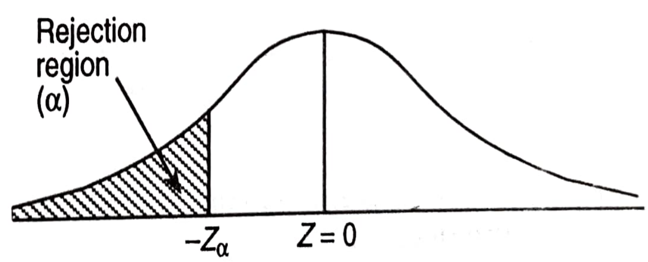
RIGHT-TAILED TEST

(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	$\lvert Z_\alpha \rvert = 2·58$	$\lvert Z_\alpha \rvert = 1·96$	$\lvert Z_\alpha \rvert = 1·645$
Right-tailed test	$Z_\alpha = 2·33$	$Z_\alpha = 1·645$	$Z_\alpha = 1·28$
Left-tailed test	$Z_\alpha = -2·33$	$Z_\alpha = -1·645$	$Z_\alpha = -1·28$

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

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 $|Z| > 1·96$, H_0 is rejected at 5% level of significance.

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

From the normal probability tables, we have :

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

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

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 : $p(x) = {}^nC_x P^x Q^{n-x}$; $x = 0, 1, 2, ..., n$.

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

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

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.*,

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

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).

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

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 :

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We have $Z = \dfrac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$, 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 $\hat{P} = p = \frac{3{,}240}{9{,}000}$ 0·36 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) : $\hat{P} = p = 0{\cdot}13,\ \hat{Q} = q = 0{\cdot}87.$

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

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 :

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

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 $= 0{\cdot}5 \Rightarrow Q = 1 - P = 0{\cdot}5.$

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

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

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

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

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

the covariance terms vanish since the sample observations are independent.

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\Rightarrow \qquad P\left[\,|\,\bar{x}-\mu\,|<1.96\times(\sigma/\sqrt{n})\right]=0.95 \qquad \ldots(**)$$

From (*) and (**), we get $\dfrac{1.96\times\sigma}{\sqrt{n}}=10{,}000 \Rightarrow \dfrac{1.96\times 30{,}000}{\sqrt{n}}=10{,}000$

$$\therefore \qquad n=(1.96\times 3)^2=(5.88)^2=34.56\simeq 35$$

Aliter. Using Remark to Example 14·19,

$$n=\left(\frac{z_\alpha\cdot\sigma}{E}\right)^2=\left(\frac{1.96\times 30{,}000}{10{,}000}\right)^2\simeq 35.$$

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),

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}}\sim N(0,1) \qquad \ldots(14\cdot11)$$

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 $H_0:\mu_1=\mu_2$,

$$Z=\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{\{(1/n_1)+(1/n_2)\}}}\sim N(0,1) \qquad \ldots[14\cdot11a]$$

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 :

$$\hat{\sigma}^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2} \qquad \ldots[14\cdot11b]$$

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

Hence the sample size of each of the two groups should be increased by at 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 :

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

But in case of large samples, the S.E. of the difference of the sample standard deviations is given by :

$$\text{S.E.}\,(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

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

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

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

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

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

(i) Is the difference between the means significant ?

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

CHAPTER FIFTEEN

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

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

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

CHAPTER OUTLINE

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

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

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,

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

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

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

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 :

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

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

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 :

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

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 :

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

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.*,

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

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

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. $\sigma/\sqrt{n}$.

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

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

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

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

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

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

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

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

$$= (1-2t)^{-n/2}, \; |\, 2t \,| < 1 \qquad \ldots(15·4)$$

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)

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

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

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

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

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

$$\mu_r' = 2^r\, \Gamma[(n/2)+r] / \Gamma(n/2) \qquad \ldots(15·4b)$$

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

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

$\therefore \quad \kappa_1$ = Coefficient of t in $K(t) = n$, $\qquad \kappa_2$ = Coefficient of $\frac{t^2}{2\,!}$ in $K(t) = 2n$,

κ_3 = Coefficient of $\frac{t^3}{3\,!}$ in $K(t) = 8n$, and $\qquad \kappa_4$ = Coefficient of $\frac{t^4}{4\,!}$ in $K(t) = 48n$

In general, $\quad \kappa_r$ = Coefficient of $\frac{t^r}{r\,!}$ in $K(t) = n\, 2^{r-1}(r-1)\,!$ $\qquad \ldots(15·4c)$

Hence

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

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 : $\quad M_{X-\mu/\sigma}(t) = e^{-\mu t/\sigma} M_X(t/\sigma)$

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

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

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

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

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

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

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 :

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $n \ge 1$, 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.

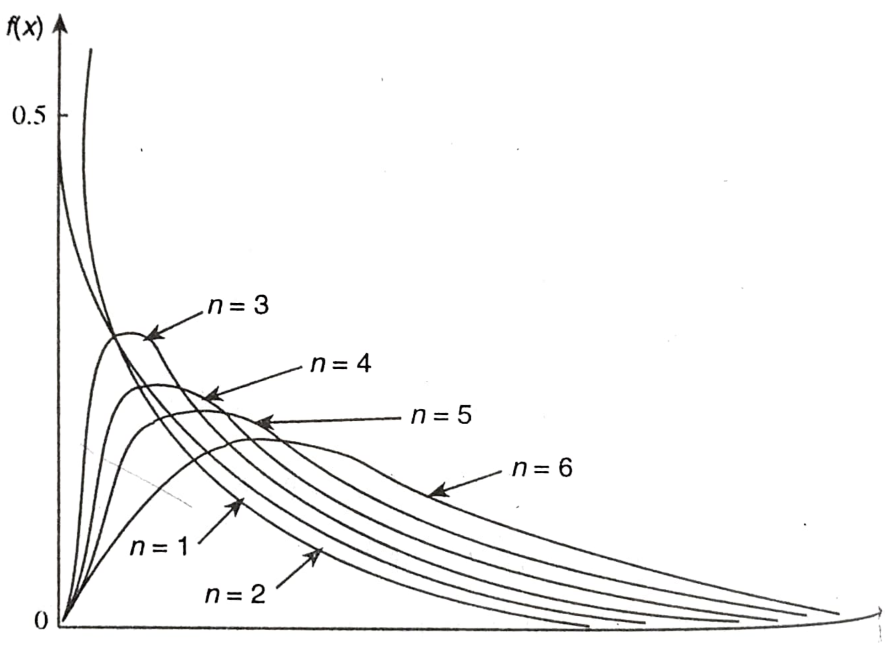


Fig. 15·1: Probability Curve of Chi-square Distribution

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 $\dfrac{X_1}{X_2}$ is a $\beta_2\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate.*

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

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

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

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

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

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

Let us make the transformation :

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof. From *Theorem 15·1*, we have

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

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

Let us transform to u and v defined as follows :

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Another convenient form of this formula is as follows :

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

$$= \sum_{i=1}^{k} (O_i^2/E_i) + \sum_{i=1}^{k} E_i - 2\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} (O_i^2/E_i) - N, \quad \text{...(15...}$$

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

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

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

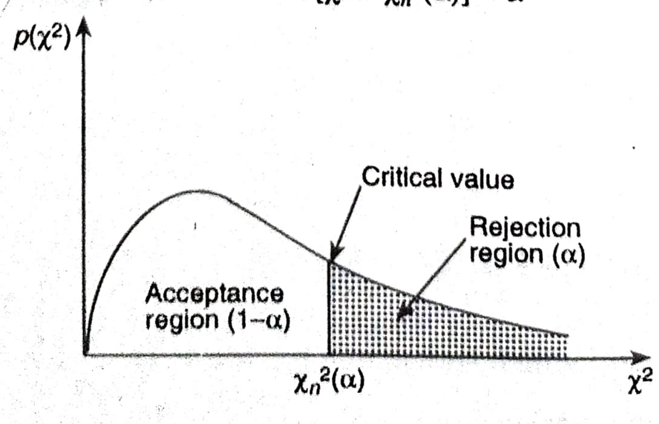
(*ii*) Constraints on the cell frequencies, if any, should be linear, *e.g.*, $\Sigma n_i = \Sigma \lambda_i$ or $\Sigma O_i = ...$

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

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

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

4. *Critical Values.* Let $\chi_n^2(\alpha)$ denote the value of chi-square for n $d.f.$ such that the area... the right of this point is α, *i.e.*, $P[\chi^2 > \chi_n^2(\alpha)] = \alpha$ ...(15...



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mu_2 = 2n\mu_0 = 2n, \qquad \mu_3 = 4(\mu_2 + n\mu_1) = 8n \qquad [\because \mu_1 = 0 \text{ and } \mu_0 = 1]$$

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

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

15·6. APPLICATIONS OF CHI-SQUARE DISTRIBUTION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and apply Normal Test.

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

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

Solution.

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

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

COMPUTATION OF SAMPLE VARIANCE

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

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

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma(X-\overline{X})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182,$$

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 $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{50 \times 225}{100} = 112.5$

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

$\therefore \quad Z = \sqrt{225} - \sqrt{99} = 15 - 9.95 = 5.05$

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

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

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

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

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

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

Example 15·11. *The demand for a particular spare part in a factory was found to vary from day-to-day. In a sample study the following information was obtained :*

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

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

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

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

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

TABLE 15·2 : CALCULATIONS FOR χ^2

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

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

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

The tabulated $\chi^2_{0\cdot05}$ 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 Observed (f_i)	Expected (e_i)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
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

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

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* Observed (f_i)	Expected (e_i)	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
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

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

$d.f. = 4 - 1 = 3$, tabulated $\chi^2_{0\cdot05}$ for 3 $d.f.$ $= 7{\cdot}815$

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 :

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

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

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 :

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

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

TABLE 15·5 : CALCULATIONS FOR χ^2

No. of male births	Frequency Observed (f_i)	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

$$\begin{aligned} f(0) &= 392 \times e^{-0.482} = 392 \times \text{Antilog}(-0.482 \log_{10} e) \\ &= 392 \times \text{Antilog}(-0.482 \times \log_{10} 2.7183) \qquad (\because e = 2.7183) \\ &= 392 \times \text{Antilog}(-0.482 \times 0.4343) = 392 \times \text{Antilog}(-0.2093) \\ &= 392 \times \text{Antilog}(\bar{1}.7907) = 392 \times 0.6176 = 242.1 \end{aligned}$$

$$f(1) = m \times f(0) = 0.482 \times 242.1 = 116.69, \quad f(2) = \frac{m}{2} \times f(1) = 0.241 \times 116.69 = 28.12$$

$$f(3) = \frac{m}{3} \times f(2) = \frac{0.482}{3} \times 28.12 = 4.518, \quad f(4) = \frac{m}{4} \times f(3) = \frac{0.482}{4} \times 4.51 = 0.544$$

$$f(5) = \frac{m}{5} \times f(4) = \frac{0.482}{5} \times 0.544 = 0.052, \quad f(6) = \frac{m}{6} \times f(5) = \frac{0.482}{6} \times 0.052 = 0.004$$

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)	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 } 15	0·5 } 5·1	98·01	19·217
5	2	0·1		
6	1	0		
Total	392	392		40·937

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

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

(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 frequ-encies; 3 $d.f.$ are lost because of pooling the last four expected cell frequencies which are less than five.)

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

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

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

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

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

B \ A	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)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_j	(A_1B_j)	(A_2B_j)	...	(A_iB_j)	...	(A_rB_j)	(B_j)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_s	(A_1B_s)	(A_2B_s)	...	(A_iB_s)	...	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	...	(A_i)	...	(A_r)	N

TABLE 15.

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

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

AND υ IS DEGREES OF FREEDOM (*d.f.*)

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

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

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

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

CHAPTER SIXTEEN

Exact Sampling Distributions-II (t, F and z Distributions)

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

1. Define Student's *t*, Fisher's *t*, F and z statistics and derive their probability distributions.
2. Obtain the various constants of t and F distributions and discuss their important properties.
3. Understand and appreciate various applications of t and F distributions in Statistics.
4. Understand and derive the relationship between t, F and chi-square distributions.
5. Derive the sampling distribution of correlation coefficient in a random sample from an uncorrelated bivariate normal population.
6. Define Fisher's Z-transformation for correlation coefficient and discuss its various applications in Statistics.

CHAPTER OUTLINE

16·1. INTRODUCTION

The entire large sample theory was based on the application of *"Normal Test"* (*cf.* § 14·9). However, if the sample size n is small, the distribution of the various statistics, *e.g.*, $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ or $Z = (X - nP)/\sqrt{nPQ}$ etc., are far from normality and as such *'normal test'* cannot be applied if n is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R.A. Fisher (1926), are used. In the following sections we shall discuss: (*i*) *t-test*, (*ii*) *F-test*, and (*iii*) *Fisher's z-transformation.*

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that *"the population(s) from which sample(s) is (are) drawn is (are) normal, i.e., the parent population(s) is (are) normally distributed."*

16·2. STUDENT'S 't' DISTRIBUTION

Let x_i $(i = 1, 2, \ldots, n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic :

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \ldots(16\cdot1)$$

where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $\ldots(16\cdot1a)$

is an unbiased estimate of the population variance σ^2, and it follows Student's t distribution with $\upsilon = (n - 1)$ $d.f.$ with probability density function :

$$f(t) = \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}\,; -\infty < t < \infty \qquad \ldots(16\cdot2)$$

Remarks 1. A statistic t following Student's t-distribution with n $d.f.$ will be abbreviated as $t \sim t_n$.

2. If we take $\nu = 1$ in (16·2), we get :

$$f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1+t^2)} = \frac{1}{\pi} \cdot \frac{1}{(1+t^2)}\,; -\infty < t < \infty \qquad [\because \Gamma(1/2) = \sqrt{\pi}]$$

which is the $p.d.f.$ of standard Cauchy distribution. Hence, when $\nu = 1$, Student's t distribution reduces to Cauchy distribution.

16·2·1. Derivation of Student's t-distribution. The expression (16·1) can be re-written as :

$$t^2 = \frac{n(\bar{x}-\mu)^2}{S^2} = \frac{n(\bar{x}-\mu)^2}{ns^2/(n-1)} \quad \Rightarrow \quad \frac{t^2}{(n-1)} = \frac{(\bar{x}-\mu)^2}{\sigma^2/n} \cdot \frac{1}{ns^2/\sigma^2} = \frac{(\bar{x}-\mu)^2/(\sigma^2/n)}{ns^2/\sigma^2}$$

Since x_i $(i = 1, 2, \ldots, n)$ is a random sample from the normal population with mean μ and variance σ^2, $\quad \bar{x} \sim N(\mu, \sigma^2/n) \quad \Rightarrow \quad \dfrac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$

Hence $\dfrac{(\bar{x}-\mu)^2}{\sigma^2/n}$, being the square of a standard normal variate is a chi-square variate with 1 $d.f.$

Also $\dfrac{ns^2}{\sigma^2}$ is a χ^2-variate with $(n-1)$ $d.f.$ (*c.f. Theorem 15·5*).

Further since $\bar{x}$ and s^2 are independently distributed (*c.f. Theorem* 15·5), $\dfrac{t^2}{n-1}$ being the ratio of two independent χ^2-variates with 1 and $(n-1)$ $d.f.$ respectively, is a $\beta_2\left(\frac{1}{2}, \frac{n-1}{2}\right)$ variate and its distribution is given by :

$$dF(t) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2/\nu)^{\frac{1}{2}-1}}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}\, d(t^2/\nu),\ 0 \le t^2 < \infty \qquad [\text{where } \nu = (n-1)]$$

$$= \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}\, dt\,; -\infty < t < \infty$$

the factor 2 disappearing since the integral from $-\infty$ to ∞ must be unity. This is the required probability density function as given in (16·2) of Student's t-distribution with $\nu = (n-1)$ $d.f.$

Remarks on Student's 't'. 1. ***Importance of Student's t-distribution in Statistics.*** W.S. Gosset, who wrote under pseudonym (pen-name) of Student defined his t in a slightly different way, *viz.*, $t = (\bar{x}-\mu)/s$ and investigated its sampling distribution, somewhat empirically, in a paper entitled *'The Probable Error of the Mean'*, published in 1908. Prof. R.A. Fisher, later on defined his own 't' and gave a rigorous proof for its sampling distribution in 1926. The salient feature of 't' is that both the statistic and its sampling distribution are functionally independent of σ, the population standard deviation.

The discovery of 't' is regarded as a landmark in the history of statistical inference. Before Student gave his 't', it was customary to replace σ^2 in $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, by its unbiased estimate S^2 to give $t = \dfrac{\bar{x}-\mu}{S/\sqrt{n}}$ and then normal test was applied even for small samples. It has been found that although the distribution of t is asymptotically normal for large n (c.f. § 16·2·5), it is far from normality for small samples. The Student's t ushered in an era of exact sample distributions (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

2. *Confidence or Fiducial Limits for* μ. If $t_{0\cdot05}$ is the tabulated value of t for $\nu = (n-1)$ *d.f.* at 5% level of significance, *i.e.*, $P(|t| > t_{0\cdot05}) = 0\cdot05 \Rightarrow P(|t| \le t_{0\cdot05}) = 0\cdot95$,

the 95% confidence limits for μ are given by :

$$|t| \le t_{0\cdot05}, \quad i.e., \quad \left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| \le t_{0\cdot05} \quad \Rightarrow \quad \bar{x} - t_{0\cdot05} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_{0\cdot05} \frac{S}{\sqrt{n}}$$

Thus, 95% confidence limits for μ are : $\bar{x} \pm t_{0\cdot05} \cdot (S/\sqrt{n})$ ...(16·2a)

Similarly, 99% confidence limits for μ are : $\bar{x} \pm t_{0\cdot01}\, (S/\sqrt{n})$ ...(16·2b)

where $t_{0\cdot01}$ is the tabulated value of t for $\nu = (n-1)$ *d.f.* at 1% level of significance.

16·2·2. Fisher's 't' (*Definition*). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a $N(0, 1)$ and χ^2 is an independent chi-square variate with n *d.f.*, then Fisher's t is given by :

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \qquad ...(16·3)$$

and it follows Student's 't' distribution with n degrees of freedom.

16·2·3. Distribution of Fisher's 't'. Since ξ and χ^2 are independent, their joint probability differential is given by :

$$dF(\xi, \chi^2) = \frac{1}{\sqrt{2\pi}} \exp(-\xi^2/2) \, \frac{\exp(-\chi^2/2)\,(\chi^2)^{\frac{n}{2}-1}}{2^{n/2}\,\Gamma(n/2)}\, d\xi\, d\chi^2$$

Let us transform to new variates t and u by the substitution :

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \quad \text{and} \quad u = \chi^2 \quad \Rightarrow \quad \xi = t\sqrt{u/n} \text{ and } \chi^2 = u$$

Jacobian of transformation J is given by :

$$J = \frac{\partial(\xi, \chi^2)}{\partial(t, u)} = \begin{vmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{n}}$$

The joint *p.d.f* $g(t, u)$ of t and u becomes :

$$g(t, u) = \frac{1}{\sqrt{2\pi}\; 2^{n/2}\,\Gamma(n/2)\sqrt{n}} \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{\frac{n}{2}-\frac{1}{2}}\, du$$

Since $\psi^2 \ge 0$ and $-\infty < \xi < \infty$, $u \ge 0$ and $-\infty < t < \infty$.

Integrating *w.r. to* 'u' over the range 0 to ∞, the marginal *p.d.f.* $g_1(.)$ of t becomes :

$$g_1(t) = \frac{1}{\sqrt{2\pi}\; 2^{n/2}\,\Gamma(n/2)\sqrt{n}} \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{(n-1)/2}\, du\right]$$

$$= \frac{1}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)\sqrt{n}} \cdot \frac{\Gamma[(n+1)/2]}{\left[\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right]^{(n+1)/2}}$$

$$= \frac{\Gamma[(n+1)/2]}{\sqrt{n}\,\Gamma(n/2)\Gamma(1/2)} \cdot \frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}, \; -\infty < t < \infty$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}, \quad -\infty < t < \infty$$

which is the probability density function of Student's t-distribution with n *d.f.*

Remarks 1. In Fisher's 't' the *d.f.* is the same as the *d.f.* of chi-square variate.

2. Student's 't' may be regarded as a particular case of Fisher's 't' as explained below.

Since $\bar{x} \sim N(\mu, \sigma^2/n)$, $\quad \xi = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ ...(*) and $\chi^2 = \dfrac{ns^2}{\sigma^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2/\sigma^2$ ...(**)

is independently distributed as chi-square variate with $(n-1)$ *d.f.* Hence Fisher's t is given by :

$$t = \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{\Sigma(x_i-\bar{x})^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{S} = \frac{\bar{x}-\mu}{S/\sqrt{n}} \quad ...(***)$$

and it follows Student's t-distribution with $(n-1)$ *d.f.* (*c.f.* Remark 1 above.)

Now, (***) is same as Student's 't' defined in (16·1). Hence Student's 't' is a particular case of Fisher's 't'.

16·2·4. Constants of t-distribution. Since $f(t)$ is symmetrical about the line $t = 0$, all the moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2r+1} \text{ (about origin)} = 0 \; ; r = 0, 1, 2, \ldots$$

In particular, μ_1' (about origin) $= 0 =$ Mean

Hence central moments coincide with moments about origin.

$$\therefore \quad \mu_{2r+1} = 0, \; (r = 0, 1, 2, \ldots) \quad ...(16·4)$$

The moments of even order are given by :

$$\mu_{2r} = \mu'_{2r} \text{ (about origin)} = \int_{-\infty}^{\infty} t^{2r} f(t)\, dt = 2\int_0^{\infty} t^{2r} f(t)\, dt$$

$$= 2 . \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}} \int_0^{\infty} \frac{t^{2r}}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} dt$$

This integral is absolutely convergent if $2r < n$.

Put $\quad 1 + \dfrac{t^2}{n} = \dfrac{1}{y} \quad \Rightarrow \quad t^2 = \dfrac{n(1-y)}{y} \quad \Rightarrow \quad 2t dt = -\dfrac{n}{y^2} dy$

When $t = 0, y = 1$ and when $t = \infty, y = 0$. Therefore,

$$\mu_{2r} = \frac{2}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{(1/y)^{(n+1)/2}} \cdot \frac{-n}{2ty^2} dy$$

$$= \frac{n}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 (t^2)^{(2r-1)/2}\, y^{[(n+1)/2]-2}\, dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} y^{\{(n+1)/2\}-2}\, dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2}-r-1} (1-y)^{r-\frac{1}{2}}\, dy = \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot B\left(\frac{n}{2}-r, r+\frac{1}{2}\right), n > 2r$$

$$= n^r \frac{\Gamma[(n/2)-r]\,\Gamma(r+\frac{1}{2})}{\Gamma(1/2)\,\Gamma(n/2)} \quad \ldots (16\text{·}4a)$$

$$= n^r \frac{\left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\ldots\frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-r\right)}{\Gamma(1/2)\,[(n/2)-1][(n/2)-2]\ldots[(n/2)-r]\Gamma[(n/2)-r]}$$

$$= n^r \frac{(2r-1)(2r-3)\ldots3\cdot1}{(n-2)(n-4)\ldots(n-2r)}, \frac{n}{2} > r \quad \ldots (16\text{·}4b)$$

In particular

$$\mu_2 = n \cdot \frac{1}{(n-2)} = \frac{n}{n-2}, (n > 2) \quad \ldots (16\text{·}4c)$$

and

$$\mu_4 = n^2 \frac{3\cdot1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)}, (n > 4) \quad \ldots (16\text{·}4d)$$

Hence

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3\left(\frac{n-2}{n-4}\right); (n > 4).$$

Remarks 1. As $n \to \infty$, $\beta_1 = 0$ and $\beta_2 = \lim\limits_{n\to\infty} 3\left(\frac{n-2}{n-4}\right) = 3 \lim\limits_{n\to\infty}\left[\frac{1-(2/n)}{1-(4/n)}\right] = 3$ ...(16·4e)

2. Changing r to $(r-1)$ in [14·4(b)], dividing and simplifying, we shall get the recurrence relation for the moments as $\frac{\mu_{2r}}{\mu_{2r-2}} = \frac{n(2r-1)}{(n-2r)}, \frac{n}{2} > r$ ...(16·4f)

3. ***Moment Generating Function of t-distribution.*** From [16·4(b)] we observe that if $t \sim t_n$, then all the moments of order $2r < n$ exist but the moments of order $2r \geq n$ do not exist. Hence the *m.g.f.* of t-distribution does not exist.

Example 16·1. *Express the constants y_0, a and m of the distribution :*

$$dF(x) = y_0\left(1 - \frac{x^2}{a^2}\right)^m dx, \; -a \leq x \leq a \quad \ldots (*)$$

in terms of its μ_2 and β_2.

Show that if x is related to a variable t by the equation :

$$x = \frac{at}{\{2(m+1) + t^2\}^{1/2}}, \quad \ldots (**)$$

then t has Student's distribution with $2(m+1)$ degrees of freedom. Use the transformation to calculate the probability that $t \geq 2$ when the degrees of freedom are 2 and also when 4.

Solution. First of all, we shall determine the constant y_0 from the consideration that total probability is unity.

$$\therefore \quad y_0 \int_{-a}^{a} \left(1 - \frac{x^2}{a^2}\right)^m dx = 1 \quad \Rightarrow \quad 2y_0 \int_0^a \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

($\because$ Integrand is an even function of x)

$$\Rightarrow \quad 2y_0 \int_0^{\pi/2} \cos^{2m}\theta \,.\, a\cos\theta \, d\theta = 1, \qquad (x = a\sin\theta)$$

$$\Rightarrow \quad 2ay_0 \int_0^{\pi/2} \cos^{2m+1}\theta \, d\theta = 1$$

But we have the Beta integral, $\quad 2\int_0^{\pi/2} \sin^p\theta \cos^q\theta \, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \quad \ldots(1)$

$$\therefore \quad ay_0 . 2\int_0^{\pi/2} \cos^{2m+1}\theta \sin^0\theta \, d\theta = 1 \quad \Rightarrow \quad ay_0\, B(m+1, \tfrac{1}{2}) = 1 \qquad \text{[Using (1)]}$$

$$\Rightarrow \quad y_0 = \frac{1}{a\,B\left(m+1 \cdot \frac{1}{2}\right)} \qquad \ldots(2)$$

Since the given probability function is symmetrical about the line $x = 0$, we have as in § 16·2·4, $\quad \mu_{2r+1} = \mu'_{2r+1} = 0\,;\, r = 0, 1, 2, \ldots \quad$ [$\because$ Mean = Origin]

The moments of even order are given by :

$$\mu_{2r} = \mu_{2r}' \text{ (about origin)} = \int_{-a}^{a} x^{2r} f(x)\, dx = y_0 \int_{-a}^{a} x^{2r}\left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^a x^{2r}\left(1 - \frac{x^2}{a^2}\right)^m dx = 2y_0 \int_0^{\pi/2} (a\sin\theta)^{2r} \cos^{2m}\theta .\; a\cos\theta\, d\theta, \qquad (x = a\sin\theta)$$

$$= y_0\, a^{2r+1} .\, 2\int_0^{\pi/2} \sin^{2r}\theta \,.\, \cos^{2m+1}\theta\, d\theta = y_0\, a^{2r+1}\, B(r + \tfrac{1}{2}, m+1) \qquad \text{[Using (1)]}$$

$$= a^{2r} \frac{B(r+\frac{1}{2}, m+1)}{B(m+1, \frac{1}{2})} = a^{2r} . \frac{\Gamma\left(r+\frac{1}{2}\right)\Gamma\left(m+\frac{3}{2}\right)}{\Gamma\left(m+r+\frac{3}{2}\right)\Gamma\left(\frac{1}{2}\right)} \qquad \ldots(***)$$

In particular, $\quad \mu_2 = a^2 . \dfrac{\Gamma\{m+(3/2)\} . \frac{1}{2}\Gamma(1/2)}{\{m+(3/2)\}\,\Gamma\{m+(3/2)\}\,\Gamma(1/2)} = \dfrac{a^2}{2m+3}$

$$\Rightarrow \quad a^2 = (2m+3)\mu_2 \qquad \ldots(3)$$

Also $\quad \mu_4 = a^4 \dfrac{\Gamma(5/2)}{\Gamma\{m+(7/2)\}} \times \dfrac{\Gamma\{m+(3/2)\}}{\Gamma(1/2)} = \dfrac{3a^4}{(2m+5)(2m+3)} \qquad$ (On simplification)

$\therefore \quad \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3(2m+3)}{(2m+5)} \quad \Rightarrow \quad m = \dfrac{9-5\beta_2}{2(\beta_2-3)} \qquad$ (On simplification) ... (4)

Equations (2), (3) and (4) express the constants y_0, a and m in terms of μ_2 and β_2.

$$x = \frac{at}{[2(m+1)+t^2]^{1/2}} \quad \Rightarrow \quad \frac{x^2}{a^2} = \frac{t^2}{2(m+1)+t^2}$$

i.e., $\quad 1 - \dfrac{x^2}{a^2} = \dfrac{2(m+1)}{2(m+1)+t^2} = \left(1 + \dfrac{t^2}{n}\right)^{-1}, \qquad (n = 2m+2)$

Also $\quad dx = a\left[\dfrac{dt}{(n+t^2)^{1/2}} - t . \dfrac{1}{2}\dfrac{2t\,dt}{(n+t^2)^{3/2}}\right] = a\dfrac{1}{(n+t^2)^{1/2}}\left(1 - \dfrac{t^2}{n+t^2}\right)dt$

$$= \frac{an}{(n+t^2)^{3/2}}\,dt = \frac{a}{\sqrt{n}} \cdot \frac{1}{[1+(t^2/n)]^{3/2}}\,dt$$

$$= 1 - \frac{1}{2B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^x z^{(n/2)-1} (1-z)^{(1/2)-1} \, dz \qquad \left[\text{where } x = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

$$= 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right), \qquad \left[x = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

where $I_x(p, q)$ is defined in (*).

Example 16·4. *Show that for t-distribution with n d.f., mean deviation about mean is given by :*

$$\sqrt{n}\ \Gamma\left[(n-1)/2\right] / \sqrt{\pi}\, \Gamma(n/2)$$

Solution. $E(t) = 0$.

$$\text{M.D. (about mean)} = \int_{-\infty}^{\infty} |t| f(t)\, dt = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|t|\, dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{2}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{t\,dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}} = \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{dy}{(1+y)^{(n+1)/2}}, \qquad \left(\frac{t^2}{n} = y\right)$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^{\infty} \frac{y^{1-1}}{(1+y)^{\frac{n-1}{2}+1}}\, dy = \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot B\left(\frac{n-1}{2}, 1\right) = \frac{\sqrt{n}\, \Gamma[(n-1)/2]}{\sqrt{\pi}\, \Gamma(n/2)}$$

16·2·5. Limiting Form of t-distribution. *As $n \to \infty$, the p.d.f. of t-distribution with n d.f. viz.,*

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \to \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right), -\infty < t < \infty$$

Proof.
$$\lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} = \lim_{n \to \infty} \frac{1}{\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\Gamma(1/2)\, \Gamma(n/2)} = \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}$$

$$\left[\because \Gamma(1/2) = \sqrt{\pi} \text{ and } \lim_{n \to \infty} \frac{\Gamma(n+k)}{\Gamma(n)} = n^k, \text{ (c.f. Remark to § 16·8)}\right]$$

$$\therefore \quad \lim_{n \to \infty} f(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \lim_{n \to \infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-\frac{1}{2}} \times \lim_{n \to \infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}}$$

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

Hence for large d.f. t-distribution tends to standard normal distribution.

16·2·6. Graph of t-distribution. The *p.d.f.* of t-distribution with n *d.f.* is :

$$f(t) = C.\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, -\infty < t < \infty$$

Since $f(-t) = f(t)$, the probability curve is symmetrical about the line $t = 0$. As t increases, $f(t)$ decreases rapidly and tends to zero as $t \to \infty$, so that t-axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2}, n > 2\ ; \quad \beta_2 = \frac{3(n-2)}{(n-4)}, n > 4$$

Hence for $n > 2$, $\mu_2 > 1$ *i.e.*, the variance of t-distribution is greater than that of standard normal distribution and for $n > 4$, $\beta_2 > 3$ and thus t-distribution is more flat on the top than the normal curve. In fact, for small n, we have

$$P(|t| \geq t_0) \geq P(|Z| \geq t_0), \quad Z \sim N(0, 1)$$

i.e., the tails of the t-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we have also seen [§ 16·2·5], that for large n (*d.f.*), t-distribution tends to standard normal distribution.

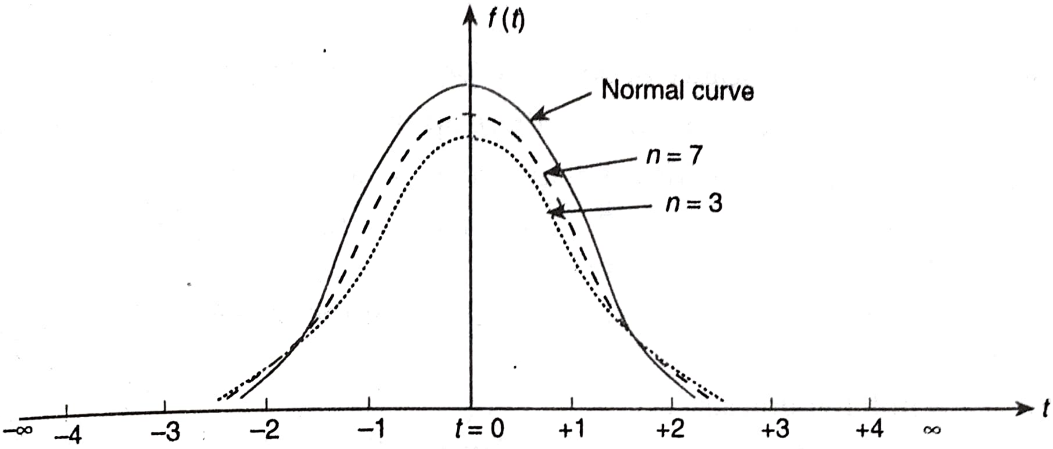


Fig. 16·1 :Graph of t-distribution

16·2·7. Critical Values of t. The critical (or significant) values of t at level of significance α and *d.f.* v for *two-tailed test* are given by the equation :

$$P[\,|t| > t_v(\alpha)] = \alpha \qquad \text{...(16·5)}$$

$$\Rightarrow \quad P[\,|t| \leq t_v(\alpha)] = 1 - \alpha \qquad \text{...(16·5}a\text{)}$$

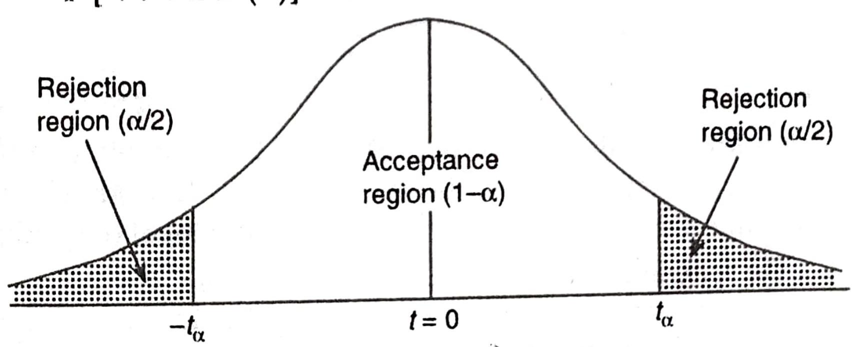


Fig. 16·2 : Critical values of t-distribution

The values $t_v(\alpha)$ have been tabulated in Fisher and Yates' Tables, for different values of α and v and are given in *Table I* at the end of the chapter.

Since t-distribution is symmetric about $t = 0$, we get from (16·5)

$$P(t > t_v(\alpha)] + P[t < -t_v(\alpha)] = \alpha \;\Rightarrow\; 2P[t > t_v(\alpha)] = \alpha$$

$$\Rightarrow \quad P[t > t_v(\alpha)] = \alpha/2 \qquad \therefore\; P[t > t_v(2\alpha)] = \alpha \qquad \text{...(16·5}b\text{)}$$

$t_v(2\alpha)$ (from the Tables at the end of the chapter) gives the significant value of t for a single-tail test [Right-tail or Left-tail-since the distribution is symmetrical], at level of significance α and v *d.f.*

Hence the significant values of t at level of significance 'α' for a single-tailed test can be obtained from those of two-tailed test by looking the values at level of significance 2α.

For example,

$t_8(0·05)$ for single-tail test $= t_8(0·10)$ for two-tail test $= 1·86$

$t_{15}(0·01)$ for single-tail test $= t_{15}(0·02)$ for two-tail test $= 2·60$.

16·3. APPLICATIONS OF t-DISTRIBUTION

The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(i) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean :

(ii) To test the significance of the difference between two sample means.

(iii) To test the significance of an observed sample correlation coefficient and sample regression coefficient.

(iv) To test the significance of observed partial correlation coefficient.

In the following sections we will discuss these applications in detail, one by one.

16·3·1. t-Test for Single Mean. Suppose we want to test :

(i) if a random sample x_i ($i = 1, 2, \ldots, n$) of size n has been drawn from a normal population with a specified mean, say μ_0, or

(ii) if the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis, H_0 :

(i) The sample has been drawn from the population with mean μ_0 or

(ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ_0,

the statistic
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}, \qquad \ldots(16·6)$$

where
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \ldots[16·6(a)]$$

follows Student's t-distribution with $(n-1)$ $d.f.$

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t| >$ tabulated t, null hypothesis is rejected and if calculated $|t| <$ tabulated t, H_0 may be accepted at the level of significance adopted.

Remarks 1. ***On computation of S^2 for numerical problems.*** If $\bar{x}$ comes out in integers, the formula (16·6a) can be conveniently used for computing S^2. However, if $\bar{x}$ comes in fractions then the formula (16·6a) for computing S^2 is very cumbersome and is not recommended. In that case, step deviation method, given below, is quite useful.

If we take $d_i = x_i - A$, where A is any arbitrary number, then

$$S^2 = \frac{1}{n-1}\left[\Sigma(x_i - \bar{x})^2\right] = \frac{1}{n-1}\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right] \qquad \ldots(16·6b)$$

$$= \frac{1}{n-1}\left[\Sigma d_i^2 - \frac{(\Sigma d_i)^2}{n}\right], \text{ since variance is independent of change of origin.} \qquad \ldots(16·6c)$$

Also, in this case
$$\bar{x} = A + \frac{\Sigma d_i}{n}. \qquad \ldots(16·6d)$$

2. We know, the sample variance : $s^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \quad \Rightarrow \quad ns^2 = (n-1) S^2$

$$\therefore \quad \frac{S^2}{n} = \frac{s^2}{n-1} \qquad \ldots(16·6e)$$

Hence for numerical problems test statistic (16·6) on using [16·6(e)] becomes

$$= \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1} \qquad \ldots(16·6f)$$

3. Assumption for Student's t-test. The following assumptions are made in the Student's t-test :

(i) The parent population from which the sample is drawn is normal.

(ii) The sample observations are independent, *i.e.*, the sample is random.

(iii) The population standard deviation σ is unknown.

Example 16·5. *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.*

Solution. Here we are given :

$$\mu = 0{\cdot}700 \text{ inche}, \quad \bar{x} = 0{\cdot}742 \text{ inche}, \quad s = 0{\cdot}040 \text{ inche} \quad \text{and} \quad n = 10$$

Null Hypothesis, $H_0 : \mu = 0{\cdot}700$, *i.e., the product is conforming to specifications.*

Alternative Hypothesis, $H_1 : \mu \neq 0{\cdot}700$

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \mu}{\sqrt{S^2/n}} = \dfrac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{(n-1)}$

$$\therefore \quad t = \frac{\sqrt{9}\,(0{\cdot}742 - 0{\cdot}700)}{0{\cdot}040} = 3{\cdot}15$$

How to proceed further. Here the test statistic 't' follows Student's t-distribution with $10 - 1 = 9$ *d.f.* We will now compare this calculated value with the tabulated value of t for 9 *d.f.* and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(i) If calculated 't', *viz.*, $3{\cdot}15 > t_0$, we say that the value of t is significant. This implies that $\bar{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(ii) If calculated $t < t_0$, we say that the value of t is not significant, *i.e.*, there is no significant difference between $\bar{x}$ and μ. In other words, the deviation $(\bar{x} - \mu)$ is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, *i.e.*, we may take the product conforming to specifications.

Example 16·6. *The mean weekly sales of soap bars in departmental stores was 146·3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153·7 and showed a standard deviation of 17·2. Was the advertising campaign successful ?*

Solution. We are given : $n = 22$, $\bar{x} = 153{\cdot}7$, $s = 17{\cdot}2$.

Null Hypothesis. The advertising campaign is not successful, *i.e.*, $H_0 : \mu = 146{\cdot}3$

Alternative Hypothesis, $H_1 : \mu > 146{\cdot}3$ (Right-tail).

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$

$$\therefore \quad t = \frac{153{\cdot}7 - 146{\cdot}3}{\sqrt{(17{\cdot}2)^2/21}} = \frac{7{\cdot}4 \times \sqrt{21}}{17{\cdot}2} = 9{\cdot}03$$

Conclusion. Tabulated value of t for 21 *d.f.* at 5% level of significance for *single-tailed test* is 1·72. Since calculated value is much greater than the tabulated value, it is

3. *Assumption for Student's t-test.* The following assumptions are made in the Student's *t*-test :

(*i*) The parent population from which the sample is drawn is normal.
(*ii*) The sample observations are independent, *i.e.*, the sample is random.
(*iii*) The population standard deviation σ is unknown.

Example 16·5. *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.*

Solution. Here we are given :

$$\mu = 0{\cdot}700 \text{ inche}, \quad \bar{x} = 0{\cdot}742 \text{ inche}, \quad s = 0{\cdot}040 \text{ inche} \quad \text{and} \quad n = 10$$

Null Hypothesis, $H_0 : \mu = 0{\cdot}700$, *i.e., the product is conforming to specifications.*

Alternative Hypothesis, $H_1 : \mu \neq 0{\cdot}700$

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \mu}{\sqrt{S^2/n}} = \dfrac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{(n-1)}$

$$\therefore \qquad t = \frac{\sqrt{9}\,(0{\cdot}742 - 0{\cdot}700)}{0{\cdot}040} = 3{\cdot}15$$

How to proceed further. Here the test statistic '*t*' follows Student's *t*-distribution with 10 – 1 = 9 *d.f.* We will now compare this calculated value with the tabulated value of *t* for 9 *d.f.* and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(*i*) If calculated '*t*', *viz.*, $3{\cdot}15 > t_0$, we say that the value of *t* is significant. This implies that $\bar{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(*ii*) If calculated $t < t_0$, we say that the value of *t* is not significant, *i.e.*, there is no significant difference between $\bar{x}$ and μ. In other words, the deviation $(\bar{x} - \mu)$ is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, *i.e.*, we may take the product conforming to specifications.

Example 16·6. *The mean weekly sales of soap bars in departmental stores was 146·3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153·7 and showed a standard deviation of 17·2. Was the advertising campaign successful ?*

Solution. We are given : $n = 22, \bar{x} = 153{\cdot}7, s = 17{\cdot}2$.

Null Hypothesis. The advertising campaign is not successful, *i.e.*, $H_0 : \mu = 146{\cdot}3$

Alternative Hypothesis, $H_1 : \mu > 146{\cdot}3$ (Right-tail).

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$

$$\therefore \qquad t = \frac{153{\cdot}7 - 146{\cdot}3}{\sqrt{(17{\cdot}2)^2/21}} = \frac{7{\cdot}4 \times \sqrt{21}}{17{\cdot}2} = 9{\cdot}03$$

Conclusion. Tabulated value of *t* for 21 *d.f.* at 5% level of significance for *single-tailed test* is 1·72. Since calculated value is much greater than the tabulated value, it is highly significant. Hence we reject the null hypothesis and conclude that the advertising campaign was definitely successful in promoting sales.

Example 16·7. *A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.*

Solution. Null hypothesis, H_0 : *The data are consistent with the assumption of a mean I.Q. of 100 in the population, i.e.,* $\mu = 100$.

Alternative hypothesis, $H_1 : \mu \neq 100$.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{(\bar{x} - \mu)}{\sqrt{S^2/n}} \sim t_{(n-1)}$,

where $\bar{x}$ and S^2 are to be computed from the sample values of I.Q.'s.

TABLE 16·1 : CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
70	− 27·2	739·84
120	22·8	519·84
110	12·8	163·84
101	3·8	14·44
88	− 9·2	84·64
83	− 14·2	201·64
95	− 2·2	4·84
98	0·8	0·64
107	9·8	96·04
100	2·8	7·84
Total 972		1833·60

Here $n = 10$, $\bar{x} = \dfrac{972}{10} = 97{\cdot}2$ and $S^2 = \dfrac{1833{\cdot}60}{9} = 203{\cdot}73$

$$\therefore \qquad |t| = \frac{|97{\cdot}2 - 100|}{\sqrt{203{\cdot}73/10}} = \frac{2{\cdot}8}{\sqrt{20{\cdot}37}} = \frac{2{\cdot}8}{4{\cdot}514} = 0{\cdot}62$$

Tabulated $t_{0{\cdot}05}$ for $(10 - 1)$, *i.e.,* 9 *d.f.* for two-tailed test is 2·262.

Conclusion. Since calculated t is less than tabulated $t_{0{\cdot}05}$ for 9 *d.f.*, H_0 may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by :

$$\bar{x} \pm t_{0{\cdot}05}\, S/\sqrt{n} = 97{\cdot}2 \pm 2{\cdot}262 \times 4{\cdot}514 = 97{\cdot}2 \pm 10{\cdot}21 = 107{\cdot}41 \text{ and } 86{\cdot}99$$

Hence the required 95% confidence interval is [86·99, 107·41].

Remark. Aliter for computing $\bar{x}$ ***and*** S^2. Here we see that $\bar{x}$ comes in fractions and as such the computation of $(x - \bar{x})^2$ is quite laborious and time consuming. In this case we use the method of step deviations to compute $\bar{x}$ and S^2, as given below.

x	$d = x - 90$	d^2
70	– 20	400
120	30	900
110	20	400
101	11	121
88	– 2	4
83	– 7	49
95	5	25
98	8	64
107	17	289
100	10	100
Total	$\Sigma d = 72$	$\Sigma d^2 = 2{,}352$

Here $d = x - A$, where $A = 90$. Therefore

$$\bar{x} = A + \frac{1}{n}\Sigma d = 90 + \frac{72}{10} = 97{\cdot}2 \text{ and } S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{9}\left[2352 - \frac{(72)^2}{10}\right] = 203{\cdot}73.$$

Example 16·8. *The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? Test at 5% significance level assuming that for 9 degrees of freedom $P\ (t > 1{\cdot}83) = 0{\cdot}05$.*

Solution. *Null Hypothesis,* $H_0 : \mu = 64$ inches

Alternative Hypothesis, $H_1 : \mu > 64$ inches

TABLE 16·2 : CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	70	67	62	68	61	68	70	64	64	66	Total 660
$x - \bar{x}$	4	1	– 4	2	–5	2	4	–2	–2	0	0
$(x - \bar{x})^2$	16	1	16	4	25	4	16	4	4	0	90

$$\bar{x} = \frac{\Sigma x}{n} = \frac{660}{10} = 66; \quad S^2 = \frac{1}{n-1}\Sigma (x - \bar{x})^2 = \frac{90}{9} = 10$$

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{66 - 64}{\sqrt{10/10}} = 2,$$

which follows Student's t-distribution with $10 - 1 = 9$ $d.f.$

Tabulated value of t for 9 $d.f.$ at 5% level of significance for single (right) tail-test is 1·833. (This is the value $t_{0{\cdot}10}$ for 9 $d.f.$ in the two-tailed tables given at the end of the chapter.)

Conclusion. Since calculated value of t is greater than the tabulated value, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

Example 16·9. *A random sample of 16 values from a normal population showed a mean of 41·5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43·5 inches for the population is not reasonable. Obtain 95 per cent and 99 per cent fiducial limits for the same.*

You may use the following information from statistical tables :

$$v = 15, \begin{cases} P = 0{\cdot}05, \ t = 2{\cdot}131 \\ P = 0{\cdot}01, \ t = 2{\cdot}947 \end{cases}$$

Solution. We are given $n = 16, \bar{x} = 41{\cdot}5$ inches and $\Sigma(x - \bar{x})^2 = 135$ sq. inches.

$$\therefore \quad S^2 = \frac{1}{n-1}\Sigma(x - \bar{x})^2 = \frac{135}{15} = 9 \quad \Rightarrow \quad S = 3$$

Null Hypothesis, $H_0 : \mu = 43{\cdot}5$ inches, *i.e.,* the data are consistent with the assumption that the mean height in the population is 43·5 inches.

Alternative Hypothesis, $H_1 : \mu \neq 43{\cdot}5$ inches.

Test Statistic. Under H_0, the test statistic is : $\quad t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$

$$\therefore \quad |\, t \,| = \frac{|\, 41{\cdot}5 - 43{\cdot}5 \,|}{3/4} = \frac{8}{3} = 2{\cdot}667$$

Here number of degrees of freedom is $(16 - 1) = 15$.

We are given : $t_{0{\cdot}05}$ for 15 *d.f.* = 2·131 and $t_{0{\cdot}01}$ for 15 *d.f.* = 2·947.

Conclusion. Since calculated $|\, t \,|$ is greater than 2·131, null hypothesis is rejected at 5% level of significance and we conclude that the assumption of mean of 43·5 inches for the population is not reasonable.

Remark. Since calculated $|\, t \,|$ is less than 2·947, null hypothesis ($\mu = 43{\cdot}5$) may be accepted at 1% level of significance.

95% fiducial limits for μ : (*d.f.* = 15)

$$\bar{x} \pm t_{0{\cdot}05} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}131 \times \frac{3}{4} = 41{\cdot}5 \pm 1{\cdot}598 \quad \Rightarrow \quad 39{\cdot}902 < \mu < 43{\cdot}098$$

99% fiducial limits for μ : (*d.f.* = 15)

$$\bar{x} \pm t_{0{\cdot}01} \times \frac{S}{\sqrt{n}} = 41{\cdot}5 \pm 2{\cdot}947 \times \frac{3}{4} = 43{\cdot}71 \text{ and } 39{\cdot}29 \quad \Rightarrow \quad 39{\cdot}29 < \mu < 43{\cdot}71$$

16·3·2. t-Test for Difference of Means. Suppose we want to test if two independent samples x_i $(i = 1, 2, \ldots, n_1)$ and y_j, $(j = 1, 2, \ldots, n_2)$ of sizes n_1 and n_2 have been drawn from two normal populations with means μ_X and μ_Y respectively.

Under the null hypothesis (H_0) *that the samples have been drawn from the normal populations with means* μ_X *and* μ_Y *and under the assumption that the population variance are equal, i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \qquad \ldots(16{\cdot}7)$$

where $\quad \bar{x} = \dfrac{1}{n_1}\sum\limits_{i=1}^{n_1} x_i, \qquad \bar{y} = \dfrac{1}{n_2}\sum\limits_{j=1}^{n_2} y_j$

and $\quad S^2 = \dfrac{1}{n_1 + n_2 - 2}\left[\sum\limits_i (x_i - \bar{x})^2 + \sum\limits_j (y_j - \bar{y})^2\right] \qquad \ldots(16{\cdot}7a)$

is an unbiased estimate of the common population variance σ^2, follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ *d.f.*

Proof. *Distribution of t defined in (16·7).*

$$\xi = \frac{(\bar{x}-\bar{y}) - E(\bar{x}-\bar{y})}{\sqrt{V(\bar{x}-\bar{y})}} \sim N(0, 1)$$

But $\quad E(\bar{x}-\bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$

$$V(\bar{x}-\bar{y}) = V(\bar{x}) + V(\bar{y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} = \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \qquad \text{(By assumption)}$$

[The covariance term vanishes since samples are independent.]

$$\therefore \quad \xi = \frac{(\bar{x}-\bar{y}) - (\mu_X-\mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(*)$$

Let
$$\chi^2 = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]$$

$$= \left[\sum_i (x_i-\bar{x})^2/\sigma^2\right] + \left[\sum_j (y_j-\bar{y})^2/\sigma^2\right] = \frac{n_1 s_X^2}{\sigma^2} + \frac{n_2 s_Y^2}{\sigma^2} \qquad \ldots(**)$$

Since $n_1 s_X^2/\sigma^2$ and $n_2 s_Y^2/\sigma^2$ are independent χ^2-variates with (n_1-1) and (n_2-1) *d.f.* respectively, by the additive property of chi-square distribution, χ^2 defined in (**) is a χ^2-variate with $(n_1-1)+(n_2-1)$, *i.e.*, n_1+n_2-2 *d.f.* Further, since sample mean and sample variance are independently distributed, ξ and χ^2 are independent random variables. Hence Fisher's t statistic is given by

$$t = \frac{\xi}{\sqrt{\dfrac{\chi^2}{n_1+n_2-2}}}$$

$$= \frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \times \frac{1}{\left[\dfrac{1}{n_1+n_2-2}\left\{\sum_i (x_i-\bar{x})^2 + \sum_j (y_j-\bar{y})^2\right\}/\sigma^2\right]^{1/2}}$$

$$= \frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{S\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \text{ where } S^2 = \frac{1}{n_1+n_2-2}\left[\sum_i (x_i-\bar{x})^2 + \sum_j (y_j-\bar{y})^2\right]$$

and it follows Student's t-distribution with (n_1+n_2-2) *d.f.* (*c.f.* Remark 1, § 16·2·3).

Remarks 1. S^2, defined in (16·7*a*) is an unbiased estimate of the common population variance σ^2, since

$$E(S^2) = \frac{1}{n_1+n_2-2}E\left[\sum_i (x_i-\bar{x})^2 + \sum_j (y_j-\bar{y})^2\right] = \frac{1}{n_1+n_2-2}E\left[(n_1-1)S_X^2 + (n_2-1)S_Y^2\right]$$

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

2. An important deduction which is of much practical utility is discussed below :

Suppose we want to test if : (*a*) two independent samples x_i $(i = 1, 2, \ldots, n_1)$, and y_j $(j = 1, 2, \ldots, n_2)$, have been drawn from the populations with same means, or (*b*) the two sample means $\bar{x}$ and $\bar{y}$ differ significantly or not.

Under the null hypothesis, H_0 that (*a*) *samples have been drawn from the populations with same means, i.e.,* $\mu_X = \mu_Y$, *or* (*b*) *the sample means* $\bar{x}$ *and* $\bar{y}$ *do not differ significantly,* the statistic :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad [\because \mu_X = \mu_Y, \text{ under } H_0] \qquad \ldots(16\cdot8)$$

where symbols are defined in (16·7*a*), follows Student's *t*-distribution with $(n_1 + n_2 - 2)$ *d.f.*

3. *On the assumption of t-test for difference of means.* Here we make the following three fundamental assumptions :

(*i*) Parent populations, from which the samples have been drawn are normally distributed.

(*ii*) The population variances are equal and unknown, *i.e.*, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), where σ^2 is unknown.

(*iii*) The two samples are random and independent of each other.

Thus before applying *t*-test for testing the equality of means it is theoretically desirable to test the equality of population variances by applying *F*-test. (*c.f* § 16·6·1) If the variances do not come out to be equal then *t*-test becomes invalid and in that case Behren's '*d*'-test based on fiducial intervals is used. For practical problems, however, the assumptions (*i*) and (*ii*) are taken for granted.

16·3·3. Paired t-test for Difference of Means. Let us now consider the case when (*i*) the sample sizes are equal, *i.e.*, $n_1 = n_2 = n$ (say), and (*ii*) the two samples are not independent but the sample observations are paired together, *i.e.*, the pair of observations (x_i, y_i), $(i = 1, 2, \ldots, n)$ corresponds to the same (*i*th) sample unit. The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let x_i and y_i $(i = 1, 2, \ldots, n)$ be the readings, in hours of sleep, on the *i*th individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed in § 16·3·2, we apply the paired *t*-test given below.

Here we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, \ldots, n)$.

Under the null hypothesis, H_0 that *increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments,* the statistic :

$$t = \frac{\bar{d}}{S/\sqrt{n}} \qquad \ldots(16\cdot9)$$

where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 \qquad \ldots(16\cdot9a)$$

follows Student's *t*-distribution with $(n - 1)$ *d.f.*

Example 16·10. *Below are given the gain in weights (in kgs.) of pigs fed on two diets A and B.*

Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution. Null hypothesis, $H_0 : \mu_X = \mu_Y$, *i.e., there is no significant difference between the mean increase in weight due to diets A and B.*

Alternative hypothesis, $H_1 : \mu_X \neq \mu_Y$ (two-tailed).

Diet A			Diet B		
x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
25	−3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	−8	64
34	6	36	10	−20	400
24	−4	16	47	17	289
14	−14	196	31	1	1
32	4	16	40	10	100
24	−4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	−12	144
25	−3	9	21	−9	81
			35	5	25
			29	−1	1
			22	−8	64
$\Sigma x = 336$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(x-\bar{x})^2 = 380$	$\Sigma y = 450$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(y-\bar{y})^2 = 1{,}410$

$$\bar{x} = \frac{336}{12} = 28,\ \bar{y} = \frac{450}{15} = 30,\ S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2\right] = 71{\cdot}6$$

and $n_1 = 12$, $n_2 = 15$

Under null hypothesis (H_0) :

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

$$\therefore \quad t = \frac{28-30}{\sqrt{71{\cdot}6\left(\frac{1}{12}+\frac{1}{15}\right)}} = \frac{-2}{\sqrt{10{\cdot}74}} = -0{\cdot}609$$

Tabulated $t_{0{\cdot}05}$ for $(12 + 15 - 2) = 25$ *d.f.* is 2·06.

Conclusion. Since calculated $|t|$ is less than tabulated t, H_0 may be accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

Remark. Here $\bar{x}$ and $\bar{y}$ come out to be integral values and hence the direct method of computing $\Sigma(x-\bar{x})^2$ and $\Sigma(y-\bar{y})^2$ is used. In case $\bar{x}$ and (or) $\bar{y}$ comes out to be fractional, then the step deviation method is recommended for computation of $\Sigma(x-\bar{x})^2$ and $\Sigma(y-\bar{y})^2$.

Example 16·11. *Samples of two types of electric light bulbs were tested for length of life and following data were obtained :*

	Type I	*Type II*
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1{,}234$ *hrs.*	$\bar{x}_2 = 1{,}036$ *hrs.*
Sample S.D.'s	$s_1 = 36$ *hrs.*	$s_2 = 40$ *hrs.*

Is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life ?

Solution. *Null Hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e.,* the two types I and II of electric bulbs are indentical.

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$, *i.e.,* type I is superior to type II.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{13},$$

where
$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right]$$
$$= \frac{1}{n_1 + n_2 - 2}\left(n_1 s_1^2 + n_2 s_2^2\right) = \frac{1}{13}\left[8 \times (36)^2 + 7 \times (40)^2\right] = 1{,}659{\cdot}08$$

$$\therefore \quad t = \frac{1234 - 1036}{\sqrt{1659{\cdot}08\left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{198}{\sqrt{1659{\cdot}08 \times 0{\cdot}2679}} = 9{\cdot}39$$

Tabulated value of t for 13 *d.f.* at 5% level of significance for right (single)-tailed test is 1·77. [This is the value of $t_{0.10}$ for 13 *d.f.* from two-tail tables given at the end of the chapter.]

Conclusion. Since calculated 't' is much greater than tabulated 't', it is highly significant and H_0 is rejected. Hence the two types of electric bulbs differ significantly. Further, since $\bar{x}_1$ is much greater than $\bar{x}_2$, we conclude that type I is definitely superior to type II.

Example 16·12. *The heights of six randomly chosen sailors are (in inches) : 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss, the light that these data throw on the suggestion that sailors are on the average taller than soldiers.*

Solution. If the heights of sailors and soldiers be represented by the variables X and Y respectively then the *Null Hypothesis* is, $H_0 : \mu_X = \mu_Y$, *i.e., the sailors are not on the average taller than the soldiers.*

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$ (Right-tailed).

Under H_0, the test statistic is : $t = \dfrac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$

Sailors			*Soldiers*		
X	$d = X - A$ $= X - 68$	d^2	Y	$D = Y - B$ $= Y - 66$	D^2
63	−5	25	61	−5	25
65	−3	9	62	−4	16
68	0	0	65	−1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
			71	5	25
Total	0	60	72	6	36
			73	7	49
			Total	18	186

$$\bar{x} = A + \frac{\sum d}{n_1} = 68 + 0 = 68 \qquad \bar{y} = B + \frac{\sum D}{n_2} = 66 + \frac{18}{10} = 67 \cdot 8$$

$\therefore$ and $\sum(x-\bar{x})^2 = \sum d^2 - \frac{(\sum d)^2}{n_1} = 60 - 0 = 60$ and $\sum(y-\bar{y})^2 = \sum D^2 - \frac{(\sum D)^2}{n_2} = 186 - \frac{324}{10} = 153 \cdot 6$

$$S^2 = \frac{1}{n_1+n_2-2}\left[\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2\right] = \frac{1}{14}(60 + 153 \cdot 6) = 15 \cdot 2571$$

$$\therefore \quad t = \frac{68 - 67 \cdot 8}{\sqrt{15 \cdot 2571}\left(\frac{1}{6}+\frac{1}{10}\right)^{1/2}} = \frac{0 \cdot 2}{\sqrt{15 \cdot 2571 \times 0 \cdot 2667}} = 0 \cdot 099$$

Tabulated $t_{0 \cdot 05}$ for 14 *d.f.* for single-tail test is 1·76.

Conclusion. Since calculated t is much less than 1·76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

Example 16·13. *To test the claim that the resistance of electric wire can be reduced by at least 0·05 ohm by alloying, 25 values obtained for each alloyed wire and standard wire produced the following results :*

	Mean	*Standard deviation*
Alloyed wire	*0·083 ohm*	*0·003 ohm*
Standard wire	*0·136 ohm*	*0·002 ohm*

Test at 5% level whether or not the claim is substantiated.

Solution. *Null Hypothesis* $H_0 : \mu_1 - \mu_2 \geq 0 \cdot 05$, [*i.e.,* the claim is substantiated]

Alternative Hypothesis $H_1 : \mu_1 - \mu_2 < 0 \cdot 05$ (Left-tailed, test)

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{25 \times (0 \cdot 003)^2 + 25 \times (0 \cdot 002)^2}{25 + 25 - 2} = \frac{0 \cdot 000225 + 0 \cdot 0001}{48} = 0 \cdot 0000067$

$$\therefore \quad t = \frac{(0 \cdot 083 - 0 \cdot 136) - 0 \cdot 05}{\sqrt{\left\{0 \cdot 0000067\left(\frac{1}{25}+\frac{1}{25}\right)\right\}}} = -\frac{0 \cdot 103}{0 \cdot 00071} = -145 \cdot 07$$

The (critical) tabulated value of t for 48 *d.f.*, at 5% level of significance for left-tailed test is – 1·645.

Conclusion. Since calculated value of t is much less than tabulated value of t, it falls in the rejection region. We, therefore, reject the null hypothesis and conclude that the claim is not substantiated.

Example 16·14. *A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure :*

5, 2, 8, –1, 3, 0, –2, 1, 5, 0, 4 and 6

Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure ?

Solution. Here we are given the increments in blood pressure, *i.e.*, $d_i \ (= x_i - y_i)$.

Null Hypothesis, $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in the blood pressure readings of the patients before and after the drug. In other words, the given increments are just by chance (fluctuations of sampling) and not due to the stimulus.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$, *i.e.*, the stimulus results in an increase in blood pressure.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$

d	5	2	8	−1	3	0	−2	1	5	0	4	6	31
d^2	25	4	64	1	9	0	4	1	25	0	16	36	185

$$\bar{d} = \frac{1}{n} \Sigma d = 2{\cdot}58 \quad \text{and} \quad S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = 9{\cdot}5382$$

$$\therefore \qquad t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2{\cdot}58 \times \sqrt{12}}{\sqrt{9{\cdot}5382}} = \frac{2{\cdot}58 \times 3{\cdot}464}{3{\cdot}09} = 2{\cdot}89$$

Tabulated $t_{0{\cdot}05}$ for 11 *d.f.* for single-tail test is 1·80. [This is the value of $t_{0{\cdot}10}$ for 11 *d.f.* in the table for two-tail test given at the end of the chapter.]

Conclusion. Since calculated $t > t_{0{\cdot}05}$, H_0 is rejected at 5% level of significance. Hence we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

Example 16·15. *In a certain experiment to compare two types of animal foods A and B, the following results of increase in weights were observed in animals :*

Animal number		1	2	3	4	5	6	7	8	*Total*
Increase weight in lb	*Food A*	49	53	51	52	47	50	52	53	407
	Food B	52	55	52	53	50	54	54	53	423

(*i*) *Assuming that the two samples of animals are independent, can we conclude that food B is better than food A ?*

(*ii*) *Also examine the case when the same set of eight animals were used in both the foods.*

Solution. *Null Hypothesis*, H_0 : If the increase in weights due to foods A and B are denoted by X and Y respectively, then $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in increase in weights due to diets A and B.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$ (Left-tailed).

(*i*) If the two samples of animals be assumed to be independent, then we will apply *t*-test for difference of means to test H_0.

Test Statistic. Under $H_0 : \mu_X = \mu_Y$, the test criterion is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Food A			Food B		
X	d = X − 50	d^2	Y	D = Y − 52	D^2
49	−1	1	52	0	0
53	3	9	55	3	9
51	1	1	52	0	0
52	2	4	53	1	1
47	−3	9	50	−2	4
50	0	0	54	2	4
52	2	4	54	2	4
53	3	9	53	1	1
Total	7	37		7	23

$$\therefore \bar{x} = 50 + \frac{7}{8} = 50{\cdot}875 \qquad \bar{y} = 52 + \frac{7}{8} = 52{\cdot}875$$

and
$$\Sigma(x-\bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1} = 37 - \frac{49}{8} = 30{\cdot}875 \qquad \Sigma(y-\bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2} = 23 - \frac{49}{8} = 16{\cdot}875$$

$$S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2\right] = \frac{1}{14}(30{\cdot}875 + 16{\cdot}875) = 3{\cdot}41$$

$$\therefore \quad t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{50{\cdot}875 - 52{\cdot}875}{\sqrt{3{\cdot}41\left(\frac{1}{8}+\frac{1}{8}\right)}} = -2{\cdot}17$$

Tabulated $t_{0{\cdot}05}$ for $(8 + 8 - 2) = 14$ *d.f.* for one-tail test is 1·76.

Conclusion. The critical region for the left-tail test is $t < -1{\cdot}76$. Since calculated t is less than −1·76, H_0 is rejected at 5% level of significance. Hence we conclude that the foods *A* and *B* differ significantly as regards their effect on increase in weight. Further, since $\bar{y} > \bar{x}$, food *B* is superior to food *A*.

(*ii*) If the same set of animals is used in both the cases, then the readings *X* and *Y* are not independent but they are paired together and we apply the paired *t*-test for testing H_0.

Under $H_0 : \mu_X = \mu_Y$, the statistic is : $t = \dfrac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$

X	49	53	51	52	47	50	52	53	Total
Y	52	55	52	53	50	54	54	53	
d = X − Y	−3	−2	−1	−1	−3	−4	−2	0	−16
d^2	9	4	1	1	9	16	4	0	44

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-16}{8} = -2 \text{ and } S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{7}\left(44 - \frac{256}{8}\right) = 1{\cdot}714$$

$$\therefore \quad |t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{2}{\sqrt{1{\cdot}7143/8}} = \frac{2}{0{\cdot}4629} = 4{\cdot}32$$

Tabulated $t_{0{\cdot}05}$ for $(8-1) = 7$ *d.f.* for one-tail test is 1·90.

Conclusion. Here also the observed value of 't' is significant at 5% level of significance and we conclude that food B is superior to food A.

Example 16·16. *Two laboratories carry out independent estimates of a particular chemicals in a medicine produced by a certain firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The following data is obtained :*

No. of samples	10
Mean value of the difference of estimates	0·6
Sum of the squares of the differences (from their means)	20

Is the difference significant ? (Value of t at 5% level for 9 d.f. is 2·262.)

Solution. Let d stand for the difference between the estimates of the chemical between the two halves of each batch, and $\bar{d}$ the mean value of the difference of estimates. In usual notations, we are given :

$$n = 10, \ \bar{d} = 0{\cdot}6, \ \Sigma(d - \bar{d})^2 = 20$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$, *i.e.*, the difference is insignificant.

Alternative hypothesis, $H_1 : \mu_1 \neq \mu_2$.

Test Statistic. Under H_0, the test statistic is : $t = \dfrac{\bar{d}}{\sqrt{S^2/n}} \sim t_{10-1}$

where $\quad S^2 = \dfrac{1}{n-1}\Sigma(d-\bar{d})^2 = \dfrac{20}{9} = 2{\cdot}22 \quad \therefore \quad t = \dfrac{0{\cdot}6}{\sqrt{2{\cdot}22/10}} = \dfrac{0{\cdot}6}{0{\cdot}471} = 1{\cdot}274.$

The tabulated value of t at 5% level for 9 *d.f.*, is 2·262 (given).

Conclusion. Since calculated value of t is less than tabulated value of t, it is not significant. Hence, we may accept the null hypothesis and conclude that the difference is not significant.

16·3·4. t-test for Testing the Significance of an Observed Sample Correlation Coefficient. If r is the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis, $H_0 : \rho = 0$, *i.e., population correlation coefficient is zero*, the statistic :

$$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)} \qquad \ldots(16{\cdot}10)$$

follows Student's t-distribution with $(n-2)$ *d.f.* (*c.f. Remark to* § 16·4).

If the value of t comes out to be significant, we reject H_0 at the level of significance adopted and conclude that $\rho \neq 0$, *i.e.*, 'r' is significant of correlation in the population.

If t comes out to be non-significant, then H_0 may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

16·5. F-DISTRIBUTION

Definition. *If X and Y are two independent chi-square variates with υ_1 and υ_2 d.f. respectively, then F-statistic is defined by*

$$F = \frac{X/\nu_1}{Y/\nu_2} \qquad \text{...(16·13)}$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (υ_1, υ_2) *d.f.* with probability function given by :

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{\left(1 + \frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, 0 \le F < \infty \qquad \text{...[16·13(a)]}$$

Remarks 1. The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom υ_1 and υ_2.

2. A statistic F following Snedecor's F-distribution with (υ_1, υ_2) *d.f.* will be denoted by $F \sim F(\upsilon_1, \upsilon_2)$.

16·5·1 Derivation of Snedecor's F-distribution. Since X and Y are independent chi-square variates with ν_1 and ν_2 *d.f.* respectively, their joint probability density function is given by :

$$f(x, y) = \left\{\frac{1}{2^{\nu_1/2}\,\Gamma(\nu_1/2)} \exp(-x/2)\, x^{(\nu_1/2)-1}\right\} \times \left\{\frac{1}{2^{\nu_2/2}\,\Gamma(\nu_2/2)} \exp(-y/2)\, y^{(\nu_2/2)-1}\right\}$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\{-(x+y)/2\} \times x^{(\nu_1/2)-1}\, y^{(\nu_2/2)-1}, 0 \le (x, y) < \infty$$

Let us make the following transformation of variables :

$F = \dfrac{x/\nu_1}{y/\nu_2}$ and $u = y$, so that $0 \le F < \infty, 0 < u < \infty$ $\qquad \therefore \quad x = \dfrac{\nu_1}{\nu_2} Fu$ and $y = u$

Jacobian of transformation J is given by :
$$J = \frac{\partial(x, y)}{\partial(F, u)} = \begin{vmatrix} \frac{\nu_1}{\nu_2} u & 0 \\ \frac{\nu_1}{\nu_2} F & 1 \end{vmatrix} = \frac{\nu_1 u}{\nu_2}$$

Thus the joint *p.d.f.* of the transformed variables is :

$$g(F, u) = \frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\left\{-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2}F\right)\right\} \times \left(\frac{\nu_1}{\nu_2} Fu\right)^{(\nu_1/2)-1} u^{(\nu_2/2)-1}\, |J|$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\left\{-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2}F\right)\right\} \times u^{[(\nu_1+\nu_2)/2]-1}\, F^{(\nu_1/2)-1};\ 0 < u < \infty, 0 \le F < \infty$$

Integrating *w.r. to u* over the range 0 to ∞, the *p.d.f.* of F becomes :

$$g_1(F) = \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2/2)}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \times \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} u^{[(\nu_1+\nu_2)/2]-1}\, du\right]$$

$$= \frac{(\nu_1/\nu_2)^{(\nu_1/2)} F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \times \frac{\Gamma[(\nu_1+\nu_2)/2]}{\left[\frac{1}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right]^{(\nu_1+\nu_2)/2}}$$

$$\therefore \quad g_1(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, 0 \le F < \infty$$

which is the required probability function of F-distribution with (ν_1, ν_2) *d.f.*

Aliter.
$$F = \frac{x/\nu_1}{y/\nu_2}$$

$\therefore \quad \dfrac{\nu_1}{\nu_2} F = \dfrac{x}{y}$, being the ratio of two independent chi-square variates with ν_1 and ν_2 *d.f.* respectively is a $\beta_2\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$ variate. Hence the probability function of F is given by :

$$dP(F) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{\left(\frac{\nu_1}{\nu_2}F\right)^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}} d\left(\frac{\nu_1}{\nu_2}F\right)$$

$$\Rightarrow \quad f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, 0 \le F < \infty$$

16·5·2. Constants of F-distribution.

$$\mu'_r \text{ (about origin)} = E(F^r) = \int_0^\infty F^r f(F)\, dF$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty F^r \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}} dF \qquad \ldots(*)$$

To evaluate the integral, put : $\dfrac{\nu_1}{\nu_2}F = y$, so that $dF = \dfrac{\nu_2}{\nu_1}dy$

$$\therefore \quad \mu_r' = \frac{[\nu_1/\nu_2]^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2}{\nu_1}y\right)^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy$$

$$= \frac{\left(\frac{v_2}{v_1}\right)^r}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \int_0^\infty \frac{y^{r+(v_1/2)-1}}{(1+y)^{(v_1/2)+r+\{(v_2/2)-r\}}}\, dy$$

$$= \left(\frac{v_2}{v_1}\right)^r \cdot \frac{1}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot B\left(r + \frac{v_1}{2}, \frac{v_2}{2} - r\right),\ v_2 > 2r \qquad \ldots(16 \cdot 14)$$

Aliter for (16·14). (16·14) could also be obtained by substituting $\frac{v_1}{v_2} F = \tan^2 \theta$ in (*) and using the Beta integral :

$$2\int_0^{\pi/2} \sin^p \theta \cos^q \theta \, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right)$$

$$\therefore \quad \mu_r' = \left(\frac{v_2}{v_1}\right)^r \cdot \frac{\Gamma[r + (v_1/2)]\,\Gamma[(v_2/2) - r]}{\Gamma(v_1/2)\,\Gamma(v_2/2)}; r < \frac{v_2}{2} \quad \Rightarrow \quad v_2 > 2r \qquad \ldots(16 \cdot 15)$$

In particular

$$\mu'_1 = \frac{v_2}{v_1} \cdot \frac{\Gamma[1 + (v_1/2)]\,\Gamma[(v_2/2) - 1]}{\Gamma(v_1/2)\,\Gamma(v_2/2)} = \frac{v_2}{v_2 - 2},\ v_2 > 2 \qquad \ldots(16 \cdot 15a)$$

$$[\because \Gamma(r) = (r-1)\,\Gamma(r-1)]$$

Thus the mean of F-distribution is independent of v_1.

$$\mu_2' = \left(\frac{v_2}{v_1}\right)^2 \cdot \frac{\Gamma[(v_1/2) + 2]\,\Gamma[(v_2/2) - 2]}{\Gamma(v_1/2)\,\Gamma(v_2/2)}$$

$$= \left(\frac{v_2}{v_1}\right)^2 \cdot \frac{[(v_1/2) + 1]\,(v_1/2)}{[(v_2/2) - 1]\,[(v_2/2) - 2]} = \frac{v_2^2(v_1 + 2)}{v_1(v_2 - 2)\,(v_2 - 4)},\ v_2 > 4.$$

$$\therefore \ \mu_2 = \mu_2' - \mu_1'^2 = \frac{v_2^2(v_1 + 2)}{v_1(v_2 - 2)(v_2 - 4)} - \frac{v_2^2}{(v_2 - 2)^2} = \frac{2v_2^2\,(v_2 + v_1 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)},\ v_2 > 4 \qquad \ldots(16 \cdot 15b)$$

Similarly, on putting $r = 3$ and 4 in μ_r', we get μ_3' and μ_4' respectively, from which the central moments μ_3 and μ_4 can be obtained.

Remark. It has been proved that for large degrees of freedom, v_1 and v_2, F tends to $N[1, 2\{(1/v_1) + (1/v_2)\}]$ variate.

16·5·3. Mode and Points of Inflexion of F-distribution. We have

$$\log f(F) = C + \{(v_1/2) - 1\} \log F - \left(\frac{v_1 + v_2}{2}\right) \log \{1 + (v_1/v_2) F\},$$

where C is a constant independent of F.

$$\frac{\partial}{\partial F}[\log f(F)] = \left(\frac{v_1}{2} - 1\right) \cdot \frac{1}{F} - \frac{(v_1 + v_2)}{2} \cdot \frac{1}{\left(1 + \frac{v_1}{v_2} F\right)} \cdot \frac{v_1}{v_2}$$

$$f'(F) = \frac{\partial}{\partial F} f(F) = 0 \quad \Rightarrow \quad \frac{v_1 - 2}{2F} - \frac{v_1\,(v_1 + v_2)}{2(v_2 + v_1 F)} = 0$$

Hence

$$F = \frac{v_2\,(v_1 - 2)}{v_1\,(v_2 + 2)}$$

It can be easily verified that at this point $f''(F) < 0$. Hence mode $= \frac{v_2\,(v_1 - 2)}{v_1\,(v_2 + 2)}$ ...(16·16)

***Remarks* 1.** Since $F > 0$, mode exists if and only if $\nu_1 > 2$.

2. $$\text{Mode} = \left(\frac{\nu_2}{\nu_2+2}\right)\cdot\left(\frac{\nu_1-2}{\nu_1}\right)$$

Hence mode of *F*-distribution is always less than unity.

3. The points of inflexion of *F*-distribution exist for $\upsilon_1 > 4$ and are equidistant from mode.

Proof. We have $$\frac{\nu_1}{\nu_2}F = \frac{X}{Y} \sim \beta_2(l, m), \quad \ldots(*)$$

where $l = \upsilon_1/2$ and $m = \upsilon_2/2$. We now find the points of inflexion of Beta distribution of second kind with parameters l and m. If $X \sim \beta_2(l, m)$, its *p.d.f.* is :

$$f(x) = \frac{1}{\beta(l, m)}\cdot\frac{x^{l-1}}{(1+x)^{l+m}} \; ; 0 \le x < \infty \quad \ldots(**)$$

Points of inflexion are the solution of $f''(x) = 0$ and $f'''(x) \ne 0$

From (**), $\log f(x) = -\log\beta(l, m) + (l-1)\log x - (l+m)\log(1+x)$

Differentiating twice *w.r. to x*, we get

$$\frac{f'(x)}{f(x)} = \frac{l-1}{x} - \frac{l+m}{1+x} \quad \ldots(***)$$

and $$\frac{f(x)f''(x) - [f'(x)]^2}{[f(x)]^2} = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

If $f''(x) = 0$, then $$-\left[\frac{f'(x)}{f(x)}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \Rightarrow -\left[\frac{l-1}{x} - \frac{l+m}{1+x}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

[On using (***)]

$$\Rightarrow \frac{l-1}{x^2}\left(l-1-1\right) - 2\frac{(l-1)(l+m)}{x(1+x)} + \frac{l+m}{(1+x)^2}\times(l+m+1) = 0$$

$$\Rightarrow (l-1)(l-2)(1+x)^2 - 2x(1+x)(l-1)(l+m) + x^2(l+m)(l+m+1) = 0 \quad \ldots(****)$$

which is a quadratic in x. It can be easily verified that at these values of x, $f'''(x) \ne 0$, if $l > 2$.

The roots of (****) give the points of inflexion of $\beta_2(l, m)$ distribution. The sum of the points of inflexion is equal to the sum of roots of (****) and is given by :

$$-\left[\frac{\text{Coeff. of } x \text{ in } (****)}{\text{Coeff. of } x^2 \text{ in } (****)}\right] = -\left[\frac{2(l-1)(l-2) - 2(l-1)(l+m)}{(l-1)(l-2) - 2(l-1)(l+m) + (l+m)(l+m+1)}\right]$$

$$= \frac{2(l-1)[(l+m)-(l-2)]}{(l-1)(l-2) - (l-1)(l+m) - (l-1)(l+m) + (l+m)(l+m+1)}$$

$$= \frac{2(l-1)(m+2)}{(l-1)[(l-2-l-m)] + (l+m)(l+m+1-l+1)}$$

$$= \frac{2(l-1)(m+2)}{-(l-1)(m+2) + (l+m)(m+2)} = \frac{2(l-1)}{l+m-l+1} = \frac{2(l-1)}{(m+1)}$$

$\therefore$ Sum of points of inflexion of $\left(\frac{\nu_1}{\nu_2}F\right)$ distribution $= \dfrac{2(l-1)}{(m+1)} = \dfrac{2\left(\frac{\nu_1}{2}-1\right)}{\left(\frac{\nu_2}{2}+1\right)} = \dfrac{2(\nu_1-2)}{(\nu_2+2)}$

$\Rightarrow$ Sum of points of inflexion of $F(\upsilon_1, \upsilon_2)$ distribution

$$= \frac{\nu_2}{\nu_1}\cdot\frac{2(\nu_1-2)}{(\nu_2+2)}, \text{ provided } l = \frac{\nu_1}{2} > 2 = 2\,\frac{\nu_2(\nu_1-2)}{\nu_1(\nu_2+2)} = 2 \text{ Mode, provided } \nu_1 > 4$$

Hence the points of inflexion of $F(\upsilon_1, \upsilon_2)$ distribution, when they exist, (*i.e.*, when $\upsilon_1 > 4$) are equidistant from the mode.

4. Karl Pearson's coefficient of skewness is given by : $S_k = \dfrac{\text{Mean} - \text{Mode}}{\sigma} > 0$, since mean > 1 and mode < 1. Hence F-distribution is highly positively skewed.

5. The probability $f(F)$ increases steadily at first until it reaches its peak (corresponding to the modal value which is less than 1) and then decreases slowly so as to become tangential at $F = \infty$, i.e., F-axis is an asymptote to the right tail.

Example 16·20. *When $v_1 = 2$, show that the significance level of F corresponding to a significant probability p is :* $F = \dfrac{v_2}{2}\left[p^{-(2/v_2)} - 1\right]$ *where v_1 an v_2 have their usual meanings.*

Solution. When $v_1 = 2$,

$$f(F) = \frac{1}{B\left(1, \frac{v_2}{2}\right)} \cdot \frac{2}{v_2} \cdot \frac{1}{\left(1 + \frac{2}{v_2} F\right)^{(v_2/2)+1}}$$

$$= \frac{\Gamma\left(\frac{v_2}{2} + 1\right)}{\Gamma(1)\Gamma(v_2/2)} \times \frac{2/v_2}{\left(\frac{2}{v_2}\right)^{(v_2/2)+1}\left(F + \frac{v_2}{2}\right)^{(v_2/2)+1}} = \frac{\left(\frac{v_2}{2}\right)^{(v_2/2)+1}}{\left(F + \frac{v_2}{2}\right)^{(v_2/2)+1}}$$

Hence $$p = \int_F^\infty f(F)\, dF = \left[\frac{v_2}{2}\right]^{(v_2/2)+1} \times \int_F^\infty \frac{dF}{\left(F + \frac{v_2}{2}\right)^{(v_2/2)+1}}$$

$$= \left(\frac{v_2}{2}\right)^{(v_2/2)+1} \times \left|\frac{\left(F + \frac{v_2}{2}\right)^{-(v_2/2)}}{-\frac{v_2}{2}}\right|_F^\infty = \left[\frac{\left(\frac{v_2}{2}\right)}{F + \frac{v_2}{2}}\right]^{v_2/2} = \frac{1}{\left(1 + \frac{2}{v_2} F\right)^{v_2/2}}$$

$$\Rightarrow \quad p^{-(2/v_2)} = 1 + \frac{2F}{v_2} \quad \Rightarrow \quad F = \frac{v_2}{2}\left[p^{-(2/v_2)} - 1\right].$$

Example 16·21. *X is a binomial variate with parameters n and p and F_{v_1, v_2} is an F-statistic with v_1 and v_2 d.f. Prove that :*

$$P(X \le k - 1) = P\left[F_{2k,\, 2(n-k+1)} > \frac{n-k+1}{k} \cdot \frac{p}{1-p}\right].$$

Solution. If $X \sim B(n, p)$, then we have

$$P(X \le k-1) = (n-k+1) \cdot \binom{n}{k-1} \int_0^q t^{n-k}(1-t)^{k-1}\, dt$$

$$= \frac{1}{\beta(n-k+1, k)} \int_0^q t^{n-k}(1-t)^{k-1}\, dt \qquad \ldots (*)$$

[See Example 8·31.]

$$P = P_r\left[F_{2k, 2(n-k+1)} > \frac{n-k+1}{k}\left(\frac{p}{1-p}\right)\right] = \int_{\frac{n-k+1}{k} \cdot \frac{p}{q}}^\infty p[F_{2k,\, 2(n-k+1)}]\, dF$$

$$= \frac{1}{B(k, n-k+1)} \int_{\frac{n-k+1}{k} \cdot \frac{p}{q}}^\infty \frac{[k/(n-k+1)]^k \cdot F^{k-1}\, dF}{\left(1 + \frac{kF}{n-k+1}\right)^{n+1}} \qquad \ldots (**)$$

Put $1 + \dfrac{kF}{n-k+1} = \dfrac{1}{y} \quad \Rightarrow \quad F = \dfrac{n-k+1}{k}\left(\dfrac{1-y}{y}\right)$ and $dF = \dfrac{n-k+1}{k} \cdot \dfrac{-dy}{y^2}$

$F = \infty \;\Rightarrow\; y = 0$ and $F = \dfrac{n-k+1}{k} \cdot \dfrac{p}{q} \quad \Rightarrow \quad \dfrac{1}{y} = \dfrac{q+p}{q} = \dfrac{1}{q} \quad \Rightarrow \quad y = q$

Substituting in (**), we get :

$$P = \frac{1}{B(k, n-k+1)} \int_q^0 \left(\frac{1-y}{y}\right)^{k-1} \cdot y^{n+1} \left(\frac{-dy}{y^2}\right)$$

$$= \frac{1}{B(k,n-k+1)} \int_0^q y^{n-k}(1-y)^{k-1} dy \qquad \ldots (***)$$

From (*), (**) and (***), we get the result.

Example 16·22. *If $F(n_1, n_2)$ represents an F-variate with n_1 and n_2 d.f., prove that $F(n_2, n_1)$ is distributed as $1/F(n_1, n_2)$ variate. Deduce that*

$$P[F(n_1, n_2) \geq c] = P\left[F(n_2, n_1) \leq \frac{1}{c}\right]$$

Or

Show how probability points of $F(n_2, n_1)$ can be obtained from those of $F(n_1, n_2)$.

Solution. Let X and Y be independent chi-square variates with n_1 and n_2 d.f. respectively. Then by definition, we have

$$F = \frac{(X/n_1)}{(Y/n_2)} \sim F(n_1, n_2) \quad \text{and} \quad \frac{1}{F} = \frac{(Y/n_2)}{(X/n_1)} \sim F(n_2, n_1) \qquad \ldots (*)$$

Hence the result.

We have : $P\left[F(n_1, n_2) \geq c\right] = P\left[\dfrac{1}{F(n_1, n_2)} \leq \dfrac{1}{c}\right] = P\left[F(n_2, n_1) \leq \dfrac{1}{c}\right]$ [From (*)]

Remark. Probability Points of $F(n_2, n_1)$ from Those of $F(n_1, n_2)$ Distribution.

Let $P\left[F(n_1, n_2) \geq c\right] = \alpha$, *i.e.*, let c be the upper α-significant point of $F(n_1, n_2)$ distribution.

$$\therefore \quad 1 - \alpha = 1 - P\left[F(n_1, n_2) \geq c\right] = 1 - P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$\Rightarrow \quad \alpha = P\left[F(n_2, n_1) \leq \frac{1}{c}\right] = 1 - P\left[F(n_2, n_1) \geq \frac{1}{c}\right] \qquad \text{[From (*)]}$$

$$\therefore \quad P\left[F(n_2, n_1) \geq \frac{1}{c}\right] = 1 - \alpha$$

Thus $(1-\alpha)$ significant points of $F(n_2, n_1)$ distribution are the reciprocal of α-significant points of $F(n_1, n_2)$ distributions, *e.g.*, $F_{8,4}(0·05) = 6·04 \quad \Rightarrow \quad F_{4,8}(0·95) = \dfrac{1}{6·04}$.

Example 16·23. *Prove that if $n_1 = n_2$, the median of F-distribution is at $F = 1$ and that the quartiles Q_1 and Q_3 satisfy the condition $Q_1 Q_3 = 1$.*

Solution. Since $n_1 = n_2 = n$, (say), the median (M) of $F(n_1, n_2) = F(n, n)$ distribution is given by : $\quad P\left[F(n, n) \leq M\right] = 0·5 \qquad \ldots (*)$

(c) Since $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent, $X_1^2 \sim \chi^2_{(1)}$ and $X_2^2 \sim \chi^2_{(1)}$, are also independent. Hence by definition of F-statistic,

$$\frac{X_1^2/1}{X_2^2/1} \sim F_{(1, 1)} \quad \Rightarrow \quad \frac{X_1^2}{X_2^2} \sim F_{(1, 1)}$$

(d) X_1/X_2, being the ratio of two independent standard normal variates is a standard Cauchy variate.

16·6. APPLICATIONS OF F-DISTRIBUTION

F-distribution has the following applications in statistical theory.

16·6·1. F-test for Equality of Two Population Variances. Suppose we want to test (i) whether two independent samples x_i, $(i = 1, 2, \ldots, n_1)$ and y_j, $(j = 1, 2, \ldots, n_2)$ have been drawn from the normal populations with the same variance σ^2 (say), or (ii) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) that (i) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e., the population variances are equal, or (ii) Two independent estimates of the population variance are homogeneous,* the statistic F is given by :

$$F = \frac{S_X^2}{S_Y^2} \qquad \ldots(16·17)$$

where $\quad S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2 \quad$ and $\quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \qquad \ldots(16·17a)$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (ν_1, ν_2) *d.f.* where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

Proof. $F = \frac{S_X^2}{S_Y^2} = \left[\frac{n_1}{n_1 - 1} s_X^2\right] \Big/ \left[\frac{n_2}{n_2 - 1} s_Y^2\right]$

$$= \left[\frac{n_1 s_X^2}{\sigma_X^2} \cdot \frac{1}{(n_1 - 1)}\right] \Big/ \left[\frac{n_2 s_Y^2}{\sigma_Y^2} \cdot \frac{1}{(n_2 - 1)}\right] \qquad (\because \sigma_X^2 = \sigma_Y^2 = \sigma^2, \text{ under } H_0)$$

Since $\frac{n_1 s_X^2}{\sigma_X^2}$ and $\frac{n_2 s_Y^2}{\sigma_Y^2}$ are independent chi-square variates with $(n_1 - 1)$ and $(n_2 - 1)$ *d.f.* respectively, F follows Snedecor's F-distribution with $(n_1 - 1, n_2 - 1)$ *d.f.* (*c.f.* § 16·5).

Remarks 1. In (16·17), greater of the two variances S_X^2 and S_Y^2 is to be taken in the numerator and n_1 corresponds to the greater variance.

By comparing the calculated value of F obtained by using (16·17) for the two given samples, with the tabulated value of F for (n_1, n_2) *d.f.* at certain level of significance (5% or 1%), H_0 is either rejected or accepted.

2. *Critical values of F-distribution.* The available F-tables (given in Table II-A and II-B at the end of the chapter) give the critical values of F for the right-tailed test, *i.e.*, the critical region is determined by the right-tail areas. Thus the significant value $F_\alpha(n_1, n_2)$ at level of significance α and (n_1, n_2) *d.f.* is determined by :

$$P[F > F_\alpha(n_1, n_2)] = \alpha, \qquad \ldots(*)$$

as shown in the diagram on page 16·37.

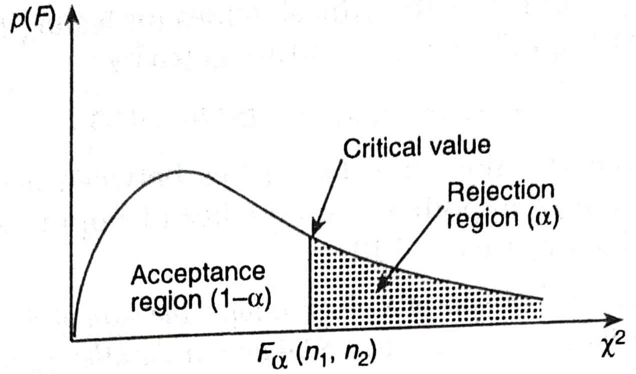


Fig. 16·3 : Critical Values of F-Distribution

From the *Remark to Example 16·23*, we have the following reciprocal relation between the upper and lower α-significant points of *F*-distribution :

$$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)} \quad \Rightarrow \quad F_\alpha(n_1, n_2) \times F_{1-\alpha}(n_2, n_1) = 1 \qquad \ldots(**)$$

The critical values of F for left tail–test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 < \sigma_2^2$ are given by $F < F_{n_1-1, n_2-1}(1-\alpha)$, and for the two – tailed test, $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$ are given by $F > F_{n_1-1, n_2-1}(\alpha/2)$ and $F < F_{n_1-1, n_2-1}(1-\alpha/2)$ [For details, see Chapter Eighteen.]

Example 16·25. *Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0·8 and 0·5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level. [Assume that $P(F_{10,8} \geq 3{\cdot}35) = 0{\cdot}05$ and $P(F_{8,10} \geq 3{\cdot}07) = 0{\cdot}05$.]*

Solution. We want to test *Null Hypothesis*, $H_0 : \sigma_X^2 = \sigma_Y^2$ against the *Alternative Hypothesis*, $H_1 : \sigma_X^2 \neq \sigma_Y^2$ (Two-tailed).

We are given : $n_1 = 11, n_2 = 9, s_X = 0{\cdot}8$ and $s_Y = 0{\cdot}5$.

Under the null hypothesis, $H_0 : \sigma^2_X = \sigma^2_Y$, the statistic :

$$F = \frac{S_X^2}{S_Y^2} \text{ follows } F \text{ distribution with } (n_1 - 1, n_2 - 1)\ d.f.$$

$$n_1 s_X^2 = (n_1 - 1) S_X^2 \Rightarrow S_X^2 = \left(\frac{n_1}{n_1 - 1}\right) s_X^2 = \left(\frac{11}{10}\right) \times (0{\cdot}8)^2 = 0{\cdot}704$$

Similarly, $$S_Y^2 = \left(\frac{n_2}{n_2 - 1}\right) s_Y^2 = \left(\frac{9}{8}\right) \times (0{\cdot}5)^2 = 0{\cdot}28125$$

$$\therefore \quad F = \frac{0{\cdot}704}{0{\cdot}28125} = 2{\cdot}5$$

The significant values of F for two-tailed test at level of significance $\alpha = 0{\cdot}10$ are :

$$\left.\begin{array}{ll} F > & F_{10,8}(\alpha/2) = F_{10,8}(0{\cdot}05) \\ \text{and} \quad F < & F_{10,8}(1 - \alpha/2) = F_{10,8}(0{\cdot}95) \end{array}\right\} \qquad \ldots(*)$$

We are given the tabulated (significant) values :

$$P(F_{10,8} \geq 3{\cdot}35) = 0{\cdot}05 \quad \Rightarrow \quad F_{10,8}(0{\cdot}05) = 3{\cdot}35 \qquad \ldots(**)$$

Also $$P(F_{8,10} \geq 3{\cdot}07) = 0{\cdot}05 \quad \Rightarrow \quad P\left(\frac{1}{F_{8,10}} \leq \frac{1}{3{\cdot}07}\right) = 0{\cdot}05$$

$$\Rightarrow \quad P(F_{10,8} \leq 0{\cdot}326) = 0{\cdot}05 \quad \Rightarrow \quad P(F_{10,8} \geq 0{\cdot}326) = 0{\cdot}95 \qquad \ldots(***)$$

Hence from (*), (**) and (***), the critical values for testing $H_0 : \sigma_X^2 = \sigma_Y^2$, against $H_1 : \sigma_X^2 \neq \sigma_Y^2$ at level of significance $\alpha = 0{\cdot}10$ are given by :

$$F > 3{\cdot}35 \quad \text{and} \quad F < 0{\cdot}326 = 0{\cdot}33$$

Since, the calculated value of F (=2·5) lies between 0·33 and 3·35, it is not significant and hence null hypothesis of equality of population variances may be accepted at level of significance $\alpha = 0{\cdot}10$.

Example 16·26. *In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84·4 and in the other sample of 10 observations it was 102·6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3·29.*

Solution. Here $n_1 = 8, n_2 = 10$ and $\Sigma(x - \bar{x})^2 = 84{\cdot}4, \ \Sigma(y - \bar{y})^2 = 102{\cdot}6$

$$\therefore \quad S_X^2 = \frac{1}{n_1 - 1}\Sigma(x - \bar{x})^2 = \frac{84{\cdot}4}{7} = 12{\cdot}057$$

$$S_Y^2 = \frac{1}{n_2 - 1}\Sigma(y - \bar{y})^2 = \frac{102{\cdot}6}{9} = 11{\cdot}4$$

Under $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e.,* the estimates of σ^2 given by the samples are homogeneous, the test statistic is :

$$F = \frac{S_X^2}{S_Y^2} = \frac{12{\cdot}057}{11{\cdot}4} = 1{\cdot}057$$

Tabulated $F_{0{\cdot}05}$ for (7, 9) *d.f.* is 3·29.

Since calculated $F < F_{0{\cdot}05}$, H_0 may be accepted at 5% level of significance.

Example 16·27. *Two random samples gave the following results :*

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

[*Given* : $F_{0{\cdot}05}(9, 11) = 2{\cdot}90$, $F_{0{\cdot}05}(11, 9) = 3{\cdot}10$ (approx.) and $t_{0{\cdot}05}(20) = 2{\cdot}086$, $t_{0{\cdot}05}(22) = 2{\cdot}07$]

Solution. A normal population has two parameters, *viz.,* mean μ and variance σ^2. To test if two independent samples have been drawn from the same normal population, we have to test (*i*) the equality of population means, and (*ii*) the equality of population variances.

Null Hypothesis : The two samples have been drawn from the same normal population, *i.e.,*

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad \sigma_1^2 = \sigma_2^2.$$

Equality of means will be tested by applying t-test and equality of variances will be tested by applying F-test. Since t-test assumes $\sigma_1^2 = \sigma_2^2$, we shall first apply F-test and then t-test. In usual notations, we are given :

$$n_1 = 10, \quad n_2 = 12; \quad \bar{x}_1 = 15, \quad \bar{x}_2 = 14, \quad \Sigma(x_1 - \bar{x}_1)^2 = 90, \quad \Sigma(x_2 - \bar{x}_2)^2 = 108.$$

F-test : Here

$$S_1^2 = \frac{1}{n_1 - 1}\Sigma(x_1 - \bar{x}_1)^2 = \frac{90}{9} = 10, \quad S_2^2 = \frac{1}{n_2 - 1}\Sigma(x_2 - \bar{x}_2)^2 = \frac{108}{11} = 9{\cdot}82$$

Since $S_1^2 > S_2^2$, under $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) = F(9, 11)$$

$$\therefore \quad F = \frac{S_1^2}{S_2^2} = \frac{10}{9{\cdot}82} = 1{\cdot}018$$

Tabulated $F_{0{\cdot}05}$ (9, 11) = 2·90. Since calculated F is less than tabulated F, it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Since $\sigma_1^2 = \sigma_2^2$, we can now apply t test for testing $H_0 : \mu_1 = \mu_2$.

t-test : Under $H_0' : \mu_1 = \mu_2$, against alternative hypothesis, $H_1' : \mu_1 \neq \mu_2$, the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2} = t_{20}$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right] = \frac{1}{20}(90 + 108) = 9{\cdot}9$$

$$\therefore \quad t = \frac{15 - 14}{\sqrt{9{\cdot}9\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{1}{\sqrt{9{\cdot}9 \times \frac{11}{60}}} = \frac{1}{\sqrt{1{\cdot}815}} = 0{\cdot}742)$$

Tabulated $t_{0{\cdot}05}$ for 20 $d.f.$ = 2·086. Since $| t | < t_{0{\cdot}05}$, it is not significant. Hence the hypothesis $H_0' : \mu_1 = \mu_2$ may be accepted. Since both the hypotheses, *i.e.*, $H_0' : \mu_1 = \mu_2$ and $H_0 : \sigma_1^2 = \sigma_2^2$ are accepted, we may regard that the given samples have been drawn from the same normal population.

16·6·2. F-test for Testing the Significance of an Observed Multiple Correlation Coefficient. If R is the observed multiple correlation coefficient of a variate with k other variates in a random sample of size n from a $(k + 1)$ variate population, then Prof. R.A. Fisher proved that under the null hypothesis (H_0) *that the multiple correlation coefficient in the population is zero, the statistic* :

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} \qquad \ldots(16{\cdot}18)$$

conforms to F-distribution with $(k, n - k - 1)$ $d.f.$

16·6·3. F-test for Testing the Significance of an Observed Sample Correlation Ratio η_{YX}. Under the null hypothesis that *population correlation ratio is zero,* the test statistic is :

$$F = \frac{\eta^2}{1 - \eta^2} \cdot \frac{N - h}{h - 1} \sim F(h - 1, N - h) \qquad \ldots(16{\cdot}19)$$

where N is the size of the sample (from a bi-variate normal population) arranged in h arrays.

16·6·4. F-test for Testing the Linearity of Regression. For a sample of size N arranged in h arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is :

$$F = \frac{\eta^2 - r^2}{1 - \eta^2} \cdot \frac{N - h}{h - 2} \sim F(h - 2, N - h) \qquad \ldots(16\cdot20)$$

16·6·5. F-test for Equality of Several Means. This test is carried out by the technique of Analysis of Variance, which plays a very important and fundamental role in Design of Experiments in Agricultural Statistics.

[For a detailed discussion of the Analysis of Variance Technique, see Fundamentals of Applied Statistics by the same authors.]

16·7. RELATION BETWEEN t AND F DISTRIBUTIONS

In F-distribution with (υ_1, υ_2) *d.f.* [*c.f.* 16·5 (*a*)], take $\upsilon_1 = 1$, $\upsilon_2 = \upsilon$ and $t^2 = F$, *i.e.,* $dF = 2t\,dt$. Thus the probability differential of F transforms to :

$$dG(t) = \frac{(1/\nu)^{1/2}}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} 2t\,dt, \; 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} dt, \; -\infty < t < \infty$$

the factor 2 disappearing since the total probability in the range $(-\infty, \infty)$ is unity. This is the probability function of Student's t-distribution with υ *d.f.* Hence we have the following relation between t and F distributions.

If a statistic t follows Student's t distribution with n d.f., then t^2 follows Snedecor's F-distribution with (1, n) d.f. Symbolically,

$$\left.\begin{array}{ll} \text{if} & t \sim t_{(n)} \\ \text{then} & t^2 \sim F_{(1, n)} \end{array}\right\} \qquad \ldots(16\cdot21)$$

Aliter Proof of (16·21). If $\xi \sim N(0, 1)$ and $X \sim \chi^2_{(n)}$ are independent *r.v.'s*, then :

$$U = \xi^2 \sim \chi^2_{(1)} \text{ [Square of a } S.N.V.]$$

and
$$t = \frac{\xi}{\sqrt{X/n}} \sim t_{(n)} \quad \Rightarrow \quad t^2 = \frac{\xi^2}{(X/n)} = \frac{(\xi^2/1)}{(X/n)},$$

being the ratio of two independent chi-square variates divided by their respective degrees of freedom is $F(1, n)$ variate. Hence $t^2 \sim F(1, n)$.

With the help of relation (16·21), all the uses of t-distribution can be regarded as the applications of F-distribution also, *e.g.,* for test for a single mean, instead of computing $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$, we may compute $F = t^2 = \dfrac{n(\bar{x} - \mu)^2}{S^2}$ and then apply F-test with $(1, n)$ *d.f.*, and so on.

Similarly, we can write the test statistic F from § 16·3·4, § 16·3·5 and § 16·3·6 for testing the significance of an observed sample correlation coefficient, regression coefficient and partial correlation coefficient respectively.

Example 16·28. Given : $P[F(10, 12) > 2.753] = 0.05 = P[F(1, 12) > 4.747]$,

find $P[F(12, 10) > (2.753)^{-1}]$, and $P(-\sqrt{4.747} < t_{12} < \sqrt{4.747})$

Solution.
$$P[F(12, 10) > (2.753)^{-1}] = P\left[\frac{1}{F(12, 10)} < 2.753\right] = P[F(10, 12) < 2.753]$$
$$= 1 - P[F(10, 12) > 2.753] = 1 - 0.05 = 0.95$$
$$P(-\sqrt{4.747} < t_{12} < \sqrt{4.747}) = P(t^2_{12} < 4.747) = P[F(1, 12) < 4.747]$$
$$= 1 - P[F(1, 12) > 4.747] = 1 - 0.05 = 0.95$$

16·8. RELATION BETWEEN F AND χ^2 DISTRIBUTION

In $F(n_1, n_2)$ distribution if we let $n_2 \to \infty$, then $\chi^2 = n_1 F$ follows χ^2-distribution with n_1 d.f. ... (16·22)

Proof. We have

$$f(F) = \frac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma[(n_1+n_2)/2]}{\left(1 + \frac{n_1}{n_2}F\right)^{(n_1+n_2)/2}}, \quad 0 < F < \infty$$

In the limit as $n_2 \to \infty$, we have

$$\frac{\Gamma[(n_1+n_2)/2]}{n_2^{n_1/2}\,\Gamma(n_2/2)} \to \frac{(n_2/2)^{n_1/2}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}} \quad \dots (*)$$

$$\left[\because \frac{\Gamma(n+k)}{\Gamma(n)} \to n^k \text{ as } n \to \infty. \text{ (c.f. Remark below)}\right]$$

Also
$$\lim_{n_2\to\infty}\left(1 + \frac{n_1}{n_2}F\right)^{(n_1+n_2)/2} = \lim_{n_2\to\infty}\left[\left(1+\frac{n_1}{n_2}F\right)^{n_2}\right]^{1/2} \times \lim_{n_2\to\infty}\left(1+\frac{n_1}{n_2}F\right)^{n_1/2}$$
$$= \exp(n_1F/2) = \exp(\chi^2/2) \qquad (\because n_1F = \chi^2) \quad \dots (**)$$

Hence in the limit, on using (*) and (**), the *p.d.f.* of $\chi^2 = n_1F$ becomes :

$$dP(\chi^2) = \frac{(n_1/2)^{n_1/2}\, e^{-\chi^2/2}}{\Gamma(n_1/2)} \cdot \left(\frac{\chi^2}{n_1}\right)^{(n_1/2)-1} d\left(\frac{\chi^2}{n_1}\right)$$
$$= \frac{1}{2^{n_1/2}\,\Gamma(n_1/2)} \cdot e^{-\chi^2/2}\,(\chi^2)^{(n_1/2)-1}\, d\chi^2, \quad 0 < \chi^2 < \infty$$

which is the p.d.f. of chi-square distribution with n_1 *d.f.*

Remark.
$$\lim_{n\to\infty}\frac{\Gamma(n+k)}{\Gamma(n)} = \lim_{n\to\infty}\frac{(n+k-1)!}{(n-1)!}, \text{ (for large } n) \simeq \lim_{n\to\infty}\frac{\sqrt{2\pi}\, e^{-(n+k-1)}\,(n+k-1)^{n+k-(1/2)}}{\sqrt{2\pi}\, e^{-(n-1)}\,(n-1)^{n-(1/2)}}$$

(On using Stirling's approximation for $n!$ as $n \to \infty$.)

$$= e^{-k}\lim_{n\to\infty}\frac{n^{n+k-\frac{1}{2}}\left(1+\frac{k-1}{n}\right)^{n+k-\frac{1}{2}}}{n^{n-\frac{1}{2}}\left(1-\frac{1}{n}\right)^{n-\frac{1}{2}}} = e^{-k}\,n^k\,\frac{\lim\limits_{n\to\infty}\left(1+\frac{k-1}{n}\right)^n \lim\limits_{n\to\infty}\left(1+\frac{k-1}{n}\right)^{k-\frac{1}{2}}}{\lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right)^n \lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right)^{-\frac{1}{2}}}$$

$$= e^{-k}\,n^k\left[\frac{e^{(k-1)}}{e^{-1}} \cdot \frac{1}{1}\right] = n^k$$

$$\therefore \quad \lim_{n\to\infty}\frac{\Gamma(n+k)}{\Gamma n} = n^k \quad \dots (16\cdot22a)$$

16·9. FISHER'S z-DISTRIBUTION

In G.W. Snedecor's F-distribution with (ν_1, ν_2) *d.f.*, if we put

$$F = \exp(2Z) \quad \Rightarrow \quad Z = \frac{1}{2}\log_e F, \qquad \ldots(16\cdot23)$$

the distribution of Z becomes

$$g(z) = p(F).\left|\frac{dF}{dz}\right| = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{(e^{2z})^{(\nu_1/2)-1}\, 2e^{2z}}{\left[1 + \frac{\nu_1}{\nu_2} e^{2z}\right]^{(\nu_1+\nu_2)/2}}$$

$$= 2\, \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{e^{\nu_1 z}}{\left[1 + \frac{\nu_1}{\nu_2} e^{2z}\right]^{(\nu_1+\nu_2)/2}}\ ; -\infty < z < \infty \qquad \ldots(16\cdot24)$$

which is the probability function of Fisher's z-distribution with (ν_1, ν_2) *d.f.* The tables of significant values z_0 of z which will be exceeded in random sampling with probabilities 0·05 and 0·01, *i.e.*, $P(z > z_0) = 0\cdot05$ and $P(z > z_0') = 0\cdot01$ corresponding to various *d.f.* (ν_1, ν_2) were published by Fisher (*c.f.* Statistical Methods for Research Workers) in 1925. From these tables, Snedecor (1934-38) by using (16·23) deduced the tables of significant values of the variance ratio which he denoted by F in honour of Prof. R.A. Fisher.

Remark. With the help of relation (16·23), all the applications of F-distribution may be regarded as the applications of z-distribution also.

16·9·1. Moment Generating Function of z-distribution.

$$M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} g(z)\, dz = \int_0^{\infty} F^{t/2} f(F)\, dF \qquad [\because e^{2z} = F]$$

Since μ_r' (about origin) for F-distribution is $\int_0^{\infty} F^r f(F)\, dF$, we can find *m.g.f.* of the z-distribution by putting $r = t/2$ in the expression for μ_r' for F-distribution.

Hence
$$M_Z(t) = \left(\frac{\nu_2}{\nu_1}\right)^{t/2} \cdot \frac{\Gamma\{(\nu_1+t)/2\}\,\Gamma\{(\nu_2-t)/2\}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \qquad [\textit{c.f.} \text{ Equation } (16\cdot15)]$$

$$\Rightarrow \quad K_Z(t) = \log M_Z(t)$$

$$= \frac{t}{2}(\log \nu_2 - \log \nu_1) + \log\Gamma\{(\nu_1+t)/2\} + \log\Gamma\{(\nu_2-t)/2\} - \log\Gamma(\nu_1/2) - \log\Gamma(\nu_2/2)$$

Using Stirling's approximation for $n\,!$, when n is large, *viz.*,

$$\lim_{n\to\infty} \Gamma(n+1) = \lim_{n\to\infty} n\,! \approx \sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}$$

$$\Rightarrow \quad \log\Gamma(n+1) = (n + \tfrac{1}{2})\log n - n + \log\sqrt{2\pi}\,, \text{ we get}$$

$$\kappa_1 = \mu_1' = \frac{1}{2}\left(\frac{1}{\nu_2} - \frac{1}{\nu_1}\right), \qquad \kappa_2 = \mu_2 = \frac{1}{2}\left(\frac{1}{\nu_2} + \frac{1}{\nu_1} + \frac{1}{\nu_1^2} + \frac{1}{\nu_2^2}\right)$$

$$\kappa_3 = \mu_3 = \frac{1}{2}\left[\left(\frac{1}{\nu_2^2} - \frac{1}{\nu_1^2}\right) + \left(\frac{1}{\nu_2^3} + \frac{1}{\nu_1^3}\right)\right]$$

$$\kappa_4 = \mu_4 - 3\mu_2^2 = \frac{1}{\nu_1^3} + \frac{1}{\nu_2^3} + 3\left(\frac{1}{\nu_1^4} + \frac{1}{\nu_2^4}\right),$$

whence β_1 and β_2 can be found.

Remark. z-distribution tends to normal distribution with mean $\frac{1}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)$ and variance $\frac{1}{2}\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right)$, as υ_1 and υ_2 become large.

16·10. FISHER'S Z-TRANSFORMATION

To test the significance of an observed sample correlation coefficient from an uncorrelated bivariate normal population, t-test is used. But in random sample of size n from a bivariate normal population in which $\rho \neq 0$, Prof. R.A. Fisher proved that the distribution of 'r' is by no means normal and in the neighbourhood of $\rho = \pm 1$, its probability curve is extremely skewed even for large n. If $\rho \neq 0$, Fisher suggested the following transformation :

$$Z = \frac{1}{2}\log_e \frac{1+r}{1-r} = \tanh^{-1} r \qquad \ldots(16{\cdot}25)$$

and proved that even for small samples, the distribution of Z is approximately

normal with mean : $$\xi = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} = \tanh^{-1}\rho \qquad \ldots(16{\cdot}25a)$$

and variance $1/(n-3)$ and for large values of n, say > 50, the approximation is fairly good.

Remark. The values of Z have been tabulated for different values of r and are given in Table III at the end of the chapter.

16·10·1. Applications of Z-Transformation. Z-transformation has the following applications in Statistics.

(1) To test if an observed value of 'r' differs significantly from a hypothetical value ρ of the population correlation coefficient.

H_0 : *There is no significant difference between r and ρ. In other words, the given sample has been drawn from a bivariate normal population with correlation coefficient ρ.* If we take

$$Z = \tfrac{1}{2}\log_e \{(1+r)/(1-r)\} \text{ and } \xi = \tfrac{1}{2}\log_e \{(1+\rho)/(1-\rho)\}, \text{ then under } H_0,$$

$$Z \sim N\left(\xi, \frac{1}{n-3}\right) \quad \Rightarrow \quad \frac{Z-\xi}{\sqrt{1/(n-3)}} \sim N(0, 1)$$

Thus if $(Z-\xi)\sqrt{(n-3)} > 1{\cdot}96$, H_0 is rejected at 5% level of significance and if it is greater than 2·58, H_0 is rejected at 1% level of significance, where Z and ξ are defined in (16·25) and (16·25a).

Remark. Z defined in equation (16·25) should not be confused with the Z used in Fisher's z-distribution (*c.f.* § 16·9).

Example 16·29. *A correlation coefficient of 0·72 is obtained from a sample of 29 pairs of observations.*

(i) Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0·8 ?

(ii) Obtain 95% confidence limits for ρ in the light of the information provided by the sample.

Solution. (*i*) H_0 : There is no significant difference between $r = 0{\cdot}72$; and $\rho = 0{\cdot}80$, *i.e.*, the sample can be regarded as drawn from the bivariate normal population with $\rho = 0{\cdot}8$. Here

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1{\cdot}1513\log_{10}\left(\frac{1+r}{1-r}\right) = 1{\cdot}1513\log_{10}6{\cdot}14 = 0{\cdot}907$$

$$\xi = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) = 1{\cdot}1513\log_{10}\frac{(1+0{\cdot}8)}{(1-0{\cdot}8)} = 1{\cdot}1513\times 0{\cdot}9541 = 1{\cdot}1$$

$$\text{S.E.}(Z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{26}} = 0{\cdot}196$$

Under H_0, the test statistic is : $U = \dfrac{Z-\xi}{1/\sqrt{n-3}} \sim N(0, 1)$

$$\therefore \qquad U = \frac{(0{\cdot}907 - 1{\cdot}100)}{0{\cdot}196} = -0{\cdot}985$$

Since $|U| < 1{\cdot}96$, it is not significant at 5% level of significance and H_0 may be accepted. Hence the sample may be regarded as coming from a bivariate normal population with $\rho = 0{\cdot}8$.

(*ii*) 95% confidence limits for ρ on the basis of the information supplied by the sample, are given by :

$$|U| \le 1{\cdot}96 \quad \Rightarrow \quad |Z-\xi| \le 1{\cdot}96 \times \frac{1}{\sqrt{n-3}} = 1{\cdot}96 \times 0{\cdot}196$$

$$\Rightarrow \quad |0{\cdot}907 - \xi| \le 0{\cdot}384 \quad \text{or} \quad 0{\cdot}907 - 0{\cdot}384 \le \xi \le 0{\cdot}907 + 0{\cdot}384$$

$$\Rightarrow \quad 0{\cdot}523 \le \xi \le 1{\cdot}291 \quad \text{or} \quad 0{\cdot}523 \le \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}291$$

$$\Rightarrow 0{\cdot}523 \le 1{\cdot}1513\log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}291 \quad \text{or} \quad \frac{0{\cdot}523}{1{\cdot}1513} \le \log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le \frac{1{\cdot}291}{1{\cdot}1513}$$

$$\therefore \qquad 0{\cdot}4543 \le \log_{10}\left(\frac{1+\rho}{1-\rho}\right) \le 1{\cdot}1213 \qquad \ldots(*)$$

Now $\log_{10}\left(\dfrac{1+\rho}{1-\rho}\right) = 0{\cdot}4543$

$$\Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}\,(0{\cdot}4543) = 2{\cdot}846$$

$$\therefore \quad \rho = \frac{2{\cdot}846-1}{2{\cdot}846+1} = \frac{1{\cdot}846}{3{\cdot}846} = 0{\cdot}4799$$

and $\log_{10}\left(\dfrac{1+\rho}{1-\rho}\right) = 1{\cdot}1213$

$$\Rightarrow \frac{1+\rho}{1-\rho} = \text{Antilog}\,(1{\cdot}1213) = 13{\cdot}22$$

$$\therefore \quad \rho = \frac{13{\cdot}22-1}{13{\cdot}22+1} = \frac{12{\cdot}22}{14{\cdot}22} = 0{\cdot}86$$

Hence, substituting in (*), we get $\qquad 0{\cdot}48 \le \rho \le 0{\cdot}86$

(2) To test the significance of the difference between two independent sample correlation coefficients. Let r_1 and r_2 be the sample correlation coefficients observed in two independent samples of sizes n_1 and n_2 respectively, then,

$$Z_1 = \log_e\left(\frac{1+r_1}{1-r_1}\right) \quad \text{and} \quad Z_2 = \frac{1}{2}\log_e\left(\frac{1+r_2}{1-r_2}\right)$$

Under the null hypothesis, H_0 : *that sample correlation coefficients do not differ significantly, i.e., the samples are drawn from the same bivariate normal population or from different populations with same correlation coefficient* ρ, *(say), the statistic :*

$$Z = \frac{(Z_1 - Z_2) - E(Z_1 - Z_2)}{\text{S.E.}\,(Z_1 - Z_2)} \sim N(0,1)$$

$$E(Z_1 - Z_2) = E(Z_1) - E(Z_2) = \xi_1 - \xi_2 = 0$$

$$\left[\because \xi_1 = \xi_2 = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \text{ (under } H_0)\right]$$

and $$\text{S.E.}(Z_1 - Z_2) = \sqrt{V(Z_1) + V(Z_2)} = \sqrt{\left\{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right\}}$$

[Covariance term vanishes since samples are independent.]

Under H_0, the test statistic is :

$$Z = \frac{Z_1 - Z_2}{\sqrt{\left\{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right\}}} \sim N(0, 1)$$

By comparing this value with 1·96 or 2·58, H_0 may be accepted or rejected at 5% and 1% levels of significance respectively.

(3) To obtain pooled estimate of ρ. Let $r_1, r_2, \ldots, r_k$ be observed correlation coefficients in k-independent samples of sizes $n_1, n_2, \ldots, n_k$ respectively from a bivariate normal population. The problem is to combine these estimates of ρ to get a pooled estimate for the parameter. If we take

$Z_i = \frac{1}{2}\log_e\left(\frac{1+r_i}{1-r_i}\right)$; $i = 1, 2, \ldots, k$; then Z_i ; $i = 1, 2, \ldots, k$ are independent normal variates with variances $\frac{1}{(n_i - 3)}$; $i = 1, 2, \ldots, k$ and common mean $\xi = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$.

The weighted mean, (say $\bar{Z}$), of these Z_i's is given by :

$$\bar{Z} = \sum_{i=1}^{k} w_i Z_i \Big/ \sum_{i=1}^{k} w_i, \text{ where } w_i \text{ is the weight of } Z_i.$$

Now $\bar{Z}$ is also an unbiased estimate of ξ, since

$$E(\bar{Z}) = \frac{1}{\Sigma w_i}\left(E \sum_{i=1}^{k} w_i Z_i\right) = \frac{1}{\Sigma w_i}\left[\sum_i w_i E(Z_i)\right] = \frac{1}{\Sigma w_i}\left(\sum_i w_i \xi\right) = \xi$$

and $$V(\bar{Z}) = \frac{1}{(\Sigma w_i)^2} V(\Sigma w_i Z_i) = \frac{1}{(\Sigma w_i)^2}\left[\Sigma w_i^2 V(Z_i)\right] \qquad \ldots (*)$$

The weights w_i's, ($i = 1, 2, \ldots, n$) are so chosen that $\bar{Z}$ has minimum variance.

In order that $V(\bar{Z})$ is minimum for variations in w_i, we should have

$$\frac{\partial}{\partial w_i} V(\bar{Z}) = 0 ; \quad i = 1, 2, \ldots, k.$$

$$\Rightarrow \quad \frac{(\Sigma w_i)^2 2 w_i V(Z_i) - \left[\Sigma w_i^2 V(Z_i)\right] 2(\Sigma w_i)}{(\Sigma w_i)^4} = 0 \quad \text{or} \quad w_i V(Z_i) = \frac{\Sigma w_i^2 V(Z_i)}{\Sigma w_i}, \text{ a constant.}$$

$$\therefore \quad w_i \propto \frac{1}{V(Z_i)} = (n_i - 3) ; i = 1, 2, \ldots, k. \qquad \ldots (**)$$

Hence the minimum variance estimate of ξ is given by :

TABLE I.

SIGNIFICANT VALUES $t_\nu(\alpha)$ of t-Distribution (TWO-TAIL AREAS)

$$P[\,|\,t\,|>t_\nu(\alpha)]=\alpha$$

d.f.	*Probability (Level of Significance)*					
(ν)	0·50	0·10	0·05	0·02	0·01	0·001
1	1·00	6·31	12·71	31·82	63·66	636·62
2	0·82	2·92	4·30	6·97	6·93	31·60
3	0·77	2·35	3·18	4·54	5·84	12·94
4	0·74	2·13	2·78	3·75	4·60	8·61
5	0·73	2·02	2·57	3·37	4·03	6·86
6	0·72	1·94	2·45	3·14	3·71	5·96
7	0·71	1·90	2·37	3·00	3·50	5·41
8	0·71	1·86	2·31	2·90	3·36	5·04
9	0·70	1·83	2·26	2·82	3·25	4·78
10	0·70	1·81	2·23	2·76	3·17	4·59
11	0·70	1·80	2·20	2·72	3·11	4·44
12	0·70	1·78	2·18	2·68	3·06	4·32
13	0·69	1·77	2·16	2·65	3·01	4·22
14	0·69	1·76	2·15	2·62	2·98	4·14
15	0·69	1·75	2·13	2·60	2·95	4·07
16	0·69	1·75	2·12	2·58	2·92	4·02
17	0·69	1·74	2·11	2·57	2·90	3·97
18	0·69	1·73	2·10	2·55	2·88	3·92
19	0.69	1.73	2.09	2.54	2.86	3·88
20	0·69	1·73	2·09	2·53	2·85	3·85
21	0·69	1·72	2·08	2·52	2·83	3·83
22	0·69	1·72	2·07	2·51	2·82	3·79
23	0·69	1·71	2·07	2·50	2·81	3·77
24	0·69	1·71	2·06	5·49	2·80	3·75
25	0·68	1·71	2·06	2·49	2·79	3·73
26	0·68	1·71	2·06	2·48	2·78	3·71
27	0·68	1·70	2·05	2·47	2·77	3·69
28	0·68	1·70	2·05	2·47	2·76	3·67
29	0·68	1·70	2·05	2·46	2·76	3·66
30	0·68	1·70	2·04	2·46	2·75	3·65
∞	0·67	1·65	1·96	2·33	2·58	3·29

TABLE II-A

SIGNIFICANT VALUES OF THE VARIANCE-RATIO F-DISTRIBUTION (RIGHT TAIL AREAS) 5 PER CENT POINTS

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	238.90	243.90	249.00	254.30
2	18.51	19.00	19.16	19.25	19.30	19.35	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.65
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.96
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.865	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.365	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	4.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.76
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.60
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.30
120	3.92	3.87	2.68	2.45	2.29	2.17	2.02	1.83	1.62	1.25
240	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

TABLE II-B

SIGNIFICANT VALUES OF THE VARIANCE RATIO F-DISTRIBUTION (RIGHT TAIL AREAS) — 1 PER CENT POINTS

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999.5	5403	5625	5764	5859	5982	6106	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.95	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

TABLE III—TRANSFORMATION FROM r to $Z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$

r	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.0000	0.0100	0.0200	0.0300	0.0400	0.0500	0.0601	0.0701	0.0802	0.0902
·1	0.1003	0.1104	0.1206	0.1307	0.1409	0.1511	0.1614	0.7117	0.1820	0.1923
·2	0.2027	0.2132	0.2237	0.2342	0.2448	0.2554	0.2661	0.2769	0.2877	0.2986
·3	0.3005	0.3205	0.3316	0.3428	0.3541	0.3654	0.3769	0.3884	0.4001	0.4118
·4	0.4236	0.4356	0.4477	0.4599	0.4722	0.4847	0.4973	0.5101	0.5230	0.5361
·5	0.5493	0.5627	0.5763	0.5901	0.6042	0.6184	0.6328	0.6475	0.6625	0.6777
·6	0.6931	0.7089	0.7250	0.7414	0.7582	0.7753	0.7928	0.8107	0.8291	0.8480
·7	0.8673	0.8872	0.9076	0.9287	0.9505	0.9730	0.9962	1.0203	1.0454	1.0714
·8	1.0996	1.1270	1.1568	1.1881	1.2212	1.2562	1.2933	1.3331	1.3758	1.4219
·9	1.4722	1.5275	1.5890	1.6584	1.7380	1.8318	1.9459	2.0923	2.2976	2.6467