CHAPTER 27

Sampling & Inference

27.1 (1) INTRODUCTION

We know that a small section selected from the population is called a *sample* and the process of drawing a sample is called *sampling*. It is essential that a sample must be a *random* selection so that each member of the population has the same chance of being included in the sample. Thus the fundamental assumption underlying theory of sampling is *Random sampling*.

A special case of random sampling in which each event has the same probability p of success and the chance of success of different events are independent whether previous trials have been made or not, is known as *simple sampling*.

The statistical constants of the population such as mean (μ), standard deviation (σ) etc. are called the *parameters*. Similarly, constants for the *sample* drawn from the given population *i.e.*, mean ($\bar{x}$), standard deviation (S) etc. are called the *statistic*. The population parameters are in general, not known and their estimates given by the corresponding sample statistic are used. We use the Greek letters to denote the population parameters and Roman letters for sample statistic.

(2) Objectives of sampling. Sampling aims at gathering the maximum information about the population with the minimum effort, cost and time. The object of sampling studies is to obtain the best possible values of the parameters under specific conditions. Sampling determines the reliability of these estimates. The logic of the sampling theory is the logic of induction in which we pass from a particular (sample) to general (population). Such a generalisation from sample to population is called **Statistical Inference.**

27.2 SAMPLING DISTRIBUTION

Consider all possible samples of size n which can be drawn from a given population at random. For each sample, we can compute the mean. The means of the samples will not be identical. If we group these different means according to their frequencies, the frequency distribution so formed is known as *sampling distribution of the mean*. Similarly we can have *sampling distribution of the standard deviation* etc.

While drawing each sample, we put back the previous sample so that the parent population remains the same. This is called *sampling with replacement* and all the subsequent formulae will pertain to sampling with replacement.

(2) Standard error. The standard deviation of the sampling distribution is called the *standard error* (*S.E.*). Thus the standard error of the sampling distribution of means is called standard error of means. The standard error is used to assess the difference between the expected and observed values. The reciprocal of the standard error is called *precision*.

If $n \geq 30$, a sample is called *large* otherwise *small*. The sampling distribution of large samples is assumed to be normal.

27.3 (1) TESTING A HYPOTHESIS*

To reach decisions about populations on the basis of sample information, we make certain assumptions about the populations involved. Such assumptions, which may or may not be true, are called *statistical hypothesis*. By testing a hypothesis is meant a process for deciding whether to accept or reject the hypothesis. The method consists in assuming the hypothesis as correct and then computing the probability of getting the observed sample. If this probability is less than a certain preassigned value the hypothesis is rejected.

(2) Errors. If a hypothesis is rejected while it should have been accepted, we say that a *Type I error* has been committed. On the other hand, if a hypothesis is accepted while it should have been rejected, we say that a *Type II error* has been made. The statistical testing of hypothesis aims at limiting the Type I error to a preassigned value (say : 5% or 1%) and to minimize the Type II error. The only way to reduce both types of errors is to increase the sample size, if possible.

(3) Null hypothesis. The hypothesis formulated for the sake of rejecting it, under the assumption that it is true, is called the *null hypothesis* and is denoted by H_0. To test whether one procedure is better than another, we assume that there is no difference between the procedures. Similarly to test whether there is a relationship between two variates, we take H_0 that there is no relationship. By accepting a null hypothesis, we mean that on the basis of the statistic calculated from the sample, we do not reject the hypothesis. It however, does not imply that the hypothesis is proved to be true. Nor its rejection implies that it is disproved.

27.4 (1) LEVEL OF SIGNIFICANCE

The probability level below which we reject the hypothesis is known as the *level of significance*. The region in which a sample value falling is rejected, is known as the *critical region*. We generally take two critical regions which cover 5% and 1% areas of the normal curve. The shaded portion in the figure corresponds to 5% level of significance. Thus the *probability of the value of the variate falling in the critical region is the level of significance*.

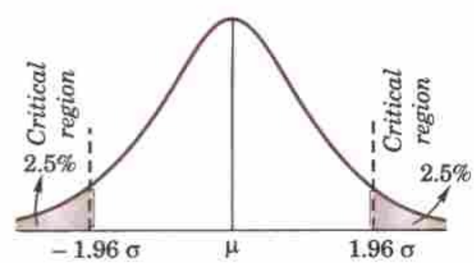

Fig. 27.1

Depending on the nature of the problem, we use a *single-tail test* or *double-tail test* to estimate the significance of a result. In a double-tail test, the areas of both the tails of the curve representing the sampling distribution are taken into account whereas in the single tail test, only the area on the right of an ordinate is taken into consideration. For instance, to test whether a coin is biased or not, double-tail test should be used, since a biased coin gives either more number of heads than tails (which corresponds to right tail), or more number of tails than heads (which corresponds to left tail only).

(2) Tests of significance. The procedure which enables us to decide whether to accept or reject the hypothesis is called the *test of significance*. Here we test whether the differences between the sample values and the population values (or the values given by two samples) are so large that they signify evidence against the hypothesis or these differences are so small as to account for fluctuations of sampling.

27.5 CONFIDENCE LIMITS**

Suppose that the sampling distribution of a statistic S is normal with mean μ and standard deviation σ. As in the Fig. 27.1 the sample statistic S can be expected to lie in the interval $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ for 95% times *i.e.*, we can be confident of finding μ in the interval $(S - 1.96\sigma, S + 1.96\sigma)$ in 95% cases. Because of this, we call

*The American statistician *J. Neyman* (1894—1981) and the English statistician *E.S. Pearson* (1895—1980)–son of *Karl Pearson* (See footnote p. 843), developed a systematic theory of tests around 1930.

***J. Neyman* developed the modern theory and terminology of confidence limits.

$(S - 1.96\sigma, S + 1.96\sigma)$ the 95% confidence interval for estimation of μ. The ends of this interval (*i.e.* $S \pm 1.96\sigma$) are called 95% *confidence limits* (or *fiducial limits*) for S. Similarly $S \pm 2.58\sigma$ are 99% confidence limits. The numbers 1.96, 2.58 etc. are called *confidence coefficients*. The values of confidence coefficients corresponding to various levels of significance can be found from the normal curve area table VI – Appendix 2.

27.6 SIMPLE SAMPLING OF ATTRIBUTES

The sampling of attributes may be regarded as the selection of samples from a population whose members possess the attribute K or not K. The presence of K may be called a success and its absence a failure.

Suppose we draw a simple sample of n items. Clearly it is same as a series of n independent trials with the same probability p of success. The probabilities of 0, 1, 2, ..., n successes are the terms in the binomial expansion of $(q + p)^n$ where $q = 1 - p$.

We know that the mean of this distribution is np and standard deviation is $\sqrt{(npq)}$ *i.e.*, the expected value of success in a sample of size n is np and the standard error is $\sqrt{(npq)}$.

If we consider the proportion of successes, then

(*i*) mean proportion of successes $= np/n = p$.

(*ii*) standard error of the proportion of successes

$$= \sqrt{\left(n \cdot \frac{p}{n} \cdot \frac{q}{n}\right)} = \sqrt{\left(\frac{pq}{n}\right)}$$

and (*iii*) precision of the proportion of successes $= \sqrt{(n/pq)}$, which varies as $\sqrt{n}$, since p and q are constants.

27.7 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

We know that the binomial distribution tends to normal for large n. Suppose we wish to test the hypothesis that the probability of success in such trial is p. Assuming it to be true, the mean μ and the standard deviation σ of the sampling distribution of number of successes are np and $\sqrt{(npq)}$ respectively.

For a normal distribution, only 5% of the members lie outside $\mu \pm 1.96\sigma$ while only 1% of the members lie outside $\mu \pm 2.58\sigma$.

If x be the observed number of successes in the sample and z is the standard normal variate then $z = (x - \mu)/\sigma$.

Thus we have the following test of significance :

(*i*) *If* $|z| < 1.96$, *difference between the observed and expected number of successes is not significant.*

(*ii*) *If* $|z| > 1.96$, *difference is significant at 5% level of significance.*

(*iii*) *If* $|z| > 2.58$, *difference is significant at 1% level of significance.*

Example 27.1. *A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased at 5% level of significance.* (V.T.U., 2007)

Solution. Suppose the coin is unbiased.

Then the probability of getting the head in a toss $= \dfrac{1}{2}$

$\therefore$ expected number of successes $= \dfrac{1}{2} \times 400 = 200$

and the observed value of successes = 216

Thus the excess of observed value over expected value = 216 – 200 = 16

Also S.D. of simple sampling $= \sqrt{npq} = \sqrt{\left(400 \times \frac{1}{2} \times \frac{1}{2}\right)} = 10$

Hence $z = \dfrac{x - np}{\sqrt{(npq)}} = \dfrac{16}{10} = 1.6$

As $z < 1.96$, the hypothesis is accepted at 5% level of significance *i.e.*, we conclude that the coin is unbiased at 5% level of significance.

Example 27.2. *A die was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate an unbiased die?* (V.T.U., 2010)

Solution. Suppose the die is unbiased.

Then the probability of throwing 5 or 6 with one die $= \frac{1}{3}$

The expected number of successes $= \frac{1}{3} \times 9000 = 3000$

and the observed value of successes = 3240

Thus the excess of observed value over expected value 3240 – 3000 = 240

Also S.D. of simple sampling $= \sqrt{npq} = \sqrt{\left(9000 \times \frac{1}{3} \times \frac{2}{3}\right)} = 44.72$

Hence $z = \frac{x - np}{\sqrt{(npq)}} = \frac{240}{44.72} = 5.4$ nearly.

As $z > 2.58$, the hypothesis has to be rejected at 1% level of significance and we conclude that the die is biased.

Example 27.3. *In a locality containing 18000 families, a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have a monthly income of ₹ 250 or less. It is desired to estimate how many out of 18,000 families have a monthly income of ₹ 250 or less. Within what limits would you place your estimate?*

Solution. Here $p = \frac{206}{840} = \frac{103}{420}$ and $q = \frac{317}{420}$

$\therefore$ standard error of the population of families having a monthly income of ₹ 250 or less

$$= \sqrt{\left(\frac{pq}{n}\right)} = \sqrt{\left(\frac{103}{420} \times \frac{317}{420} \times \frac{1}{840}\right)} = .015 = 1.5\%$$

Hence taking $\frac{103}{420}$ (or 24.5%) to be the estimate of families having a monthly income of ₹ 250 or less in the locality, the limits are $(24.5 \pm 3 \times 1.5)\%$ *i.e.*, 20% and 29% approximately.

27.8 COMPARISON OF LARGE SAMPLES

Two large samples of sizes n_1, n_2 are taken from two populations giving proportions of attributes A's as p_1, p_2 respectively.

(*a*) On the hypothesis that the populations are similar as regards the attribute A, we combine the two samples to find an estimate of the common value of proportion of A's in the populations which is given by

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

If e_1, e_2 be the standard errors in the two samples then

$$e_1^2 = \frac{pq}{n_1} \text{ and } e_2^2 = \frac{pq}{n_2}$$

If e be the standard error of the difference between p_1 and p_2, then

$$e^2 = e_1^2 + e_2^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad \therefore \quad z = \frac{p_1 \sim p_2}{e}$$

If $z > 3$, the difference between p_1 and p_2 is real one.

If $z < 2$, the difference may be due to fluctuations of simple sampling.

But if z lies between 2 and 3, then the difference is significant at 5% level of significance.

(*b*) If the proportions of A's are not the same in the two populations from which the samples are drawn, but p_1 and p_2 are the true values of proportions then S.E. e of the difference $p_1 \sim p_2$ is given by

$$e^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

If $z = \frac{p_1 \sim p_2}{e} < 3$, the difference could have arisen due to fluctuations of simple sampling.

Example 27.4. *In a city A 20% of a random sample of 900 school boys had a certain slight physical defect. In another city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant ?* (V.T.U., 2003 S)

Solution. We have $n_1 = 900, n_2 = 1600$

and $$p_1 = \frac{20}{100} = \frac{1}{5}, p_2 = \frac{18.5}{100}$$

$\therefore$ $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{180 + 296}{900 + 1600} = 0.19$$

and $$q = 1 - 0.19 = 0.81$$

Thus $$e^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.19 \times 0.81\left(\frac{1}{900} + \frac{1}{1600}\right) = 0.0017$$

giving $$e = 0.04 \text{ nearly.}$$

Also $$p_1 - p_2 = \frac{1.5}{100} = 0.015 \quad \therefore \quad z = \frac{p_1 - p_2}{e} = \frac{.015}{.04} = 0.37$$

As $z < 1$, the difference between the proportions is not significant.

Example 27.5. *In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations ?* (Coimbatore, 2001)

Solution. Here $p_1 = 0.3, p_2 = 0.25$ so that $p_1 - p_2 = 0.05$.

$\therefore$ $$e^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}$$

so that $$e = 0.0195$$

$\therefore$ $$z = \frac{p_1 - p_2}{e} = \frac{0.05}{0.0195} = 2.5 \text{ nearly}$$

Hence it is unlikely that the real difference will be hidden.

PROBLEMS 27.1

1. A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased ? (V.T.U., 2006)
2. 12 dice are thrown 3086 times and a throw of 2, 3, 4 is reckoned as a success. Suppose that 19142 throws of 2, 3, 4 have been made out. Do you think that this observed value deviates from the expected value ? If so, can the deviation from the expected value be due to fluctuations of simple sampling ?
3. Balls are drawn from a bag containing equal number of black and white balls, each ball being replaced before drawing another. In 2250 drawings 1018 black and 1232 white balls have been drawn. Do you suspect some bias on the part of the drawer ?
4. A sample of 1000 days is taken from meteorological records of a certain district and 120 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district ?
5. In a group of 50 first cousines there were found to be 27 males and 23 females. Ascertain if the observed proportions are inconsistent with the hypothesis that the sexes should be in equal proportion.
6. A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of the bad apples in the consignment as well as the standard error of the estimate. Deduce that the percentage of bad apples in the consignment almost certainly lies between 8.5 and 17.5.
7. 400 children are chosen in an industrial town and 150 are found to be under weight. Assuming the conditions of simple sampling, estimate the percentage of children who are under weight in the industrial town and assign limits within which the percentage probably lies ?

8. A machine produces 16 imperfect articles in a sample of 500. After machine is overhauled, it produces 3 imperfect articles in a batch of 100. Has the machine been improved? (*Rohtak, 2005 ; Madras, 2003*)
9. One type of aircraft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a significant difference in the two types of aircrafts so far as engine defects are concerned?
10. In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men? (*J.N.T.U., 2003*)
11. In a sample of 500 people from a state 280 take tea and rest take coffee. Can we assume that tea and coffee are equally popular in the state at 5% level of significance?

27.9 (1) SAMPLING OF VARIABLES

We now consider sampling of a variable such as weight, height, etc. Each member of the population gives a value of the variable and the population is a frequency distribution of the variable. Thus a random sample of size n from the population is same as selecting n values of the variables from those of the distribution.

(2) Sampling distribution of the mean. *If a population is distributed normally with mean μ and standard deviation σ, then the means of all positive random samples of size n, are also distributed normally with mean μ and standard error $\sigma/\sqrt{n}$.* This result shows how the precision of a sample mean increases as the sample size increases.

27.10 CENTRAL LIMIT THEOREM

This is a very important theorem regarding the distribution of the mean of a sample *if the parent population is non-normal and the sample size is large.*

If the variable X has a non-normal distribution with mean μ and standard deviation σ, then the limiting distribution of

$z = \dfrac{\bar{x}-\mu}{\sigma\sqrt{n}}$ as $n \to \infty$, *is the standard normal distribution* (*i.e.*, with mean 0 and unit S.D.)

There is no restriction upon the distribution of X except that it has a finite mean and variance. This theorem holds well for a sample of 25 or more which is regarded as large.

Thus *if the population is normal, the sampling distribution of the mean is also normal with mean μ and S.E. $\sigma/\sqrt{n}$, while for large samples the same result holds even if the distribution of the population is non-normal.* This property is of universal use and throws light on the importance of normal distribution in statistical theory.

27.11 CONFIDENCE LIMITS FOR UNKNOWN MEAN

Let the population from which a random sample of size n is drawn, have mean μ and S.D. σ. If μ is not known, there will be a range of values of μ for which observed mean $\bar{x}$ of the sample is not significant at any assigned level of probability. The relative deviation of $\bar{x}$ from μ is $(\bar{x}-\mu)/\sqrt{\sigma}$.

If $\bar{x}$ is not significant at 5% level of probability, then

$$|(\bar{x}-\mu)\sqrt{n}/\sigma| < 1.96 \quad i.e. \quad \bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}$$

Thus 95% confidence or fiducial limits for the mean of the population corresponding to given sample are $\bar{x} \pm 1.96\sigma/\sqrt{n}$.

Similarly *99% confidence limits for μ are $\bar{x} \pm 2.58\sigma/\sqrt{n}$.*

Example 27.6. *A sample of 900 members is found to have a mean of 3.4 cm. Can it be reasonably regarded as a truly random sample from a large population with mean 3.25 cm and S.D. 1.61 cm.*

Solution. Here $\bar{x}$ = 3.4 cm, n = 900, μ = 3.25 and σ = 1.61 cm

$$\therefore \qquad z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{3.4-3.25}{1.61/30} = 2.8$$

As $z > 1.96$, the deviation of the sample mean from the mean of the population is significant at 5% level of significance. Hence it cannot be regarded as a random sample.

Example 27.7. *The mean of a certain normal population is equal to the standard error of the mean of the samples of 100 from that distribution. Find the probability that the mean of the sample of 25 from the distribution will be negative ?*

Solution. If μ be the mean and σ the S.D. of the distribution, then

$$\mu = \text{S.E. of the sample means} = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10}$$

Also for a sample of size 25, we have $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{25}} = \dfrac{\bar{x} - \sigma/10}{6/5} = \dfrac{5\bar{x}}{\sigma} - \dfrac{1}{2}$

Since $\bar{x}$ is negative, $z < -\frac{1}{2}$.

$\therefore$ the probability that a normal variate $z < -\frac{1}{2}$

$$= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}} e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}} \int_{1/2}^{\infty} e^{-\frac{1}{2}z^2} dz$$

$= 0.5 - 0.915 = 0.3085$, from the tables.

Example 27.8. *An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with 90% confidence.*

Solution. S.E. of the proportion of heads $= \sqrt{\dfrac{pq}{n}} = \sqrt{\dfrac{1}{2} \cdot \dfrac{1}{2} \cdot \dfrac{1}{n}} = \dfrac{1}{2\sqrt{n}}$

90% of confidence = 45% or .45 of the total area under the normal curve on each side of the mean.

$\therefore$ the corresponding value of $z = 1.645$, from the tables.

Thus $\quad p \mp 1.645\sigma = 0.49$ or 0.51.

i.e., $\quad 0.5 - 1.645 \cdot \dfrac{1}{2\sqrt{n}} = 0.49$ and $0.5 + 1.645 \cdot \dfrac{1}{2\sqrt{n}} = 0.51$

whence $\quad \dfrac{1.645}{2\sqrt{n}} = 0.01$ or $\sqrt{n} = \dfrac{329}{4}$ or $n = 6765$ approximately.

27.12 TEST OF SIGNIFICANCE FOR MEANS OF TWO LARGE SAMPLES

(*a*) Suppose two random samples of sizes n_1 and n_2 have been *drawn from the same population* with S.D. σ. We wish to test whether the difference between the sample means $\bar{x}_1$ and $\bar{x}_2$ is significant or is merely due to fluctuations of sampling.

If the samples are independent, then the standard error e of the difference of their means is given by

$$e^2 = e_1^2 + e_2^2$$

where $e_1 = \sigma/\sqrt{n_1}$, $e_2 = \sigma/\sqrt{n_2}$ are the S.E.s of the means of the two samples.

$$\therefore \quad e = \sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \quad \text{Hence} \quad z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{(1/n_1 + 1/n_2)}}$$

is normally distributed with mean zero and S.D. 1.

Test of significance (n_1, n_2 being large) :

If $z > 1.96$, then the difference is significant at 5% level of significance.

If $z > 3$, it is highly probable that either the samples have not been drawn from the same population or the sampling is not simple.

(*b*) *If the samples are known to be drawn from different populations* with means μ_1, μ_2 and standard deviations σ_1 and σ_2. Then the standard error *e* of their means is given by

$$e = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

Assuming that the two populations have the same mean (*i.e.*, $\mu_1 = \mu_2$), the difference of the means of the samples will be normally distributed with mean zero and S.D. *e*. Now the same procedure of test of significance is applied.

Example 27.9. *The means of simple samples of sizes 1000 and 2000 are 67.5 and 68.0 cm respectively. Can the samples be regarded as drawn from the same population of S.D. 2.5 cm.* (*Madras, 2002*)

Solution. We have $\bar{x}_1 = 67.5$, $\bar{x}_2 = 68.0$

$n_1 = 1000$, $n_2 = 2000$.

On the hypothesis, that the samples are drawn from the same population of S.D. $\sigma = 2.5$, we get

$$z = \frac{\bar{x}_1 \sim \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{67.5 \sim 68.0}{2.5\sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}}$$

$$= \frac{0.5}{2.5 \times 0.0387} = \frac{0.5}{0.09675} = 5.1$$

Hence the difference between the sample means *i.e.*, 5.1 is very much greater than 1.96 and is therefore significant. Thus, the samples cannot be regarded as drawn from the same population.

Example 27.10. *A sample of height of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches while a simple sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers?*

Solution. Here $\bar{x}_1 = 67.85$, $\sigma_1 = 2.56$, $n_1 = 6400$

$\bar{x}_2 = 68.55$, $\sigma_2 = 2.52$, $n_2 = 1600$.

$\therefore$ S.E. of the difference of the mean heights is

$$e = \sqrt{\left[\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right]} = \sqrt{\left[\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}\right]}$$

$$= \sqrt{[.001024 + .003969]} = 0.005 \text{ nearly.}$$

Also difference between the means $= \bar{x}_2 - \bar{x}_1 = 0.7$, which $> 10e$. This is highly significant. Hence the data indicates that the sailors are on the average taller than the soldiers.

PROBLEMS 27.2

1. A sample of 400 items is taken from a normal population whose mean is 4 and variance 4. If the sample mean is 4.45, can the samples be regarded as a simple sample ?
2. To know the mean weights of all 10-year old boys in Delhi, a sample of 225 is taken. The mean weight of the sample is found to be 67 pounds with a S.D. of 12 pounds. Can you draw any inference from it about the mean weight of the population ?
3. A normal population has a mean 0.1 and a S.D. of 2.1. Find the probability that the mean of simple sample of 900 members will be negative.
4. If the mean breaking strength of copper wire is 575 lbs. with a standard deviation of 8.3 lbs., how large a sample must be used in order that there be one chance in 100 that the mean breaking strength of the sample is less than 572 lbs. ?

   [**Hint.** $|z| = \left|\frac{\bar{x} - \mu}{\sigma}\sqrt{n}\right| = \frac{3}{8.3}\sqrt{n}$

   Also from table IV, $z = 2.33$. Hence $n = 42$ nearly.]

5. A research worker wishes to estimate mean of a population by using sufficiently large sample. The probability is 95% that sample mean will not differ from the true mean by more than 25% of the S.D. How large a sample should be taken ?
6. The density function of a random variable x is $f(x) = ke^{-2x^2+10x}$. Find the upper 5% point of the distribution of the means of the random sample of size 25 from the above population.
7. The means of two large samples of 1000 and 2000 members are 168.75 cms. and 170 cms. respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms.
8. If 60 new entrants in a given university are found to have a mean height of 68.60 inches and 50 seniors a mean height of 69.51 inches ; is the evidence conclusive that the mean height of the seniors is greater than that of the new entrants ? Assume the standard deviation of height to be 2.48 inches.
9. A sample of 100 electric bulbs produced by manufacturer A showed a mean life time of 1190 hours and a standard deviation of 90 hours. A sample of 75 bulbs produced by manufacturer B showed a mean life time of 1230 hours, with a standard deviation of 120 hours. Is there a difference between the mean life time of two brands at a significance level of (*i*) 0.05 (*ii*) 0.01.
10. A random sample of 1000 men from North India shows that their mean wage is ₹ 5 per day with a S.D. of ₹ 1.50. A sample of 1500 men from South India gives a mean wage of ₹ 4.50 per day with a standard deviation of ₹ 2. Does the mean rate of wages varies as between the two regions ?

27.13 SAMPLING OF VARIABLES—SMALL SAMPLES

In case of large samples, sampling distribution approaches a normal distribution and values of sample statistic are considered best estimates of the parameters in a population. It will no longer be possible to assume that statistics computed from small samples are normally distributed. As such, *a new technique has been devised for small samples which involves the concept of 'degrees of freedom' which we explain below.*

Number of degrees of freedom *is the number of values in a set which may be assigned arbitrarily.* For instance, if $x_1 + x_2 + x_3 = 15$ and we assign any values of two of the variables (say : x_1, x_2), then the values of x_3 will be known. The two variables are therefore, free and independent choices for finding the third. Hence these are the degrees of freedom. *If there are n observations, the degrees of freedom (d.f.) are (n – 1).* In other words, while finding the mean of a small sample, one degree of freedom is used up and $(n-1)$ *d.f.* are left to estimate the population variance.

27.14 (1) STUDENT'S t-DISTRIBUTION

Consider a small sample of size n, drawn from a normal population with mean μ and s.d. σ. If $\bar{x}$ and σ_s be the sample mean and s.d., then the statistic, 't' is defined as

$$t = \frac{\bar{x}-\mu}{\sigma}\sqrt{n} \quad \text{or} \quad t = \frac{\bar{x}-\mu}{\sigma_s}\sqrt{(n-1)},$$

where $\nu = n - 1$ denotes the *df.* of t. If we calculate t for each sample, we obtain the sampling distribution for t. This distribution known as *Student's t-distribution**, is given by

$$y = \frac{y_0}{(1+t^2/\nu)^{(\nu+1)/2}} \qquad \text{...(1)}$$

where y_0 is constant such that the area under the curve is unity.

(2) Properties of t-distribution.

I. This curve is symmetrical about the line $t = 0$, like the normal curve, since only even powers of t appear in (1). But it is more peaked than the normal curve with the same S.D. The t-curve approaches the horizontal axis less rapidly than the normal curve. Also t-curve attains its maximum value at $t = 0$ so that its mode coincides with the mean. (Fig. 27.2)

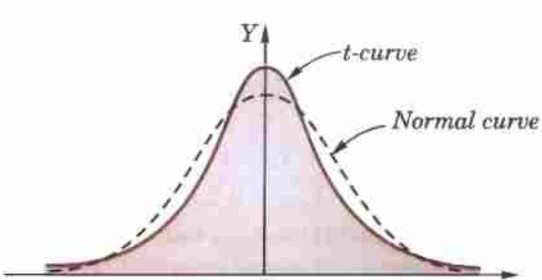

Fig. 27.2

*This distribution was first found by the English statistician *W.S. Gosset* in 1908 who wrote under the pen-name of '*Student*'. *R.A. Fisher* defined t correctly and found its distribution in 1926.

II. The limiting form of t-distribution when $\nu \to \infty$ is given by $y = y_0 e^{-\frac{1}{2}t^2}$ which is a normal curve. This shows that *t is normally distributed for large samples.*

III. The probability P that the value of t will exceed t_0 is given by

$$P = \int_{t_0}^{\infty} y \, dx$$

The values of t_0 have been tabulated for various values of P for various values of ν from 1 to 30 (Table IV – Appendix 2).

IV. Moments about the mean

All the moments of odd order about the mean are zero, due to its symmetry about the line $t = 0$.

Even order moments about the mean are

$$\mu_2 = \frac{\nu}{\nu - 2}, \quad \mu_4 = \frac{3\nu^2}{(\nu - 2)(\nu - 4)}, \ldots$$

The t-distribution is often used in tests of hypothesis about the mean when the population standard deviation σ is unknown.

27.15 SIGNIFICANCE TEST OF A SAMPLE MEAN

Given a random small sample $x_1, x_2, x_3 \ldots x_n$ from a normal population, we have to test the hypothesis that mean of the population is μ. For this, we first calculate $t = (\bar{x} - \mu)\sqrt{n}/\sigma_s$

where $$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i, \quad \sigma_s^{\,2} = \frac{1}{n-1}\sum_{1}^{n} (x_i - \bar{x})^2$$

Then find the value of P for the given *df* from the table.

If the calculated value of $t > t_{0.05}$, *the difference between* $\bar{x}$ *and* μ *is said to be significant at 5% level of significance.*

If $t > t_{0.01}$, the difference is said to be significant at 1% level of significance.

If $t < t_{0.05}$, *the data is said to be consistent with the hypothesis that* μ *is the mean of the population.*

Example 27.11. *A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressure : 5, 2, 8, – 1, 3, 0, – 2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.* (V.T.U., 2007)

Solution. Let us assume that the stimulus administered to all the 12 patients will increase the B.P. Taking the population to be normal with mean $\mu = 0$ and S.D. σ,

$$\bar{d} = \frac{5 + 2 + 8 - 1 + 3 + 0 - 2 + 1 + 5 + 0 + 4 + 6}{12} = 2.583$$

$$\sigma^2 = \frac{\Sigma d^2}{n} - \bar{d}^2 = \frac{1}{12}\,[5^2 + 2^2 + 8^2 + (-1^2) + 3^2 + 0^2 + (-2)^2 + 1^2 + 5^2 + 0^2 + 4^2 + 6^2] - (2.583)^2$$

$$= 8.744. \quad \therefore \quad \sigma = 2.9571$$

Now $$t = \frac{\bar{d} - \mu}{\sigma_s}\sqrt{(n-1)} = \frac{2.583 - 0}{2.9571}\sqrt{(12-1)} = 2.897$$

Here *d.f.* $\gamma = 12 - 1 = 11$.

For $\gamma = 11$, $t_{0.05} = 2.2$, from table IV.

Since the $|t| > t_{0.05}$, our assumption is rejected *i.e.*, the stimulus does not increase the B.P.

Example 27.12. *The nine items of a sample have the following values : 45, 47, 50, 52, 48, 47, 49, 53, 51. Does the mean of these differ significantly from the assumed mean of 47.5 ?* (V.T.U., 2010)

Solution. We find the mean and standard deviation of the sample as follows :

x	$d = x - 48$	d^2
45	– 3	9
47	– 1	1
50	2	4
52	2	4
48	0	0
47	– 1	1
49	1	1
53	5	25
51	3	9
Total	10	66

$$\therefore \qquad \bar{x} = \text{mean} = 48 + \frac{\Sigma d}{9} = 48 + \frac{10}{9} = 49.1$$

$$\sigma_s^{\,2} = \frac{\Sigma d^2}{9} - \left(\frac{\Sigma d}{9}\right)^2 = \frac{66}{9} - \frac{100}{81} = \frac{494}{81}$$

$$\therefore \qquad \sigma_s = 2.47$$

Hence $$t = \frac{\bar{x} - \mu}{\sigma_s}\sqrt{(n-1)} = \frac{49.1 - 47.5}{2047}\sqrt{8} = 1.83$$

Here $d.f.\ \nu = 9 - 1 = 8$

For $\nu = 8$, we get from table IV, $t_{0.05} = 2.31$.

As calculated value of $t < t_{0.05}$, the value of t is not significant at 5% level of significance which implies that there is no significant difference between $\bar{x}$ and μ. Thus the test provides no evidence against the population mean being 47.5.

Example 27.13. *A machinist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts shows mean diameter 0.742 inch with a standard deviation of 0.04 inch. On the basis of this sample, would you say that the work is inferior ?* (V.T.U., 2009)

Solution. Here we have $\mu = 0.700$, $\bar{x} = 0.742$, $\sigma_x = 0.040$, $n = 10$.

Taking the hypothesis that the product is not inferior *i.e.*, there is no significant difference between $\bar{x}$ and μ.

$$\therefore \qquad t = \frac{\bar{x} - \mu}{\sigma_x}\sqrt{(n-1)} = \frac{0.742 - 0.700}{0.040}\sqrt{(10-1)} = \frac{0.126}{0.040} = 3.16$$

Degrees of freedom $\rho = 10 - 1 = 9$.

For $\rho = 9$, we get from table IV, $t_{0.05} = 2.262$.

As the calculated value of $t > t_{0.05}$, the value of t is significant at 5% level of significance. This implies that $\bar{x}$ differs significantly from μ and the hypothesis is rejected. Hence the work is inferior. In fact, the work is inferior even at 2% level of significance.

Example 27.14. *Show that 95% confidence limits for the mean μ of the population are $\bar{x} \pm \frac{\sigma_s}{\sqrt{n}} t_{0.05}$. Deduce that for a random sample of 16 values with mean 41.5 inches and the sum of the squares of the deviations from the mean 135 inches^2 and drawn from a normal population, 95% confidence limits for the mean of the population are 39.9 and 43.1 inches.*

Solution. (*a*) The probability P that $t \le t_{0.05}$ is 0.95. Hence the 95% confidence limits for μ are given by

$$\left| \frac{\bar{x} - \mu}{\sigma_s}\sqrt{n} \right| \le t_{0.05}$$

or $$|\bar{x} - \mu| \le \frac{\sigma_s}{\sqrt{n}} t_{0.05} \quad \text{or} \quad \bar{x} - \frac{\sigma_s}{\sqrt{n}} t_{0.05} \le \mu \le \bar{x} + \frac{\sigma_s}{\sqrt{n}} t_{0.05}$$

We can, therefore, say with a confidence coefficient 0.95 that the confidence interval $\bar{x} \pm \frac{\sigma_s}{\sqrt{n}} t_{0.05}$ contains the population mean μ.

(*b*) Here, $n = 16, \nu = n - 1 = 15, \sigma_s = \sqrt{\frac{135}{15}} = 3.$

Also from table IV, $t_{0.05}$ (for $\nu = 15$) = 2.13

$$\therefore \qquad \frac{\sigma_s}{\sqrt{n}} t_{0.05} = \frac{3}{4} \times 2.13 = 1.6 \text{ approx.}$$

Hence the required confidence limits are 41.5 ± 1.6 *i.e.*, 39.9 and 43.1 inches.

27.16 SIGNIFICANCE TEST OF DIFFERENCE BETWEEN SAMPLE MEANS

Given two independent samples $x_1, x_2, x_3, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ with means $\bar{x}$ and $\bar{y}$ and standard deviations σ_x and σ_y from a normal population with the same variance, we have to test the hypothesis that the population means μ_1 and μ_2 are the same.

For this, we calculate
$$t = \frac{\bar{x} \sim \bar{y}}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots(1)$$

where
$$\bar{x} = \frac{1}{n_1}\sum_1^{n_1} x_i, \; \bar{y} = \frac{1}{n_2}\sum_1^{n_2} y_i$$

and
$$\sigma_s^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\sigma_x^2 + (n_2 - 1)\sigma_y^2\right] = \frac{1}{n_1 + n_2 - 2}\left\{\sum_1^{n_1} (x_i - \bar{x})^2 + \sum_1^{n_2} (y_j - \bar{y})^2\right\}$$

It can be shown that the variate t defined by (1) follows the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

If the calculated value of $t > t_{0.05}$, the difference between the sample means is said to be significant at 5% level of significance.

If $t > t_{0.01}$, the difference is said to be significant at 1% level of significance.

If $t < t_{0.05}$, the data is said to be consistent with the hypothesis, that $\mu_1 = \mu_2$.

Cor. If the two samples are of the same size and the data are paired, then t is defined by

$$t = \frac{\bar{d} - 0}{(\sigma/\sqrt{n})} \quad \text{where} \quad \sigma^2 = \frac{1}{n-1}\sum_1^n (d_i - \bar{d})^2$$

d_i = difference of the ith members of the samples ;

$\bar{d}$ = mean of the differences = $\Sigma d_i/n$; and the number of *d.f.* = $n - 1$.

Example 27.15. *Eleven students were given a test in statistics. They were given a month's further tuition and a second test of equal difficulty was held at the end of it. Do the marks give evidence that the students have benefitted by extra coaching ?*

Boys	:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Marks I test	:	*23*	*20*	*19*	*21*	*18*	*20*	*18*	*17*	*23*	*16*	*19*
Marks II test	:	*24*	*19*	*22*	*18*	*20*	*22*	*20*	*20*	*23*	*20*	*17*

(V.T.U., 2011 S)

Solution. We compute the mean and the S.D. of the difference between the marks of the two tests as under :

$$\bar{d} = \text{mean of } d\text{'s} = \frac{11}{11} = 1 \; ; \; \sigma_s^{-2} = \frac{\Sigma(d - \bar{d})^2}{n - 1} = \frac{50}{10} = 5 \quad i.e., \quad \sigma_s = 2.24$$

Assuming that the students have not been benefited by extra coaching, it implies that the mean of the difference between the marks of the two tests is zero *i.e.*, $\mu = 0$.

Then $$t = \frac{\bar{d} - \mu}{\sigma_s}\sqrt{n} = \frac{1-0}{2.24}\sqrt{11} = 1.48 \text{ nearly and } df\ \nu = 11 - 1 = 10.$$

Students	x_1	x_2	$d = x_2 - x_1$	$d - \bar{d}$	$(d - \bar{d})^2$
1	23	24	1	0	0
2	20	19	−1	−2	4
3	19	22	3	2	4
4	21	18	−3	−4	16
5	18	20	2	1	1
6	20	22	2	1	1
7	18	20	2	1	1
8	17	20	3	2	4
9	23	23	—	−1	1
10	16	20	4	3	9
11	19	17	−2	−3	9
			$\Sigma d = 11$		$\Sigma(d - \bar{d})^2 = 50$

From table IV, we find that $t_{0.05}$ (for $\nu = 10$) = 2.228. As the calculated value of $t < t_{0.05}$, the value of t is not significant at 5% level of significance *i.e.*, the test provides no evidence that the students have benefited by extra coaching.

Example 27.16. *From a random sample of 10 pigs fed on diet A, the increases in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs. For another random sample of 12 pigs fed on diet B, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs. Test whether diets A and B differ significantly as regards their effect on increases in weight ?*

Solution. We calculate the means and standard deviations of the samples as follows :

	Diet A			*Diet B*	
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
10	−2	4	7	−8	64
6	−6	36	13	−2	4
16	4	16	22	7	49
17	5	25	15	0	0
13	1	1	12	−3	9
12	0	0	14	−1	1
8	−4	16	18	3	9
14	2	4	8	−7	49
15	3	9	21	6	36
9	−3	9	23	8	64
			10	−5	25
			17	2	4
120	0	120	180	0	314

$$\bar{x} = \frac{120}{10} = 12 \text{ lbs.}, \bar{y} = \frac{180}{12} = 15 \text{ lbs.}$$

$$\sigma_s^2 = [\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2]/(n_1 + n_2 - 2)$$

$$= (120 + 314)/(10 + 12 - 2) = (434/20) = 21.1$$

$\therefore$ $$\sigma_s = 4.65$$

Assuming that the samples do not differ in weight so far as the two diets are concerned *i.e.*, $\mu_1 - \mu_2 = 0$.

Hence $$t = \frac{(\bar{y} - \bar{x}) - 0}{\sigma_s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{15 - 12}{4.65\sqrt{\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{3}{4.65}\sqrt{\frac{120}{22}} = 1.6 \text{ nearly}$$

Here $\qquad d.f.\ \nu = n_1 + n_2 - 2 = 10 + 12 - 2 = 20.$

For $\nu = 20$, we find $t_{0.05} = 2.09$ [From table IV

$\therefore$ the calculated value of $t < t_{0.05}$.

Hence the difference between the sample means is not significant *i.e.*, the two diets do not differ significantly as regards their effect on increase in weight.

PROBLEMS 27.3

1. Find the student's t for the following variable values in a sample of eight : $-4, -2, -2, 0, 2, 2, 3, 3$; taking the mean of the universe to be zero.
2. A random sample of 10 boys had the following I.Q. :

   70, 120, 110, 101, 88, 83, 95, 98, 107, 100.

   Do these data support the assumption of a population mean I.Q. of 100 (at 5% level of significance) ? (*V.T.U., 2006 ; Coimbatore, 2001*)
3. A sample of 10 measurements of the diameter of a sphere gave a mean of 12 cm and a standard deviation 0.15 cm. Find 95% confidence limits for the actual diameter.
4. A random sample of size 25 from a normal population has the mean $\bar{x} = 47.5$ and s.d. $S = 8.4$. Does this information refute the claim that the mean of the population is $\mu = 42.1$. (*J.N.T.U., 2003*)
5. A process for making certain bearings is under control if the diameter of the bearings have the mean 0.5 cm. What can we say about this process if a sample of 10 of these bearings has a mean diameter of 0.506 cm. and s.d. of 0.004 cm ?
6. A machine is supposed to produce washers of mean thickness 0.12 cm. A sample of 10 washers was found to have a mean thickness of 0.128 cm and standard deviation 0.008. Test whether the machine is working in proper order at 5% level of significance.
7. Find out the reliability of the sample mean of the following data : *Breaking strength of 10 specimens of 1.04 cms diameter hard-drawn copper wire :*

Specimen	:	1	2	3	4	5	6	7	8	9	10
Breaking Strength (kgs)	:	578	572	570	568	572	570	570	572	526	584

8. Test runs with 6 models of an experimental engine showed that they operated for 24, 28, 21, 23, 32 and 22 minutes with a gallon of fuel. If the probability of a Type I error is at the most 0.01, is this evidence against a hypothesis that on the average this kind of engine will operate for atleast 29 minutes per gallon of the same fuel. Assume normality. (*J.N.T.U., 2003*)
9. Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results :

Horse A	:	28	30	32	33	33	29	and 34
Horse B	:	29	30	30	24	27	and 29	

   Test whether you can discriminate between two horses ? (*Rohtak, 2005 ; Coimbatore, 2001*)
10. A group of 10 rats fed on a diet A and another group of 8 rats fed on a different diet B, recorded the following increase in weights :

Diet A	:	5	6	8	1	12	4	3	9	6	10	gm
Diet B	:	2	3	6	8	10	1	2	8	gm.		

   Does it show that superiority of diet A over that of B ? (*Madras, 2003*)
11. A group of boys and girls were given an intelligence test. The mean score, S.D.s and numbers in each group are as follows :

	Boys	*Girls*
Mean	124	121
S.D.	12	10
n	18	14

   Is the mean score of boys significantly different from that of girls ?
12. The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population ? (*Mumbai, 2004*)

27.17 (1) CHI-SQUARE (χ^2) TEST

When a fair coin is tossed 80 times, we expect from theoretical considerations that heads will appear 40 times and tail 40 times. But this never happens in practice *i.e.*, the results obtained in an experiment do not agree exactly with the theoretical results. *The magnitude of discrepancy between observation and theory is given by the quantity* χ^2 (pronounced as *chi-square*). If $\chi^2 = 0$, the observed and theoretical frequencies completely agree. As the value of χ^2 increases, the discrepancy between the observed and theoretical frequencies increases.

(1) Definition. *If $O_1, O_2, ..., O_n$ be a set of observed (experimental) frequencies and $E_1, E_2, ..., E_n$ be the corresponding set of expected (theoretical) frequencies, then χ^2 is defined by the relation*

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + ... + \frac{(O_n - E_n)^2}{E_n}$$

$$= \sum \frac{(O_i - E_i)^2}{E_i} \qquad ...(1)$$

with n – 1 degrees of freedom. [$\Sigma O_i = \Sigma E_i = n$ the total frequency]

(2) Chi-square distribution*

If $x_1, x_2, ..., x_n$ be n independent normal variates with mean zero and s.d. unity, then it can be shown that $x_1^2 + x_2^2 + ... + x_n^2$, is a random variate having χ^2-distribution with *ndf*.

The equation of the χ^2-curve is

$$y = y_0 e^{-\chi^2/2} (\chi^2)^{(\nu - 1)/2} \qquad ...(2)$$

where $\nu = n - 1$ (Fig. 27.3).

(3) Properties of χ^2-distribution

I. If $\nu = 1$, the χ^2-curve (2) reduces to $y = y_0 e^{-\chi^2/2}$, which is the exponential distribution.

II. If $\nu > 1$, this curve is tangential to x-axis at the origin and is positively skewed as the mean is at ν and mode at $\nu - 2$.

III. The probability P that the value of χ^2 from a random sample will exceed χ_0^2 is given by

$$P = \int_{\chi_0^2}^{\infty} y\, dx.$$

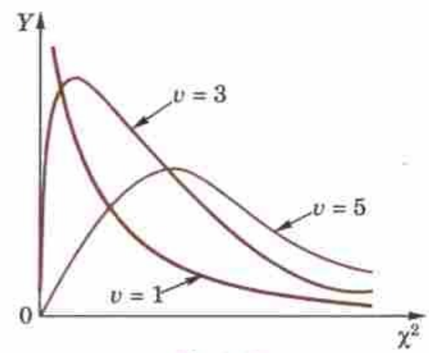

Fig. 27.3

The values of χ_0^2 have been tabulated for various values of P and for values of ν from 1 to 30. (Table-V-Appendix 2)

For $\nu > 30$, the χ^2-curve approximates to the normal curve and we should refer to normal distribution tables for significant values of χ^2.

IV. Since the equation of χ^2-curve does not involve any parameters of the population, *this distribution does not depend on the form of the population* and is therefore, very useful in a large number of problems.

V. Mean = γ and *variance* = 2γ.

27.18 GOODNESS OF FIT

The value of χ^2 is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It is also used to test how will a set of observations fit a given distribution, χ^2 therefore, provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution. As a test of goodness of fit, it can be used to study the correspondence between theory and fact.

This is a non-parametric distribution-free test since in this we make no assumption about the distribution of the parent population.

**Hamlet* discovered this distribution in 1875. *Karl Pearson* rediscovered it independently in 1900 and applied it to test 'goodness of fit'.

Procedure to test significance and goodness of fit.

(*i*) Set up a 'null hypothesis' and calculate

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

(*ii*) Find the *df* and read the corresponding values of χ^2 at a prescribed significance level from Table V.

(*iii*) From χ^2-table, we can also find the probability P corresponding to the calculated values of χ^2 for the given *d.f.*

(*iv*) If $P < 0.05$, the observed value of χ^2 is significant at 5% level of significance.

If $P < 0.01$, the value is significant at 1% level.

If $P > 0.05$, it is a good fit and the value is not significant.

Example 27.17. *In experiments on pea breeding, the following frequencies of seeds were obtained :*

Round and yellow	*Wrinkled and yellow*	*Round and green*	*Wrinkled and green*	*Total*
315	*101*	*108*	*32*	*556*

Theory predicts that the frequencies should be in proportions 9 : 3 : 3 : 1. Examine the correspondence between theory and experiment.

Solution. The corresponding frequencies are

$$\frac{9}{16}\times 556, \frac{3}{16}\times 556, \frac{3}{16}\times 556, \frac{1}{16}\times 556 \text{ i.e., } 313, 104, 104, 35.$$

Hence $$\chi^2 = \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35}$$

$$= \frac{4}{313} + \frac{9}{104} + \frac{16}{104} + \frac{9}{35} = 0.51 \quad \text{and} \quad df\ \nu = 4 - 1 = 3.$$

For $\nu = 3$, we have $\chi^2_{0.05} = 7.815$ [From Table V

Since the calculated value of χ^2 is much less than $\chi^2_{0.05}$, there is a very high degree of agreement between theory and experiment.

Example 27.18. *A set of five similar coins is tossed 320 times and the result is*

No. of heads :	*0*	*1*	*2*	*3*	*4*	*5*
Frequency :	*6*	*27*	*72*	*112*	*71*	*32*

Test the hypothesis that the data follow a binomial distribution.

(*Kottayam, 2005 ; P.T.U., 2005 ; V.T.U., 2004*)

Solution. For $\nu = 5$, we have $\chi^2_{0.05} = 11.07$.

p, probability of getting a head $= \frac{1}{2}$; q, probability of getting a tail $= \frac{1}{2}$.

Hence the theoretical frequencies of getting 0, 1, 2, 3, 4, 5 heads are the successive terms of the binomial expansion $320\,(p+q)^5$

$$= 320\,[p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5]$$

$$= 320\left[\frac{1}{32} + \frac{5}{32} + \frac{10}{32} + \frac{10}{32} + \frac{5}{32} + \frac{1}{32}\right] = 10 + 50 + 100 + 100 + 50 + 10$$

Thus the theoretical frequencies are 10, 50, 100, 100, 50, 10.

Hence

$$\chi^2 = \frac{(6-10)^2}{10} + \frac{(27-50)^2}{50} + \frac{(72-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(71-50)^2}{50} + \frac{(32-10)^2}{10}$$

$$= \frac{1}{100}(160 + 1058 + 784 + 144 + 882 + 4840) = \frac{7868}{100} = 78.68$$

and $df\ \nu = 6 - 1 = 5.$

Since the calculated value of χ^2 is much greater than $\chi^2_{0.05}$, the hypothesis that the data follow the binomial law is rejected.

Example 27.19. *Fit a Poisson distribution to the following data and test for its goodness of fit at level of significance 0.05.*

x :	0	1	2	3	4	
f :	419	352	154	56	19	(V.T.U., 2008)

Solution. Mean $m = \frac{\Sigma f x_i}{\Sigma f_i} = \frac{904}{1000} = 0.904$. $\therefore \quad e^{-m} = e^{-0.904} = 0.4049$.

Hence the theoretical frequencies are $\frac{1000 \times e^{-m}(m)^r}{r!}$, $r = 0, 1, 2, 3, 4$

i.e.,						
	x :	0	1	2	3	4	Total
	f :	404.9	366	165.4	49.8	11.3	997.4
		(406.2)				(12.6)	

In order that the total observed and expected frequencies may agree, we take the first and last theoretical frequencies as 406.2 and 12.6 instead of 404.9 and 11.3 as shown in brackets. (In case, the expected frequencies are less than 10, we group together such classes. Here of course, none of the frequencies < 10).

Hence

$$\chi^2 = \frac{(419-406.2)^2}{406.2} + \frac{(352-366)^2}{366} + \frac{(154-165.4)^2}{165.4} + \frac{(56-49.8)^2}{49.8} + \frac{(19-12.6)^2}{12.6}$$
$$= 0.403 + 0.536 + 0.786 + 0.772 + 3.251 = 5.748$$

Since the mean of the theoretical distribution has been estimated from the given data and the totals have been made to agree, there are two constraints so that the number of degrees of freedom $\nu = 5 - 2 = 3$.

For $\nu = 3$, we have $\chi^2_{0.05} = 7.82$. [From Table V]

Since the calculated value of $\chi^2 < \chi^2_{0.05}$, the agreement between the fact and theory is good and hence the Poisson distribution can be fitted to the data.

Example 27.20. *Fit a normal distribution to the following data of weights of 100 students of Delhi University and test the goodness of fit.*

Weights (kg) :	60-62	63-65	66-68	69-71	72-74
Frequency :	5	18	42	27	8

Solution. Here $N = 100$, mean $m = 67.45$ and S.D. $\sigma = 2.92$. The calculations are arranged as follows* :

Class boundary (x)	$z = (x-m)/\sigma$	*Area under normal curve from 0 to z*	*Area for each class (A)**	*Expected frequency ($f_e = N \times A$)*
59.5	– 2.72	0.4967		
			0.0413	4.13
62.5	– 1.70	0.4554		
			0.2068	20.68
65.5	– 0.67	0.2486		
			0.3892	38.92
68.5	0.36	0.1406		
			0.2771	27.71
71.5	1.39	0.4177		
			0.0743	7.43
74.5	2.41	0.4920		

* *A* is obtained by subtracting the successive areas in the third column when the corresponding values of *z* have the same sign and adding them when the *z* values have opposite signs (which occurs only once).

$$\therefore \quad \chi^2 = \frac{(5-4.13)^2}{4.13} + \frac{(18-20.68)^2}{20.68} + \frac{(42-38.92)^2}{38.92} + \frac{(27-27.71)^2}{27.71} + \frac{(8-7.43)^2}{7.43}$$
$$= 0.1833 + 0.3473 + 0.2437 + 0.0182 + 0.0437 = 0.8362$$

As regards the number of degrees of freedom (γ), there are three constraints (i) discrepancy between total observed and total estimated frequencies (ii) and (iii) mean (m) and standard deviation (σ) have been estimated from the sample data. $\therefore \quad r = 5 - 3 = 2.$

For $\gamma = 2$, $\chi^2_{0.05} = 0.103$ from table V.

Since $\chi^2 = 0.8362 > 0.103$. Hence the fit is not good.

PROBLEMS 27.4

1. Five dice were thrown 96 times and the number of times 4, 5 or 6 were thrown were :

No. of dice showing 4, 5 or 6 :	5	4	3	2	1	0
Frequency :	8	18	35	24	10	1

Find the probability of getting this result by chance ?

2. Genetic theory states that children having one parent of blood type M and other blood type N will always be one of the three types M, MN, N and that the proportions of these types will on average be 1 : 2 : 1. A report states that out of 300 children having one M parent and one N parent, 30% were found to be of type M, 45% of type MN and remainder of type N. Test the hypothesis by χ^2 test.

3. A die was thrown 60 times and the following frequency distribution was observed :

Faces :	1	2	3	4	5	6
f_0 :	15	6	4	7	11	17

Test whether the die is unbiased ?

4. The following table gives the number of aircraft accidents that occurred during the various days of the week. Find whether the accidents are uniformly distributed over the week ?

Days :	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents :	14	16	8	12	11	9	14	84

(Hissar, 2005)

5. Fit a binomial distribution to the data :

x :	0	1	2	3	4	5
f :	38	144	342	287	164	25

and test for goodness of fit, at the level of significance 0.05.

6. In 1000 extensive sets of trials for an event of small probability, the frequencies f_0 of the number x of successes proved to be :

x :	0	1	2	3	4	5	6	7
f_0 :	305	366	210	80	28	9	2	1

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

7. The frequencies of localities according to the number of deaths due to cholera during eight years in 1000 localities is as follows :

No. of deaths :	0	1	2	3	4	5	6	7
No. of localities :	314	355	204	86	29	9	3	0

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

8. Obtain the equation of the normal curve that may be fitted to the data and test the goodness of fit.

x :	4	6	8	10	12	14	16	18	20	22	24	Total
$f(x)$:	1	7	15	22	35	43	38	20	13	5	1	200

27.19 (1) F-DISTRIBUTION*

Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be the values of two independent random samples drawn from the normal populations σ^2 having equal variances.

* This distribution was introduced by the English statistician *Prof. R.A. Fisher* (1890–1962) who had greatly influenced the development of modern statistics.

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means and $s_1^2 = \frac{1}{n_1 - 1}\sum_1^{n_1}(x_i - \bar{x})^2$, $s_2^2 = \frac{1}{n_2 - 1}\sum_1^{n_2}(y_i - \bar{y})^2$ be the sample variances.

Then we define F by the relation

$$F = \frac{s_1^2}{s_2^2} \qquad (s_1^2 > s_2^2)$$

This gives F-distribution (also known as variance ratio distribution) with $\gamma_1 = n_1 - 1$ and $\gamma_2 = n - 2$ degrees of freedom. *The larger of the variances is placed in the numerator.*

(2) Properties. I. The F-distribution curve lies entirely in the first quadrant and is *unimodal.*

II. The F-distribution is independent of the population variance σ^2 and depends on γ_1 and γ_2 only.

III. $F_\alpha(\gamma_1, \gamma_2)$ is the value of F for γ_1 and γ_2 of such that the area to the right of F_α is α.

IV. It can be shown that the mode of F-distribution is less than unity.

(3) Significance test. Snedecor's F-tables give 5% and 1% points of significance for F. (Table VI – Appendix 2). 5% points of F mean that area under the F-curve to the right of the ordinate at a value of F, is 0.05. Clearly value of F at 5% significance is lower than that at 1%. F-distribution is very useful for testing the equality of population means by comparing sample variances. As such it forms the basis of *analysis of variance*.

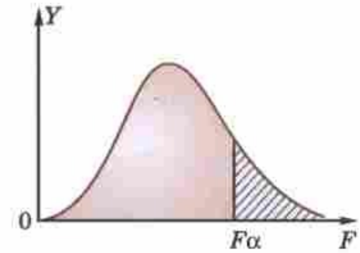

Fig. 27.4

Example 27.21. *Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 inches² and 91 inches² respectively. Can these be regarded as drawn from the same normal population ?* (V.T.U., 2002)

Solution. We have $\Sigma(x - \bar{x})^2 = 160$ and $\Sigma(y - \bar{y})^2 = 91$

$$\therefore \qquad s_1^2 = \frac{160}{8} = 20$$

and
$$s_2^2 = \frac{91}{7} = 13.$$

Hence
$$F = \frac{s_1^2}{s_2^2} = \frac{20}{13} = 1.54 \text{ nearly.}$$

For $\gamma_1 = 8$, $\gamma_2 = 7$, we have $F_{0.05} = 3.73$. [From Table VI]

Since the calculated value of $F < F_{0.05}$, the population variances are not significantly different. Thus the two samples can be regarded as drawn from two normal populations with the same variance. If the two populations are to be same, their means should also be the same which can be verified by applying t-test provided the sample means are known.

Example 27.22. *Measurements on the length of a copper wire were taken in 2 experiments A and B as under :*

A's measurements (mm) :	*12.29, 12.25, 11.86, 12.13, 12.44, 12.78, 12.77, 11.90, 12.47.*
B's measurements (mm) :	*12.39, 12.46, 12.34, 12.22, 11.98, 12.46, 12.23, 12.06.*

Test whether B's measurements are more accurate than A's. (The readings taken in both cases being unbiased)

Solution. Readings in both cases being unbiased, B's measurements will be taken more accurate if its population variance is less than that of A's measurements.

Under the hypothesis that the two populations have the same variance (*i.e.* $\sigma_1^2 = \sigma_2^2$), we have

$$F = \frac{s_1^2}{s_2^2}$$

with $\gamma_1 = n_1 - 1 = 8$ and $\gamma_2 = n_2 - 1 = 7$.

We calculate the s.d.'s of the two series as follows :

A's measurements			*B's measurements*		
x	$u = 100(x - 12)$	u^2	y	$v = 100(y - 12)$	v^2
12.29	29	841	12.39	39	1521
12.25	25	625	12.46	46	2116
11.86	– 14	196	12.34	34	1156
12.13	13	169	12.22	22	484
12.44	44	1936	11.98	– 2	4
12.78	78	6084	12.46	46	2116
12.77	77	5929	12.23	23	529
11.90	– 10	100	12.06	6	36
12.47	47	2209			
	289	18089		214	7962

$$\therefore \quad s_1^2 = \frac{1}{n_1 - 1}\left[18089 - \frac{(289)^2}{n_1}\right] = \frac{1}{8}(18089 - 9280) = 1101.1$$

$$s_2^2 = \frac{1}{n_2 - 1}\left[7962 - \frac{(214)^2}{n_2}\right] = \frac{1}{7}(7962 - 5724) = 319.7$$

$$\therefore \quad F = \frac{s_1^2}{s_2^2} = \frac{1101.1}{319.7} = 3.44$$

For $\gamma_1 = 8$ and $\gamma_2 = 7$, from table VI, $F_{0.05} = 3.73$ and $F_{0.01} = 6.84$.

Since the calculated value of F is less than both $F_{0.05}$ and $F_{0.01}$, the result is insignificant at both 5% and 1% level.

Hence there is no reason to say that B's measurements are more accurate than those of A's.

27.20 (1) FISHER'S z-DISTRIBUTION

Changing the variable F to z by the substitution $z = \frac{1}{2}\log_e F$ or $F = e^{2z}$ in the F-distribution, we get the *Fisher's z-distribution.*

It is more nearly symmetrical than F-distribution. Table showing the values of z that will be exceeded in simple sampling with probabilities 0.05 and 0.01 have been prepared for various values of v_1 and v_2.

(2) Significance test. As z-table give only critical values corresponding to right hand tail areas, therefore 5% (or 1%) points of z imply that the area to the right of the ordinate at z is 0.05 (or 0.01). In other words, *5% and 1% points of z correspond to 10% and 2% levels of significance respectively.*

Example 27.23. *Two gauge operations are tested for precision in making measurements. One operator completes a set of 26 readings with a standard deviations of 1.34 and the other does 34 readings with a standard deviations of 0.98. What is the level of significance of this difference.*

(Given that for $\nu_1 = 25$ *and* $\nu_2 = 33$, $z_{0.05} = 0.305$, $z_{0.01} = 0.432$*)*

Solution. We have $n_1 = 26$, $\sigma_x = 1.34$; $n_2 = 34$, $\sigma_y = 0.98$

$$\therefore \quad s_1^2 = \frac{n_1}{n_1 - 1}\,.\,\sigma_x^2 = \frac{26}{25}(1.34)^2 \approx (1.34)^2 \quad \text{and} \quad s_2^2 = \frac{n_2}{n_2 - 1}\,.\,\sigma_y^2 = \frac{34}{33}(0.98)^2 \approx (0.98)^2$$

$$\text{Hence} \quad F = \left(\frac{1.34}{0.98}\right)^2 = 1.8696 \quad \text{and} \quad z = \frac{1}{2}\log_e F = 1.1513 \log_{10} 1.8696 = 0.3129$$

Since the calculated value of z is just greater than $z_{0.05}$ and less than $z_{0.01}$, the difference between the standard deviation is just significant at 5% level and insignificant at 1% level.

PROBLEMS 27.5

1. Two samples of 9 and 7 individuals have variances 4.8 and 9.6 respectively. Is the variance 9.6 significantly greater than the variance 4.6 ?
2. Test for breaking strength were carried out on two lots of 5 and 9 steel wires. The variance of one lot was 230 and that of other was 492. Is there a significant difference in their variability ?
3. Show how you would use Fisher's z-test to decide whether the two sets of observations 17, 27, 18, 25, 27, 29, 27, 23, 17 and 16, 16, 20, 16, 20, 17, 15, 21, indicate samples from the same universe.
4. In two groups of ten children each, the increase in weight due to different diets during the same period, were in pounds

   3, 7, 5, 6, 5, 4, 4, 5, 3, 6

   8, 5, 7, 8, 3, 2, 7, 6, 5, 7.

   Is there a significant difference in their variability ?
5. The mean diameter of rivets produced by two firms A and B are practically the same but their standard deviations are different. For 16 rivets manufactured by firm A, the S.D. is 3.8 mm while for 22 rivets manufactured by firm B is 2.9 mm. Do you think products of firm B are of better quality than those of firm A.
6. The I.Q.'s of 25 students from one college showed a variance of 16 and those of an equal number from the other college had a variance of 8. Discuss whether there is any significant difference in variability of intelligence.
7. Two random samples from two normal populations are given below :

Sample I	:	16	26	27	23	24	22
Sample II	:	33	42	35	32	28	31

   Do the estimates of population variances differ significantly ?

Degrees of freedom	:	(5, 5)	(5, 6)	(6, 5)
5% value of F	:	5.05	4.39	4.95.

8. Two independent samples of sizes 7 and 6 have the following values :

Sample A :	28	30	32	33	33	29	34
Sample B :	29	30	30	24	27	29	

   Examine whether the samples have been drawn from normal populations having the same variance ?

   [Given that the values of F at 5% level for (6, 5) d.f. is 4.95 and for (5, 6) d.f. is 4.39].

27.21 OBJECTIVE TYPE OF QUESTIONS

PROBLEMS 27.6

Select the correct answer or fill up the blanks in each of the following questions :

1. The 'null hypothesis' implies that ....
2. The uses of t-distribution are ....
3. Type I and type II errors are such that ....
4. A single-tailed test is used when ....
5. Control limit theorem states that ....
6. A hypothesis is true, but is rejected. Then this is an error of type ....
7. If the standard deviation of a χ^2 distribution is 10, then its degree of freedom is ....
8. Range of F-distribution is ....
9. A hypothesis is false but accepted, then there is an error of type ....
10. The mean and variance of a χ^2 distribution with 8 degrees of freedom are .... and .... respectively.
11. In a t-distribution of sample size n, the degrees of freedom are ....
12. The test statistic $F = \dfrac{s_1^2}{s_2^2}$ is used when

    (*i*) $s_2^2 > s_1^2$ (*ii*) $s_2^2 < s_1^2$ (*iii*) $s_1^2 = s_2^2$ (*iv*) none of these.
13. The t-test is applicable to samples for which n is ....
14. The two main uses of χ^2-test are ....
15. Range of t-distribution is ....
16. If two samples are taken from two populations of unequal variances, we can apply t-test to test the difference of means. (True or False)
17. The Chi-square distribution is continuous. (True or False)