4 Special Discrete Probability Distributions

CHAPTER

4.1 INTRODUCTION

So far we have studied probability distribution both in case of discrete and continuous rand
variables, mathematical expectation, moments and other related characteristics of a probabili
distribution like, skewness, kurtosis etc. in general. In this chapter, we consider some speci
discrete probability distributions which occur frequently in applications. We shall also derive t
constants of the distributions studied and also will outline the specific situations, through vario
examples or otherwise, in which these distributions can be applied. The distributions studied
Section 5.2 through 5.8 are *uniform, binomial, multinomial, hypergeometric, negative binomial, geometr*
and *Poisson's distribution*. In the end, a set of review exercises and a problem set based on t
distributions studied has been given.

4.2 DISCRETE UNIFORM DISTRIBUTION

It is the simplest of all discrete probability distributions where the random variable assume
different value with equal probabilities and is defined as follows.

A random variable x is said to have a discrete uniform (or rectangular) distribution over the range [1,
if its p.d.f. can be given as

$$P[x = x_i] = \frac{1}{n}, \quad i = 1, 2, \ldots, n. \qquad \ldots(4.1$$

Here n, a positive integer, is called the *parameter* of the distribution. This distribution is suitab
in case of the random experiments when all the outcomes are equally likely, for example, throw o
an unbiased dice or random draw of a card from a well-shuffled pack.

The graphic representation of a uniform distribution by means of a histogram always turns ou
to be a set of rectangles with equal heights. That is why, sometimes the distribution is named a
rectangular distribution also.

For example, the histogram for the uniform distribution

$$f(x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6.$$

is shown in Fig. 4.1.

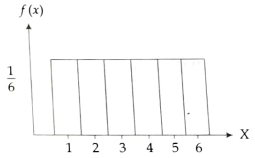

Fig. 4.1

4.2.1 Constants of the Uniform Distribution

Mean, $$E(X) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} \qquad \ldots(4.2)$$

Also, $$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{(n+1)(2n+1)}{6}$$

Hence, variance $$\sigma^2 = E(X^2) - [E(X)]^2$$

$$= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} \qquad \ldots(4.3)$$

The m.g.f. about origin,

$$M_0(t) = E(e^{tX})$$

$$= \frac{1}{n}\sum_{x=1}^{n} e^{tx} = \frac{e^t(1-e^{nt})}{n(1-e^t)} \qquad \ldots(4.4)$$

4.3 BINOMIAL DISTRIBUTION

Suppose that an experiment or a trial consists of two possible outcomes classified as *success* or *failure*. Let we define $X = 1$ when the outcome is a success and $X = 0$ when it is a failure. Then the probability mass function of X is given by

$$P\{X = 1\} = p, \quad P\{X = 0\} = 1 - p, \qquad \ldots(4.5)$$

where p, $0 \le p \le 1$, is the probability that the trial is a success.

A random variable X defined as in (4.5) is called a *Bernoulli random variable*. Obviously,

$$E(X) = 1\,p + 0\,(1-p) = p$$

Next we define *Bernoulli trials*.

Repeated independent trials in which there are only two possible outcomes say, success or failure and the probability of success remains constant throughout the trials, are called Bernoulli trials.

For example, repeated tosses of a coin, and say falling head is classified as *success* and falling tail as *failure*. Repeated draws of a card from a pack with replacement and classifying *success* as the event getting a card of heart on a draw, otherwise *failure*.

Next, we derive *binomial distribution*.

Consider a set of n independent Bernoulli trials in which the probability of success is p and of failure is $q = 1 - p$. We wish to find the probability of x successes in n such trials.

First consider the probability of x successes and $(n - x)$ failures in a specified order sequence of n Bernoulli trails. Since the trails are independent with p as probability of success, q as that of failure, thus it is obviously $p^x q^{n-x}$. Since these x successes in n trials can occur in C_x^n and all these are mutually exclusive. Hence, the requisit probability is $C_x^n p^x q^{n-x}$.

The probability distribution of the number of successes so obtained is called the Bin *probability distribution*, and the random variable X giving the number of successes is called *binomial variate*. Thus,

A r.v. X taking non-negative integral values 0, 1, 2, ... with probability mass function

$$p(x) = P\{X = x\} = C_x^n p^x q^{n-x}, \quad q = 1 - p; \quad x = 0, 1, 2, \ldots n$$

is called binomial variate and *the distribution is called* the *Binomial distribution.*

The expression (4.6) defines a probability distribution, since

$$\sum_{x=0}^{n} P\{X = x\} = \sum_{x=0}^{n} c_x^n p^x q^{n-x} = (p + q)^n = 1.$$

The two independent constants n, p are called the *parameters* of the distribution. The probab function (4.6) is sometimes denoted by $b(x; n, p)$. Thus,

$$b(x; n, p) = C_x^n p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

and, since the $(n + 1)$ terms are the successive terms in the binomial expansion of $(q + p)^n$; hence name is so.

The graphs of binomial probability distribution for $n = 10$, and $p = .1$, .9 and .5 are shown Fig. 4.2(a), (b), and (c) respectively.

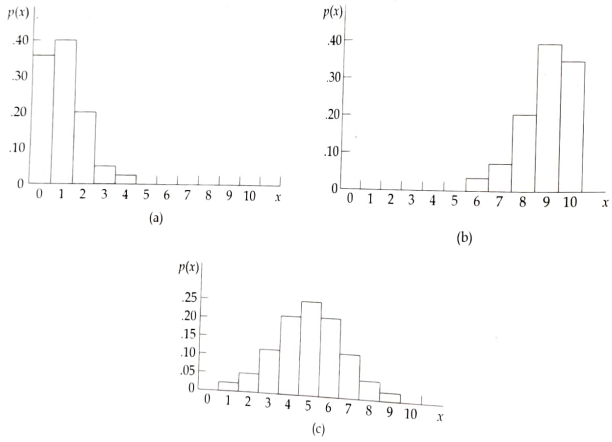

Fig. 4.2

Frequency function of the binomial distribution

Let us suppose that n trials constitute an experiment and let this experiment is repeated N times, then

$$f(x) = Np(x) = NC_x^n p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n \qquad \ldots(4.7)$$

are the expected frequencies of 0, 1, 2, ... n successes out of N. These are the successive terms in the expansion of $N(p + q)^n$.

From the examples to be studied next, we will find that the binomial distribution finds applications in many fields. In industry a quality control inspector is always interested in the *proportion defective* in an industrial process. The distribution is applicable where the process is dichotomous and the results of the process are independent with probability of success being constant from trial to trial. Also it is extensively used in pharmaceutical testing and military applications. In addition to these practical application, as we shall see that this distribution gives rise to many other special probability distributions.

Example 4.1: The probability that a newly designed electronic component will survive a given shock test is 5/6. Find the probability that exactly 3 of the next 4 components tested survive.

Solution: Assuming that the tests are independent and $p = 5/6$ for each of the 4 tests, we have

$$P\{x = 3\} = C_3^4 \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right) = 4\left(\frac{125}{1296}\right) = 0.386,$$

where x denotes the number of components survived.

Example 4.2: A machine is producing a large number of bolts. In a box of these bolts, 95% are within the permissible limits with respect to diameter. Seven bolts are drawn at random from the box. Determine the probability that, (a) two, and (b) more than or equal to two, of the seven bolts are not within the permissible limits with respect to diameter.

Solution: Let p be the probability of a bolt not being within permissible limits, then $p = 0.05$, and $q = 0.95$.

If x denotes the number of bolts with non-permissible limits out of seven selected, then

(a) $$P\{x = 2\} = C_2^7 (0.05)^2 (0.95)^5$$
$$= 21(0.0025)(0.95)^5 = 0.0406$$

(b) $$P\{x \geq 2\} = 1 - P\{x = 0\} - P\{x = 1\}$$
$$= 1 - C_0^7 (0.05)^0 (0.95)^7 - C_1^7 (0.05)(0.95)^6$$
$$= 1 - 0.630 - 0.232 = 0.138.$$

Example 4.3: A multiple choice test consists of 8 questions with 3 choices to each question of which only one is correct. A student answers each question by tossing a fair dice and marking the first choice if he gets 1 or 2, the second choice if he gets 3 or 4 and the third choice if he gets 5 or 6. To get admission the student must mark at least 75% answers correct. What is the probability that the students gets admission if there are no negative marking?

Solution: Since there are three choices to a question and probability of getting 1, 2 or 3, 4 or 5, 6 which are equally likely, and hence probability of marking correct answer $p = 1/3$ and thus $q = 2/3$.

So the probability of getting x correct answers out of 8 is given by

$$P\{X = x\} = p(x) = C_x^8\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{8-x}, \quad x = 0, 1, 2, \ldots, 8.$$

Since to get admission student must mark 75% of 8, that is 6 or more correct choices th probability is

$$P\{x \geq 6\} = p(6) + p(7) + p(8) = C_6^8\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^2 + C_7^8\left(\frac{1}{3}\right)^7\left(\frac{2}{3}\right) + C_8^8\left(\frac{1}{3}\right)^8 = 0.0197.$$

4.3.1 Constants of a Binomial Variate

The mean is,

$$E(X) = \sum_{x=0}^{n} x\, C_x^n p^x q^{n-x} = \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x} = np(p+q)^{n-1} = np$$

Also,

$$E(X^2) = E[X(X-1) + X]$$

$$= \sum_{x=0}^{n} x(x-1)c_x^n p^x q^{n-x} + \sum_{x=0}^{n} x c_x^n p^x q^{n-x}$$

$$= n(n-1)p^2(p+q)^{n-2} + np = n(n-1)p^2 + np$$

Hence, the *variance* is

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$
$$= n(n-1)p^2 + np - n^2p^2 = np - np^2 = npq.$$

The m.g.f. about origin

$$M_0(t) = E(e^{tx}) = \sum_{x=0}^{n} e^{tx} C_x^n p^x q^{n-x}$$

$$= \sum_{x=0}^{n} C_x^n (pe^t)^x q^{n-x} = (pe^t + q)^n \quad \ldots(4$$

This gives,

$$\mu_1' = [n(pe^t + q)^{n-1} pe^t]_{t=0} = np$$
$$\mu_2' = [n(n-1)(pe^t + q)^{n-2} p^2 e^{2t} + n(pe^t + q)^{n-1} pe^t]_{t=0}$$
$$= n(n-1)p^2 + np.$$

Hence, $\mu_2 = \mu_2' - (\mu_1')^2 = npq$, as obtained earlier.

We can obtain moments about the mean from the *m.g.f. about the mean* given by

$$M_{\bar{x}}(t) = e^{-npt}(pe^t + q)^n. \quad \ldots(4$$

Differentiating it successively w.r.t. t and substituting $t = 0$, we find that

$$\mu_1 = 0, \quad \mu_2 = npq, \quad \mu_3 = npq(q-p), \quad \mu_4 = npq[1 + 3(n-2)pq].$$

Thus, the *coefficients of skewness and kurtosis* are given by

$$\left.\begin{aligned} \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq} \\ \text{and,} \quad \beta_2 &= \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq} \end{aligned}\right\} \quad \ldots(4.12)$$

Hence, the binomial curve is symmetrical when $p = 1/2$. It is positively skewed for $p < 1/2$ and negatively skewed for $p > 1/2$, as already shown in Fig. 4.2(a), (b) and (c) for $p = 0.1$, 0.9 and 0.5 respectively. Further as $n \to \infty$, β_1 tends to zero and β_2 tends to 3, hence the distribution tends to be symmetric.

Next, the *mode* being the most probable number of successes in a series of n independent trials of constant probability is obtained as follows.

The chances for x successes will be greater than that of $x - 1$ and also that of $(x + 1)$, if

$$C_{x-1}^n p^{x-1} q^{n-x+1} < C_x^n p^x q^{n-x} > C_{x+1}^n p^{x+1} q^{n-x-1}$$

That is, if

$$\frac{x}{n-x+1}\frac{q}{p} < 1 > \frac{n-x}{x+1}\frac{p}{q}$$

It can be simplified to

$$np - q < x < np + p$$

or,

$$(n+1)p - 1 < x < (n+1)p.$$

Hence, the most probable number of successes is the integral part of $(n + 1)p$. In case $(n + 1)p$ is an integer, then there are two modes $(n + 1)p$ and $(n + 1)p - 1$ and the distribution is said to be *bimodal*.

4.3.2 Sum of Two Independent Binomial Variates

Let X and Y be two independent binomial variates with parameters n_1, p_1 and n_2, p_2 respectively. Then their *m.g.f.*'s are given respectively by

$$M_X(t) = (q_1 + p_1 e^t)^{n_1} \quad \text{and} \quad M_Y(t) = (q_2 + p_2 e^t)^{n_2}.$$

Consider the *mgf* of the *r.v.* $X + Y$, we have

$$M_{X+Y}(t) = M_X(t)\, M_Y(t), \text{ since } X \text{ and } Y \text{ are independent}$$

$$= (q_1 + p_1 e^t)^{n_1} (q_2 + p_2 e^t)^{n_2} \quad \ldots(4.13)$$

Since (4.13) can't be expressed in the form $(q + pe^t)^n$, thus from the uniqueness property of *m.g.f.*'s it follows that $X + Y$ is not a binomial variate.

However, if we conclude that $p_1 = p_2 = p$, (say), then from (4.13), we obtain

$$M_{X+Y}(t) = (q + pe^t)^{n_1+n_2}, \quad \ldots(4.14)$$

which is the *m.g.f.* of a binomial variate with parameters $n_1 + n_2$ and p.

Thus, the sum of the two independent binomial variates is again a binomial variate only if $p_1 = p_2$.

Example 4.4: (The probability that a patient recovers from a rare blood disease is 0.4. If 15 ... are known to have contracted this disease, then find the mean and variance of the ... survived.) Calculate $\mu \pm 2\sigma$ and use Chebyshev's inequality to interpret this interval.

Solution Let X be the number of people that survive.

Here $n = 15$ and $p = 0.4$. Hence,

$$\mu = np = 15(0.4) = 6.0$$
$$\sigma^2 = npq = 15(0.4)\,(0.6) = 3.6$$

This gives $\mu \pm 2\sigma = 6 \pm 2\,(1.897) = 6 \pm 3.794$.

Chebyshev's inequality, refer to (3.48), is

$$P[\mu - k\sigma < x < \mu + k\sigma] \geq 1 - \frac{1}{k^2}$$

For $\mu = 6$, $\sigma = 1.897$ and $k = 2$, it gives

$$P[2.206 < X < 9.794] \geq 3/4.$$

This interprets that the number of recoveries among 15 patients subjected to the given ... has a probability of at least 3/4 of falling between 2.206 and 9.794 or since the data is ... between 3 and 9 inclusive.

Example 4.5: In sampling a large number of parts manufactured by a machine, the mean ... of defectives in a sample of 20 is 2. Out of 1000 such samples, how many would be expect... contain at least three defective parts.

Solution: If p is the probability of a part being defective, then we have $np = 2$, thus $p = 2/20$ ... Therefore, probability of a non-defective part = 0.9.

Let X denote the number of defective parts in a sample of size 20, then

$$P\{X \geq 3\} = 1 - [P\{X = 0\} + P\{X = 1\} + P\{X = 2\}]$$
$$= 1 - [C_0^{20}(0.9)^{20} + C_1^{20}(0.1)\,(0.9)^{19} + C_2^{20}(0.1)^2(0.9)^{18}]$$
$$= 1 - (0.9)^{18}[0.81 + 0.09 \times 20 + 190 \times .01]$$
$$= 1 - 0.15\,[4.51] = 0.324.$$

Thus, the expected number of samples having at least three defective parts out of 1,000 sam...

$$= 1{,}000 \times 0.324 = 324.$$

Example 4.6: In a precision bombing attack there is a 50% chance that any bomb will strik... target. Two direct hits are required to destroy the target completely. How many bombs mus... dropped to give a 99% chance or better for completely destroying the target?

Solution: If p is the probability that the bomb hits the target, then $p = 1/2$.

Let n be the number of bombs needed to be dropped to ensure 99% or better chanc... completely destroy the target. Then the probability that out of n, at least two strike the target m... be greater than 0.99.

Let X be the random variable representing the number of bombs striking the target, then ... binomial variate with parameters n and $\frac{1}{2}$, and

$$p(X \geq 2) \geq 0.99$$
$$[1 - p(X \leq 1)] \geq 0.099$$
$$[1 - p(0) - p(1)] \geq 0.99$$

... if

$$1 - (1 + n)\left(\frac{1}{2}\right)^n \geq 0.99$$

... if

$$2^n \geq 100 + 100n.$$

This inequality is satisfied for $n \geq 11$, hence minimum of 11 bombs are needed to be dropped to ... get 99% or better chance to destroy the target completely.

Example 4.7: Seven coins are tossed and number of heads are noted. The experiment is repeated 128 times and the following distribution is obtained

No. of heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128

Fit a binomial distribution assuming the coin to be unbiased.

Solution: We have, $p = q = \frac{1}{2}$ and $N = 128$. Let X be a r.v. denoting the number of heads out of the seven, then X is a binomial variate with parameter 7 and 1/2, and probability mass function

$$p(x) = C_x^7 p^x q^{7-x} = C_x^7 \left(\frac{1}{2}\right)^7; \quad x = 0, 1, \ldots, 7.$$

Thus, expected frequencies are given by

$$f(x) = Np(x) = 128 C_x^7 \left(\frac{1}{2}\right)^7 = C_x^7; \quad x = 0, 1, \ldots, 7$$

Hence,

$$f(0) = 1, \quad f(1) = 7, \quad f(2) = 21, \quad f(3) = 35,$$
$$f(4) = 35, \quad f(5) = 21, \quad f(6) = 7, \quad f(7) = 1.$$

4.4 MULTINOMIAL DISTRIBUTION

In case each trial has more than 2 possible outcomes the binomial distribution becomes multinomial. For example, the drawing of a card from a pack with replacement is a multinomial distribution when we are interested in the 4 suits. Similarly selecting items with replacement from a manufactured product classified as being good, average, not acceptable is a multinomial distribution.

In general, if a given trial can result in one of the k outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$; $\sum_{i=1}^{k} p_i = 1$, and, if $X_1, X_2, \ldots, X_k$ are the r.v representing the number of occurrences of $E_1, E_2, \ldots, E_k$, then the probability that in n trials $X_1, X_2, \ldots, X_k$ take respectively the values $x_1, x_2, \ldots, x_k$, is

$$p(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k},$$

where $\sum_{i=1}^{k} x_i = n, \; 0 \le x_i \le n.$

The distribution defined by (4.15) is called *multinomial distribution*. It defines a probability distribution, since

$$\sum_x p(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \sum_x \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$
$$= (p_1 + p_2 + \ldots + p_k)^n = 1.$$

The *m.g.f.* of the multinomial distribution is given by

$$M_X(t) = M_{(X_1, X_2, \ldots, X_k)}(t_1, t_2, \ldots, t_k) = E\left[e^{\sum_{i=1}^{k} t_i x_i} \right]$$

$$= \sum_x \left[\frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} e^{\sum_{i=1}^{k} t_i x_i} \right]$$

$$= \sum_x \left[\frac{n!}{x_1! x_2! \ldots x_k!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} \ldots (p_k e^{t_k})^{x_k} \right]$$

$$= \left(p_1 e^{t_1} + p_2 e^{t_2} + \ldots + p_k e^{t_k} \right)^n$$

From (4.16)

$$M_{X_1}(t) = M_X(t_1, 0, \ldots, 0) = \left(p_1 e^{t_1} + p_2 + \ldots + p_k \right)^n$$

$$= \left[(1 - p_1) + p_1 e^{t_1} \right]^n$$

which is the *m.g.f.* of a binomial variate X_1 with parameters n_1 and p_1. Similarly, we can obtain m.g.f. for X_2, etc. Further,

$$E(X_i) = n_i p_i, \text{ and } \text{Var}(X_i) = n_i p_i q_i, \; i = 1, 2, \ldots, k.$$

Example 4.8: Out of a lot containing 5 good, 4 faulty and 3 partially faulty but working batteries three have been selected at random with replacement. Find the probability that selection consists exactly one of each type.

Solution: Let p, q and r be respectively the probabilities of selecting good, faulty and partially faulty batteries at a single draw, then $p = 5/12, \quad q = 4/12, \quad r = 3/12.$

If X_1, X_2 and X_3 are respectively the random variables giving the number of good, faulty and partially faulty batteries out of 3, then

$$P\{X_1 = 1, X_2 = 1, X_3 = 1\} = \frac{3!}{1!1!1!} \left(\frac{5}{12}\right)\left(\frac{4}{12}\right)\left(\frac{3}{12}\right) = \frac{5}{24}.$$

4.5 HYPERGEOMETRIC DISTRIBUTION

The difference between binomial distribution and hypergeometric distribution lies in the procedure sampling is made. When the population is finite and sampling is done without replacement so the events although random, become stochastically dependent, then the resulting distribution is no more binomial. It leads to hypergeometric distribution.

Let us suppose that a random sample of n items is selected without replacement from N items, k of which are classified as successes and $N - k$ as failures. If X is the random variable giving the number of successes out of the n selected, then

$$P\{X = x\} = \frac{C_x^k \times C_{n-x}^{N-k}}{C_n^N}, \; x = 1, 2, \ldots, \min(n, k), \qquad \ldots(4.17)$$

where, $$\sum_x P\{X = x\} = \frac{\sum_x C_x^k \times C_{n-x}^{N-k}}{N_{C_n}} = N_{C_n} / N_{C_n} = 1.$$

Thus (4.17) defines a probability distribution.

A random variable X with probability distribution given by (4.17) is called hypergeometric variable and the distribution is called hypergeometric probability distribution.

4.5.1 Constants of Hypergeometric Distribution

The *mean* is $$E(X) = \sum_{x=0}^{n} \frac{x C_x^k C_{n-x}^{N-k}}{C_n^N} = k \sum_{x=1}^{n} \frac{(k-1)!}{(x-1)!(k-x)!} \frac{C_{n-x}^{N-k}}{C_n^N}$$

$$= k \sum_{x=1}^{n} \frac{C_{x-1}^{k-1} C_{n-x}^{N-k}}{C_n^N}$$

$$= k \sum_{y=0}^{n-1} \frac{C_y^{k-1} C_{n-1-y}^{N-k}}{C_n^N} \quad (y = x - 1)$$

Writing $C_{n-1-y}^{N-k} = {}^{(N-1)-(k-1)}C_{n-1-y}$, and $C_n^N = \frac{N}{n} {}^{N-1}C_{n-1}$, we obtain

$$E(X) = \frac{nk}{N} \sum_{y=0}^{n-1} \frac{C_y^{k-1} \, {}^{(N-1)-(k-1)}C_{n-1-y}}{{}^{N-1}C_{n-1}} = \frac{nk}{N} \qquad \ldots(4.18)$$

Similarly, we can show that

$$E(X^2) = E[X(X-1)] + E(X)$$
$$= \frac{k(k-1)\,n(n-1)}{N(N-1)} + \frac{nk}{N}$$

Thus, the *variance* is

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$
$$= \frac{k(k-1)\,n(n-1)}{N(n-1)} + \frac{nk}{N} - \left(\frac{nk}{N}\right)^2$$
$$= \frac{Nkn(k-1)(n-1) + N(N-1)nk - (N-1)n^2k^2}{N^2(N-1)}$$
$$= nk\frac{[N(k-1)(n-1) + N(N-1) - (N-1)nk]}{N^2(N-1)}$$
$$= \frac{nk(N-k)(N-n)}{N^2(N-1)}$$
$$= \frac{N-n}{N-1}\,n\,\frac{k}{N}\left(1-\frac{k}{N}\right) \qquad \ldots(4.19)$$

From (4.18) and (4.19), we observe that as $N \to \infty$ and $\frac{k}{N} \to p$, $E(X) \to np$ and Var $(x) \to npq$.

Thus hypergeometric distribution approximates to binomial distribution when N is quite large (as compared to n), as one expects that the nature of N items changes very little in each draw. As a rule of thumb the approximation is good when $\frac{n}{N} \le 0.05$.

Hypergeometric distribution finds applications in *acceptance sampling* and electronic testing particular where the item tested is destroyed and can't be replaced in the sample. Further, the distribution can also be extended to the case where the N items can be partitioned into more than two classes.

Example 4.9: A lot consisting of 100 fuses is inspected by the following procedure. Five of these fuses are chosen at random and tested; if 4 or more work at the correct amperage, the lot is accepted. If there are 20 defective fuses in the lot, find the probability of acceptance.

Solution: Let X be the number of fuses working correctly out of the five selected, then

$$P\{X \ge 4\} = P\{X = 4\} + P\{X = 5\}$$
$$= \frac{C_4^{80} \times C_1^{20}}{C_5^{100}} + \frac{C_5^{80} \times C_0^{20}}{C_5^{100}}$$
$$= 0.42 + 0.32 = 0.74.$$

Example 4.10: A group of 10 individuals being used in a biological study contains 3 people with blood type O, 3 with blood type A and 4 with blood type B. What is the probability that a random sample of size 6 will contain 2 with type O, 2 with type A and 2 with type B?

Solution: Here the 10 items can be partitioned into three classes each of size $a_1 = 3$, $a_2 = 3$ and $a_3 = 4$. Extending the hypergeometric distribution to this case, if x_1, x_2 and x_3 are the number of people selected out of a_1, a_2 and a_3 respectively then the probability

$$P(x_1 = 2, x_2 = 2, x_3 = 2) = \frac{(C_2^3)(C_2^3)(C_2^4)}{10}$$
$$= \frac{3 \times 3 \times 6}{10 \times 3 \times 7} = 9/35.$$

4.6 NEGATIVE BINOMIAL DISTRIBUTION

Consider the sequence of independent Bernoulli trials as in case of binomial experiment, with the exception that trials will be repeated until a fixed number of successes, say k occur. Let the k th success occur at the $(x + k)$th trial. Thus, the last trial, that is, the $(x + k)$th trial must be success whose probability is p. In the preceding $(x + k - 1)$ trials, we must have $(k - 1)$ successes whose probability by binomial distribution is given by

$${}^{x+k-1}C_{k-1}\,p^{k-1}q^x.$$

Therefore, by multiplication rule the probability that kth success occurs at the $(x + k)$th trial denoted by $p(x)$ is given by

$$p(x) = {}^{x+k-1}C_{k-1}\,p^k q^x, \quad x = 0, 1, 2, \ldots \qquad \ldots(4.20)$$

Since
$${}^{x+k-1}C_{k-1} = \frac{(x+k-1)(x+k-2)\ldots(k+1)\,k}{x!}$$
$$= (-1)^x \frac{(-k)(-k-1)\ldots(-k-x+2)(-k-x+1)}{x!}$$
$$= (-1)^x\,C_x^{-k}$$

Thus, from (4.20) we have

$$\sum_{x=0} p(x) = \sum_{x=0}^{x+k-1} C_{k-1}\,p^k q^x = p^k \sum_{x=0} (-1)^x C_x^{-k} q^x$$
$$= p^k \sum_{x=0} {}^{-k}C_x\,(-q)^x = p^k\,(1-q)^{-k} = 1.$$

Hence (4.20) represents a probability distribution.

A random variable X taking on non-negative values with probability distribution given by (4.20) is called negative binomial variate and the distribution is called negative binomial distribution.

The distribution is named so since $p(x)$ is the $(x + 1)$th term in the expansion of $p^k(1 - q)^{-k}$, a binomial expression with a negative index.

4.6.1 Constants of Negative Binomial Distribution

The *m.g.f.* of negative binomial distribution is

$$M_X(t) = \sum_x p(x)\, e^{tx} = {}^{-k}C_x\, p^k \sum_x {}^{-k}C_x\, (-q)^x e^{tx}$$

$$= p^k \sum_x {}^{-k}C_x\, (-qe^t)^x$$

$$= p^k (1 - qe^t)^{-k}$$

The *mean* is

$$E(X) = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$$

$$= [p^k(-k)(1 - qe^t)^{-k-1}(-qe^t)]_{t=0}$$

$$= k\frac{q}{p}$$

$$E(X^2) = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$= [p^k(-k)(-k-1)(1 - qe^t)^{-k-2}(-qe^t)^2 + p^k(-k)(1 - qe^t)^{-k-1}(-qe^t)]_{t=0}$$

$$= k\frac{q}{p} + k(k+1)\left(\frac{q}{p}\right)^2$$

The *variance* is $\text{Var}(X) = E(X^2) - [E(X)]^2$

$$= k\frac{q}{p} + k(k+1)\left(\frac{q}{p}\right)^2 - k^2\left(\frac{q}{p}\right)^2$$

$$= k\frac{q}{p}\left(1 + \frac{q}{p}\right)$$

From (4.22) and (4.23), we observe that mean is less than variance in case of a negative binomial distribution.

Example 4.11: A drug is known to bring relief in 80% of the cases where it is used. Find the probability that the fifth patient to experience relief is the seventh patient to receive the drug during a given week.

Solution: Here $k = 5$, $x = 2$ and $p = 4/5$, thus

$$p(2) = {}^{2+5-1}C_{5-1}\left(\frac{4}{5}\right)^5\left(1 - \frac{4}{5}\right)^2$$

$$= 10 \times \frac{1024}{3125 \times 25} = 0.131.$$

Example 4.12: In a basketball championship series between the two teams A and B, the team which wins three games out of five will be the winner. Suppose that team *A* has probability 0.6 of winning over the team *B*.

(a) What is the probability that team *A* will win the series in four games?

(b) What is the probability that team *A* will win the series?

Solution: (a) Here $k = 3$, $x = 1$ and $p = 0.6$, thus

$$p(1) = {}^{1+3-1}C_{3-1}\,(0.6)^3\,(0.4)$$

$$= 3\,(.216)\,(.4) = 0.2592.$$

(b) Probability that team *A* wins the series

$$= p(x=0) + p(x=1) + p(x=2)$$

$$= {}^{3-1}C_{3-1}\,(0.6)^3 + {}^{1+3-1}C_{3-1}\,(0.6)^3(0.4) + {}^{2+3-1}C_{3-1}\,(0.6)^3(0.4)^2$$

$$= (0.6)^3 + 3\,(0.6)^3\,(0.4) + 6\,(0.6)^3\,(0.4)^2$$

$$= 0.2160 + 0.2592 + 0.2074$$

$$= 0.6826.$$

4.7 GEOMETRIC DISTRIBUTION

Suppose we have a series of independent Bernoulli trials with constant probability p of success. Then the probability that since there are x failures preceding the first success, is given by

$$p(x) = q^x p, \quad x = 0, 1, 2, \ldots; q = 1 - p. \qquad \ldots(4.24)$$

Further,

$$\sum_x p(x) = \sum_x q^x p = p[1 + q + q^2 + \ldots]$$

$$= p\frac{1}{1-q} = \frac{p}{p} = 1.$$

thus, (4.24) defines a probability distribution.

A random variable X taking non-negative integral values with probability distribution given by (4.24) is called a geometric variable and the distribution is called the geometric distribution.

We observe that (4.24) can also be obtained from (4.20) for $k = 1$. Hence, geometric distribution is a special case of the negative binomial distribution.

Example 4.13: If $p = 0.10$ is the probability of a connection during the peak business hours at a telephone exchange, then find the probability that 4 attempts are necessary for a success call.

Solution: Using the geometric distribution with $x = 4$ and $p = 0.10$, we obtain

$$P(x = 4) = (0.9)^4\,(0.1) = .06561.$$

Example 4.14: An expert shot hits a target 95% of the time. What is the probability that the expert will miss the target for the first time on the seventeenth shot?

Solution: Using the geometric distribution with $x = 16$ and $p = .05$, we obtain

$$P(x = 16) = (0.95)^{16}\,(0.05)$$

$$= 0.022.$$

4.7.1 Constants of Geometric Distribution

The *mean* is

$$E(X) = \sum_{x=0} xp(x) = p\sum_{x=0} xq^x$$

$$= p\,[q + 2q^2 + 3q^3 + \ldots]$$

$$= \frac{pq}{(1-q)^2} = \frac{q}{p}$$

$$E(X^2) = E\,[X(X-1)] + E(X)$$

$$= p\sum_{x=0} x(x-1)q^x + p\sum_{x=0} xq^x$$

$$= 2pq^2 \sum_{x=2} \frac{x(x-1)}{2.1} q^{x-2} + \frac{q}{p}$$

$$= 2pq^2 (1-q)^{-3} + \frac{q}{p} = 2\frac{q^2}{p^2} + \frac{q}{p}$$

The *variance* is

$$\text{Var}(X) = E(X)^2 - [E(X)]^2$$

$$= 2\frac{q^2}{p^2} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q^2}{p^2} + \frac{q}{p} = \frac{q}{p^2} \quad \ldots(4$$

From (4.25) and (4.26), we note that in case of geometric distribution, Var (X) > Mean (X).

The *m.g.f.* about origin is

$$M_x(t) = E(e^{tX}) = \sum_{x=0} p(x)e^{tx}$$

$$= p\sum_{x=0} q^x e^{tx}$$

$$= \frac{p}{1-qe^t} \quad \ldots(4$$

4.7.2 Lack of Memory

An important characteristic of geometric distribution is that *it lacks memory* in the sense explained below.

Suppose an event E can occur at one of the times $t = 0, 1, 2, \ldots$ and the occurrence (waiting) ti X has a geometric distribution with parameter p, that is,

$$P(X = t) = q^t p; \quad t = 0, 1, 2, \ldots$$

Let the event E has not occurred before k, that is, $X \geq k$. Define $y = X - k$ as the additional ti needed for E to occur. Then we can show that

$$P(Y = t \mid X \geq k) = P(X = t) = pq^t \quad \ldots(4.28)$$

that is, *the additional time to wait has the same distribution as the initial time to wait.*

The proof follows as given below.

Consider,

$$P(X \geq k) = \sum_{x=k} pq^x = pq^k\,[1 + q + q^2 + \ldots]$$

$$= \frac{pq^x}{1-q} = q^x \quad \ldots(4.29)$$

We have,

$$P(Y \geq t \mid X \geq k) = \frac{P(Y \geq t \cap X \geq k)}{P(X \geq k)}$$

$$= \frac{P(X - k \geq t \cap X \geq k)}{P(X \geq k)}$$

$$= \frac{P(X \geq k + t)}{P(X \geq k)} = \frac{q^{t+k}}{q^k} = q^t, \quad \text{using (4.29)}$$

Thus,

$$P(Y = t \mid X \geq k) = P(Y \geq t \mid X \geq k) - P(Y \geq t + 1 \mid X \geq k)$$

$$= q^t - q^{t+1} = q^t(1 - q) = pq^t = P(X = t)$$

4.8 POISSON DISTRIBUTION

Consider the situation when in binomial distribution n is large and p is small such that the average number of successes np is a finite constant, say equal to λ.

The probability of x successes is given by

$$p(x) = C_x^n p^x q^{n-x}.$$

Rewriting it as

$$p(x) = \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \cdot \frac{n!}{n^x(n-x)!}$$

$$= \frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{n-1}{n}\right).$$

Taking limit as $n \to \infty$, it gives

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \infty. \quad \ldots(4.30)$$

Also,

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Thus, (4.30) defines a probability distribution.

A random variable X, taking on non-negative integral values, with probability distribution (4.30) called 'Poisson variate' and the distribution is called 'Poisson distribution'.

The graphs of Poisson probability distribution for $\lambda = 0.5$, 1 and 4 are shown in Fig. 4.3(a), (b) and (c) respectively.

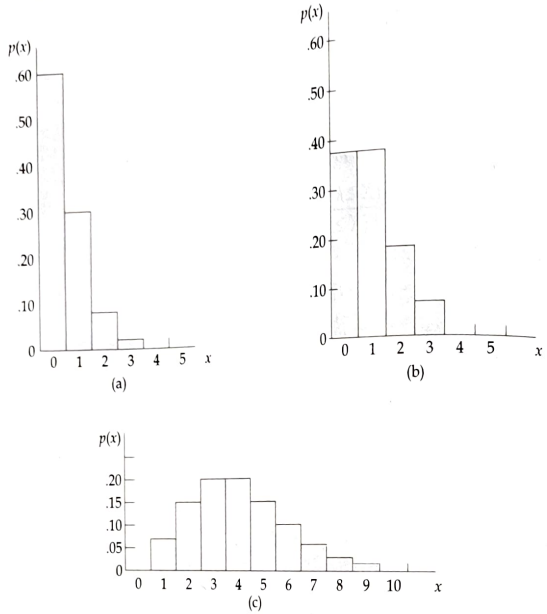

Fig. 4.3

Besides approximating binomial distribution for large n and small p (normally we a Poisson distribution when $n \geq 20$ and $p \leq 0.05$), the Poisson distribution has numerous applicati A few situations where a Poisson variate is applied are:

1. Number of printing errors per page in a printed book.
2. Number of defective fuses in a pack of 100.
3. Number of accidents per year at a busy crossing.
4. Number of wrong telephone numbers dialed in a day.
5. Number of α particles discharged in a fixed period of time from some radioactive mate
6. Number of customers arriving at a service counter on a given day.
7. Number of cars passing a crossing per minute during the busy hours of a day.
8. Number of deaths from a disease (not in the form of an epidemic) such as heart attac cancer or snake bite.

4.8.1 Constants of Poisson Distribution

The *mean* is

$$E(X) = \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \qquad ...(4.31)$$

Similarly,

$$E(X^2) = E[X(X-1) + X] = E[X(X-1)] + E(X)$$

$$= \sum_{x=0}^{\infty} x(x-1) e^{-\lambda} \frac{\lambda^x}{x!} + \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!}$$

$$= \lambda^2 \sum_{x=2}^{\infty} e^{-\lambda} \frac{\lambda^{x-2}}{(x-2)!} + \lambda \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} = \lambda^2 + \lambda$$

The *variance* is

$$\text{Var}(X) = E(X^2) - (E(X))^2$$
$$= \lambda^2 + \lambda - \lambda^2 = \lambda \qquad ...(4.32)$$

Thus, in case of Poisson variate mean is equal to variance.

The *m.g.f about origin* is

$$M_0(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} . e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} . e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \qquad ...(4.33)$$

The *m.g.f. about mean* is

$$M_{\bar{X}}(t) = e^{-\lambda t} M_0(t) = e^{-\lambda t} . e^{-\lambda} . e^{\lambda e^t}$$
$$= \exp[\lambda e^t - (1 + t)]. \qquad ...(4.34)$$

From this, we can calculate the *moments about the mean*. We can very easily check that

$$\mu_2 = \frac{d^2}{dt^2}[M_{\bar{X}}(t)]_{t=0} = \lambda, \ \mu_3 = \frac{d^3}{dt^3}[M_{\bar{X}}(t)]_{t=0} = \lambda, \text{ and } \mu_4 = \frac{d^4}{dt^4}[M_{\bar{X}}(t)]_{t=0} = 3\lambda^2 + \lambda \qquad ...(4.35)$$

Thus *the coefficients of skewness and kurtosis* are given by

$$\left.\begin{aligned} \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \\ \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{3\lambda^2 + \lambda}{\lambda^2} = 3 + \frac{1}{\lambda} \end{aligned}\right\} \qquad ...(4.36)$$

Since $\beta_1 > 0$, so the Poisson distribution is a positively skewed distribution, and also since β_2 the distribution is leptokurtic, and further, as $\lambda \to \infty$, then $\beta_1 = 0$ and $\beta_2 = 3$, the distribution tend be symmetric as already indicated in Fig. 4.3(a), (b) and (c).

Next, the *mode* is that value of x for which $\frac{e^{-\lambda}\lambda^x}{x!}$ is greater than its preceding and proceed terms, that is, x for which

$$\frac{\lambda^{x-1}e^{-\lambda}}{(x-1)!} \le \frac{\lambda^x e^{-\lambda}}{x!} \ge \frac{\lambda^{x+1}e^{-\lambda}}{(x+1)!}$$

Simplifying it gives

$$\lambda - 1 \le x \le \lambda.$$

Thus, if λ is not an integer then mode is the integral value of λ and if λ is an integer then distributio bimodal with λ and $\lambda - 1$ as its modal values.

4.8.2 Sum of Two Independent Poisson Variates

Let X and Y be two independent Poisson variates with parameters λ_1 and λ_2 respectively. Th their *m.g.f.*'s are given respectively by

$$M_X(t) = \exp[\lambda_1(e^t - 1)] \quad \text{and} \quad M_Y(t) = \exp[\lambda_2(e^t - 1)]$$

Consider the *m.g.f.* of the random variable $X + Y$, we have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\, M_Y(t), \quad \text{since } X \text{ and } Y \text{ are independent} \\ &= \exp[\lambda_1(e^t - 1)] \exp[\lambda_2(e^t - 1)] \\ &= \exp[\lambda_1 + \lambda_2)(e^t - 1)] \end{aligned}$$

which is the *m.g.f.* of a Poisson variate with paramete $(\lambda_1 + \lambda_2)$.

Hence, the sum of two independent Poisson variates with parameters λ_1 and λ_2 is also a Poisson vari with parameter $(\lambda_1 + \lambda_2)$.

Example 4.15: Past experience shows that 5% of the books bound at a certain bindery ha defective bindings. Find the probability that 2 of 100 books bound by this bindery will ha defective binding using (a) binomial distribution, (b) Poisson distribution.

Solution: (a) Using binomial

$$\begin{aligned} P(x = 2) &= {}^{100}C_2\,(0.05)^2\,(0.95)^{98} \\ &= (4750)\,(.0025)\,(0.00656) \\ &= 0.078. \end{aligned}$$

(b) Using Poisson with $\lambda = np = 100\,(0.05) = 5$

$$P(x = 2) = \frac{5^2 e^{-5}}{2!} = \frac{25 \times (.00674)}{2} = 0.084$$

Example 4.16: Find the probability that the most 5 defective fuses will be found in a box of 2 fuses if experience shows that 2% of such fuses are defective.

Solution: Here $\lambda = np = 200(.02) = 4$.

Hence, the requisite probability is

$$\begin{aligned} P(x \le 5) &= \sum_{x=0}^{5} e^{-4}\frac{4^x}{x!} = e^{-4}\left[1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{4^5}{5!}\right] \\ &= (0.0183)\,[1 + 4 + 8 + 10.6667 + 10.6667 + 8.5333] \\ &= (0.0183)\,(42.8667) = 0.7845 \end{aligned}$$

Example 4.17: Consider an experiment that consists of counting the number of α particles given off in a one-second interval by one gram of radioactive material. If past experience shows that on the average 3.2 such α-particles are given off, find the probability that more than 2 α-particles will appear?

Solution: If r.v X denotes the number of α-particles given off in a second interval, then X will be a Poisson variate with mean $\lambda = 3.2$.

Hence, the requisite probability is

$$\begin{aligned} P[x > 2] &= 1 - P[X \le 2] = 1 - \sum_{x=0}^{2} e^{-3.2}\frac{(3.2)^x}{x!} = 1 - \left(e^{-3.2} + 3.2e^{-3.2} + \frac{(3.2)^2}{2!}e^{-3.2}\right) \\ &= 1 - (0.041 + 0.130 + 0.209) = 0.62. \end{aligned}$$

Example 4.18: If the average number of road accidents reported daily in a township is 5, what proportion of days have less than 3 accidents reported? What is the probability that there will be 4 accidents reported per day in exactly three days out of five, assuming that the number accidents on different days is independent?

Solution: Let X be the number of accidents reported daily, then X is Poisson variate with mean 5. Hence, the probability that there will be less than three accidents reported is

$$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) = e^{-5} + e^{-5}\frac{5^1}{1!} + e^{-5}\frac{5}{2!} \approx 0.125.$$

Thus, over the long run on about 12.5% days the number of accidents reported will be less than or equal to 3.

Since it has been assumed that number of accidents on different days is independent, thus number of days in a five-day duration that has exactly 4 accidents reported is a binomial distribution with parameter $n = 5$ and probability of 'success' p, given by

$$p = P[X = 4] = e^{-5}\frac{5^4}{4!} \approx 0.175.$$

Thus, the probability that exact 3 out of the next five days will report 4 accidents daily

$$= C_3^5(0.175)^3(0.825)^2 = 0.0365.$$

Example 4.19: A manufacturer who produces medicine bottles finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes. Using Poisson distribution find how many boxes will contain (a) no defective, (b) at least two defectives.

Solution: We have,

$N = 100$, $n = 500$, p = Probability of a defective bottle = 0.001,

$\lambda = np = 500 \times 0.001 = 0.5$.

If the r.v. X denotes the number of defective bottles in a pack of 500, then by Poisson distribution

$$P\{X = x\} = e^{-0.5}\frac{(0.5)^x}{x!} = \frac{0.6065(0.5)^x}{x!}; \quad x = 0, 1, 2, \ldots$$

Hence, in a lot of 100 boxes the frequency of boxes with x defective bottles is given by

$$f(x) = NP\{X = x\} = \frac{100 \times 0.6065 \times (0.5)^x}{x!}$$

(i) Number of boxes with no defective

$$= 100P(X = 0) = 100 \times 0.6065 \approx 61.$$

(ii) Number of boxes with at least two defectives

$$= 100\,[P(X \geq 2)] = 100\,[1 - P(X = 0) - P(X = 1)]$$

$$= 100\,[1 - 0.6065 - 0.6065 \times 0.5] = 100 \times 0.0903 \approx 9.$$

Example 4.20: Fit a Poisson distribution to the following data which gives the number of yeast cells per square for 400 squares

No. of cells per square (x) :	0	1	2	3	4	5	6	7	8	9	10
No. of squares (f) :	103	143	98	42	8	4	2	0	0	0	0

Solution: The parameter λ of the Poisson distribution is given by

$$\lambda = \frac{1}{N}\sum f_i x_i = \frac{529}{400} = 1.32.$$

If the r.v. X denotes the number of yeast cells per square, then expected frequencies on the basis of Poisson distribution are given by

$$f(x) = NP\{X = x\} = 400\,e^{-1.32}\frac{(1.32)^x}{x!}, \; x = 0, 1, 2, \ldots, 10.$$

It gives the following frequencies:

x :	0	1	2	3	4	5	6	7	8	9	10
$f(x)$:	107	141	93	41	14	4	1	0	0	0	0

REVIEW EXERCISES

1. Define uniform distribution. Find its mean and variance. Describe a situation where this distribution is applicable.
2. Define Bernoulli trials. Find the probability that out of a sequence of n Bernoulli trials, with constant probability p for success, there are x successes.

3. Describe situations where binomial model is applicable.
4. Define a binomial variate. Find its mean and variance and m.g.f. about mean. Find the coefficient of skewness and kurtosis.
5. Show that the sum of two independent binomial variates, in general, is not a binomial variate.
6. Find the mode of the binomial distribution with parameters n and p. When is the distribution bimodal?
7. Let X denote a binomial variate with parameters n and p. Show that

$$E\left[(X - np)/\sqrt{npq}\right] = 0 \quad \text{and} \quad \text{Var}\left[(X - np)/\sqrt{npq}\right] = 1$$

Find the m.g.f. of the variate $(X - np)/\sqrt{npq}$.

(*The variate* $Y = (X - npq)/\sqrt{npq}$ *is called standard binomial variate*)

8. If $(X_1, X_2, \ldots, X_k)$ have a multinomial distribution with parameters n and p_i, $(i = 1, 2, \ldots, k)$ with $\sum_{i=1}^{k} p_i = 1$. Obtain the joint probability $p(X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_k = x_k)$. Describe a situation where multinomial distribution is applicable.
9. What is hypergeometric distribution? How does it arise? Find its mean and variance. How is it related to binomial distribution?
10. Suppose that from a population of n elements of which m are defective and $(n - m)$ are non-defective, a sample of size k is drawn without replacements. What is the probability that the sample contains exactly x defectives? Name this probability distribution.
11. Find the probability that in a sequence of Bernoulli trial with p as probability of success the kth success occurs at the $(x + k)$th trial.
12. Define a negative binomial variate. Why is it named so? Find its mean and variance. Describe a situation where it is applicable.
13. Find the probability function of X, number of times of tossing a fair coin until the first head appears. Find the mean.
14. Define a geometric variate. Derive its p.d.f. from that of a negative binomial variate.
15. Describe the *lack of memory* property of the geometric random variable.
16. Define a Poisson variate. Describe situations where Poisson variate is applicable.
17. Derive Poisson distribution as a limiting case of binomial distribution. Find its mean and variance.
18. Find the m.g.f. about mean of a Poisson variate with parameter λ. Find the first four central moments. Calculate the coefficients of skewness and kurtosis. Show that as $\lambda \to \infty$, then the distribution tends to be symmetric.
19. Let X denote a Poisson variate with parameters λ. If $Y = (X - \lambda)/\sqrt{\lambda}$, then show that $E(Y) = 0$ and Var $(Y) = 1$. Also find the m.g.f. of Y.

[*The random variable* Y *is called standard Poisson variable.*]
20. Show that the sum of two independent Poisson variates is a Poisson variate with parameter as the sum of their parameters.

PROBLEM SET

1. A student is selected from a group of 15 students to represent the group by selecting a at random from a box containing 15 tags numbered from 1 to 15. Find the probab distribution of X representing the number on the tag that is drawn. What is the probab that number drawn is more than 10?
2. It is known that disks produced by a certain company will be defective with probability independently of each other. The company sells the disk in packages of 10 and offe money-back guarantee that at most 1 of the 10 disks is defective. What proportion packages is returned? If someone buys three packages, what is the probability that exa one of them will be returned?
3. Over a long period of time it has been observed that a given shooter can hit a target o single trial with probability equal to 0.8. Suppose he fires four shots at the target. (a) W is the probability that he will hit the target exactly two times? (b) What is the probab that he will hit the target at least once?
4. The probability that a patient recovers from a rare blood disease is 0.4. If 15 people known to have contracted this disease, what is the probability that (a) at least 10 surv (b) from 3 to 8 survive, and (c) exactly 5 survive?
5. If the probability that a fluorescent light has a useful life of at least 800 hours is 0.9, find probabilities that among 20 such lights
   (a) exactly 18 will have a useful life of at least 800 hours;
   (b) at least 15 will have a useful life of at least 800 hours;
   (c) at least 2 will not have a useful life of at least 800 hours.
6. Among the 15 cities that a professional society is considering for next 3 ann conventions, 5 are in the northern part of India. To avoid arguments, the selection is left chance. If none of the cities can be chosen more than once, what are the probabilities th
   (a) none of the conventions will be held in the northern part,
   (b) all of the conventions will be held in the northern part?
7. A sortie of 20 aeroplanes is sent on operational flight. The chances that an aeroplane fails return is 5%. Find the probability that (a) one plane does not return, (b) at the most fi planes do not return, (c) what is the most probable number of returns?
8. The following data shows the results of throwing 12 fair dice 4096 times; throw of 4, 5 o being called success:

Success (x):	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f):	0	7	60	198	430	731	948	847	536	257	71	11	0

   Fit a binomial distribution and find the expected frequencies.
9. If a fair coin is tossed an even number of $2n$ times, show that the probability of obtainin more heads than tails is $\frac{1}{2}\left\{1 - C_n^{2n}\left(\frac{1}{2}\right)^{2n}\right\}$.
10. According to a genetic theory, a certain cross of guinea pigs will result in red, black, whi and gray offspring in the ratio 4: 2: 3: 1. Find the probability that among 9 offsprings will h red, 2 black, 3 white and 1 gray.

11. As a student goes to school, he encounters a traffic signal which stays green for 35 seconds, yellow for 5 sec and red for 60 seconds. Assume that he goes to school each week day between 8:00 and 8:30 a.m. and X_1, X_2, X_3 be the number of times he encounters green, yellow and red signal, respectively. Find the joint distribution for (X_1, X_2, X_3).
12. In state cup badminton championship series, the team which wins four games out of seven will be the winner. Suppose that the team A has probability 0.55 of winning over the team B and both teams A and B face each other in the championship games.
    (a) What is the probability that team A will win the series in six games?
    (b) What is the probability that the team B will win the series?
13. The probability that a person living in a certain city owns a dog is estimated to be 0.3. Find the probability that the tenth person randomly interviewed in that city is the fifth one to own a dog.
14. In a test a light switch is turned on and off until it fails. If the probability that switch will fail any time it is turned on or off is 0.001, what is the probability that the switch will fail after it has been turned on or off 1,200 times? Assuming that the conditions for the geometric distribution are met.
15. The average number of accidents on a certain section of highway is two per week. Assuming it to follow Poisson distribution, find the probability of (a) no accident on this section during a week period, (b) at most three accidents on this section during a two week period.
16. In a book of 520 pages, 390 typographical errors occur. Assuming Poisson law for the number of errors per page, find the probability that a random sample of 5 pages will contain no error.
17. Suppose that in the production of radio resistors the probability of a resistor being defective is 0.1%. The resistors are sold in lots of 200, with the guarantee that all resistors are non-defective. What is probability that a given lot will violate this guarantee?
18. The probability that a person dies when he contracts a respiratory infection is 0.002. Of the next 2000 so infected, what is the mean number that will die? What is the S.D.?
19. The probability that a student pilot passes the written test for a pilot's licence is 0.7. Find the probability that the student will pass the test, (a) on the third try, (b) before the fourth try.
20. Twenty firms are under suspicion for violation of pollution norms but all cannot be inspected. Suppose that 3 of the firms are in violation. What is the probability that, (a) inspection of 5 firms find no violation. (b) will find two violations?
21. After correcting 50 pages of the proof of a book, the proof reader finds that there are on an average 2 errors per 5 pages. How many pages would one expect to find with 0, 1, 2, 3 and 4 errors in 1000 pages of the first print of the book?
22. If X and Y are independent Poisson variates having means 1 and 3 respectively, find the variance of $3X + Y$.
23. The number of aeroplanes arriving at an airport in a 30 minute interval obeys the Poisson law with mean 25. Use Chebychev's inequality to find a lower bound for the probability that the number of planes to arrive within a given 30-minute interval will be between 15 and 35.

24. Two discrete random variable X and Y have the joint probability distribution

$$p(x, y) = \frac{9!}{x!\,y!\,(9-x-y)!}\left(\frac{1}{3}\right)^9, \quad 0 \le x \le 9, \quad 0 \le y \le 9 \quad \text{and} \quad 0 \le x + y \le 9.$$

Show that the conditional distribution of Y given $X = 3$ is binomial with parameters n and $p = 1/2$.

25. If X and Y are two independent binomial variates with parameter (n_1, p) and $(n_2$ respectively, show that

$$P\{X = r \mid (X + Y) = n\} = \frac{C_r^{n_1} \times C_{n-r}^{n_2}}{{}_{n_1+n_2}C_n}.$$

ANSWERS

1. 1/3
2. 0.5%, 0.015
3. (a) 0.1536 (b) 0.9984
4. (a) 0.0338 (b) 0.8779 (c) 0.1859
5. (a) 0.2852 (b) 0.9887 (c) 0.6083
6. (a) 0.2637 (b) 0.0220
7. (a) 0.3774 (b) 0.9997 (c) 19
8. $f(0) = 1, f(1) = 12, f(2) = 66, f(3) = 220, f(4) = 495, f(5) = 792, f(6) = 924, f(7) = 792, f(8) = 495, f(9) = 220, f(10) = 66, f(11) = 12, f(12) = 1$
11. $\frac{n!}{x_1!\,x_2!\,x_3!}(0.35)^{x_1}(0.05)^{x_2}(0.60)^{x_3}$
12. (a) 0.1853 (b) 0.3917
13. 0.0515
14. 0.3010
15. (a) 0.1353 (b) 0.4335
16. 0.0235
17. 0.1813
18. 4, 2
19. (a) 0.0630 (b) 0.9730
20. (a) 0.3991 (b) 0.1315
21. 670, 268, 54, 7, 1.

5 Special Continuous Probability Distributions

CHAPTER

5.1 INTRODUCTION

In the preceeding chapter, we have studied some specifc discrete probability distribution. Here we consider probability distributions of the continuous type which commonly occur in practice. The distributions considered through Sections 5.2 to 5.9 are *uniform*, *normal*, *log-normal*, *exponential*, *Weibull*, *gamma* and *beta distribution*. The chapter concludes with review exercises and a problem set on the distributions discussed.

5.2 CONTINUOUS UNIFORM DISTRIBUTION

It is one of the simplest continuous probability distribution with constant (uniform) probability in a closed interval, say $[a, b]$, defined as follows.

A continuous random variable X is said to have a uniform distribution over the interval $[a, b]$, if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases} \qquad \ldots(5.1)$$

Since, $$\int_a^b f(x)\,dx = \int_a^b \frac{dx}{b-a} = \frac{b-a}{b-a} = 1$$

thus, (5.1) defines a probability density function. The density function of this distribution forms a rectangle with base $(b - a)$ and height $\frac{1}{b-a}$ and that is why the distribution is oftenly called the *rectangular distribution* also. The distribution is used to model the behaviour of a continuous random variable whose values are uniformly or evenly distributed over a given interval. For example, the error x introduced by rounding an observation to the nearest inch would probably have a uniform distribution over the interval $[-0.5, 0.5]$.

and, thus the probability that the rounding error is less than 0.2 in magnitude, that is,

$$P(-.2 < x < .2) = \int_{-.2}^{.2} \frac{dx}{.5-(-.5)} = 0.4$$

5.2.1 Constants of Uniform Distribution

The *mean* is $$E(X) = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}$$ ...(5.

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{b-a}\cdot\frac{b^3-a^3}{3} = \frac{a^2+ab+b^2}{3}$$

The *variance* is $$\sigma^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2$$

$$= \frac{(b-a)^2}{12}$$ ...(5.

The *m.g.f. about origin* is

$$M_o(t) = E\left(e^{tX}\right)$$

$$= \frac{1}{b-a}\int_a^b e^{tx}\,dx$$

$$= \frac{e^{bt}-e^{at}}{t(b-a)^2}, \quad t \neq 0$$ ...(5.

We can find the higher order moments and various characteristics of the distribution using the *m.g.f.* (5.4)

Example 5.1: If X is uniformly distributed with mean 1 and variance 4/3, find $P(X < 0)$.

Solution: Let X is defined over $[a, b]$, then p.d.f is

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b,$$

We have, $$E(X) = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Thus, $$\frac{a+b}{2} = 1 \text{ and } \frac{(b-a)^2}{12} = \frac{4}{3}.$$

Solving for a and b and using the fact that $a < b$, we get $a = -1$ and $b = 3$. Therefore,

$$f(x) = \frac{1}{4}; \quad -1 \le x \le 3.$$

Hence, $$P(X < 0) = \int_{-1}^{0} f(x)dx = \frac{1}{4}[x]_{-1}^{0} = 1/4.$$

Example 5.2: The metro trains on a certain section run every 10 minutes between 5 a.m. to 10 p.m. What is the probability that a commuter entering the station at a random time during this period will have to wait at least five minutes?

Solution: Let X be the waiting time in minutes, then X is distributed uniformly over [0, 10] with p.d.f.

$$f(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10 \\ 0, & \text{otherwise} \end{cases}$$

The probability that the waiting time will be at least five minutes is

$$P(X \ge 5) = \int_5^{10} \frac{1}{10}dx = \frac{1}{2}.$$

5.3 NORMAL DISTRIBUTION

Normal distribution is the most important continuous probability distribution in the field of statistics since in applications many random variables are normal random variables or they are approximately normal particularly when the population size is large.

A continuous r.v. X with two parameters μ and σ having the p.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty \qquad ...(5.5)$$

is called the normal variate, and the distribution defined by (5.5) is called the normal distribution.

It defines a p.d.f., since

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz, \quad \left(z = \frac{x-\mu}{\sigma}\right)$$

$$= \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-\frac{1}{2}z^2} dz, \text{ integrand being an even function in } z,$$

$$= \frac{1}{\sqrt{\pi}} \int_0^\infty e^{-t} t^{-\frac{1}{2}} dt, \quad \left(t = \frac{1}{2}z^2\right)$$

$$= \frac{1}{\sqrt{\pi}} \Gamma(1/2) = 1, \text{ since } \Gamma(1/2) = \sqrt{\pi}.$$

The normal probability curve, $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ is a bell-shaped curve symmetrical a

the line $x = \mu$ and attains its maximum value of $1/\sigma\sqrt{2\pi} \approx 0.399/\sigma$ at $x = \mu$, as shown in Fig. 5

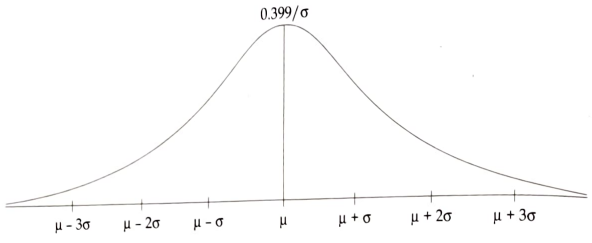

Fig. 5.1

The normal probability curve describes approximately many phenomena that occur in natu industry and research. Physical measurements such as meteorological experiments, rainf studies, error made in measuring a physical quantity are approximately normal in their behavio This distribution is often referred to as the *Gaussian distribution* also.

5.3.1 Constants of Normal Distribution

The *mean* is $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

$$= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz \left(z = \frac{x-\mu}{\sigma}\right) \quad ...(5$$

$$= \mu,$$

since $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = 1$, as proved already and $\int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz = 0$; intregrand being an od function in z.

The *variance* is $E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} dz, \quad \left(z = \frac{x-\mu}{\sigma}\right)$$

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

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left[-ze^{-\frac{1}{2}z^2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right] = \sigma^2 \quad ...(5.7)$$

Hence, the standard deviation (S.D.) is σ.

Next, the *mode* is that value of x for which $f(x)$ is maximum, that is, mode is the solution of $f'(x) = 0$ and $f''(x) < 0$.

For a normal distribution, the p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Taking logarithm both sides of this gives

$$ln\, f(x) = k - \frac{1}{2\sigma^2}(x-\mu)^2,$$

where $k = ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$ is a constant. Differentiating it w.r.t. x, we obtain

$$f'(x) = -\frac{1}{\sigma^2}(x-\mu)f(x), \text{ and } f''(x) = \frac{-f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$$

Now, $f'(x) = 0$ gives $x = \mu$, and at $x = \mu$, we obtain

$$f''(x) = -\frac{1}{\sigma^2}[f(x)]_{x=\mu} = -\frac{1}{\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} < 0.$$

Hence $x = \mu$ is the mode of the distribution.

The *median* is that value of x which divides the distribution in two equal parts.

Thus, for x to be median $\int_{-\infty}^{x} f(x)dx = \frac{1}{2}$. This gives

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

or, $$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx + \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}.$$

But, $$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{1}{2}z^2} dz = \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-\frac{1}{2}z^2} dz = \frac{1}{2}.$$

Hence, from (5.8), we have

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mu}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0,$$

which gives $x = \mu$ as the medium of the distribution. Hence, for a normal distribution

$$mean = mode = median$$

Thus, the normal distribution is symmetrical.

Moments about the mean: Odd order moments are given by

$$\mu_{2n+1} = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} (x-\mu)^{2n+1} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{\sigma^{2n+1}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^{2n+1} e^{-\frac{1}{2}z^2} dz, \quad z = \frac{x-\mu}{\sigma}$$

$= 0$, integrand being an odd function of

Hence, $\mu_{2n+1} = 0$, ...(5

Thus, in case of normal variate all odd order moments about the mean are zeros.

Even order moments are given by

$$\mu_{2n} = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} (x-\mu)^{2n} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{\sigma^{2n}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^{2n-1} e^{-\frac{1}{2}z^2} z\,dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}}\left[\left[-z^{2n-1}e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}(2n-1)z^{2n-2}e^{-\frac{1}{2}z^2} dz\right]$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}}(0-0) + (2n-1)\sigma^2\frac{\sigma^{2n-2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^{2n-2}e^{-\frac{1}{2}z^2} dz$$

$$= (2n-1)\sigma^2\mu_{2n-2}$$

Hence, $$\mu_{2n} = (2n-1)\sigma^2\mu_{2n-2} \quad ...(5.11)$$

It gives, $\mu_{2n} = (2n-1)(2n-3)\ldots 5.3.1.\sigma^{2n}$

In particular, $\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$ etc.

Hence, $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 \quad ...(5.12)$$

Thus, the normal probability curve is 'symmetric and mesokurtic'.

The *m.g.f. about origin* is

$$M_X(t) = E(e^{tX})$$

$$= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Put $\left(\frac{x-\mu}{\sigma}\right) = z$, we obtain

$$M_x(t) = e^{\mu t}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2 + tz} dz$$

$$= e^{\mu t + \frac{1}{2}t^2\sigma^2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2} dz$$

$$= e^{\mu t + \frac{1}{2}t^2\sigma^2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du, \quad u = z - t\sigma$$

$$= e^{\mu t + \frac{1}{2}t^2\sigma^2}, \text{ since } \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du = 1. \quad ...(5.13)$$

The m.g.f. about mean μ is

$$M_\mu(t) = E[e^{t(X-\mu)}]$$
$$= e^{-\mu t} E(e^{tX})$$
$$= e^{\frac{1}{2}t^2\sigma^2}, \text{ using (5.13)} \quad ...(5.14)$$

Expanding the r.h.s., we obtain

$$M_\mu(t) = 1 + \left(\frac{1}{2}t^2\sigma^2\right) + \frac{1}{2!}\left(\frac{1}{2}t^2\sigma^2\right)^2 + \ldots + \frac{1}{n!}\left(\frac{1}{2}t^2\sigma^2\right)^n + \ldots$$

This gives,

$$\mu_{2n+1} = \text{coeff. of } \frac{t^{2n+1}}{(2n+1)!} = 0$$

and,
$$\mu_{2n} = \text{coeff. of } \frac{t^{2n}}{(2n)!} = \frac{\left(\frac{1}{2}\sigma^2\right)^n (2n!)}{n!} = 1.3.5. \ldots (2n-1)\sigma^{2n}$$
as already obtained respectively in (5.10) and (5.11).

5.3.2 Sum of Two Independent Normal Variates

Let X and Y be two independent normal variates with parameter μ_1, σ_1^2 and μ_2, σ_2^2 respectively. Their *m.g.f*'s are given by

$$M_X(t) = \exp\left[\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right] \quad \text{and} \quad M_Y(t) = \exp\left[\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right]$$

Consider the *m.g.f.* of the random variable $X + Y$, we have

$$M_{X+Y}(t) = M_X(t).M_Y(t), \quad \text{since } X \text{ and } Y \text{ are independent}$$
$$= \exp\left[\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right] \exp\left[\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right]$$
$$= \exp\left[(\mu_1 + \mu_2)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2\right]$$

which is the *m.g.f.* of a normal variate with mean $(\mu_1 + \mu_2)$ and variance $\sigma_1^2 + \sigma_2^2$. ...(5.15)

Hence, the sum of the two independent normal variates is again a normal variate with mean and variance as the sum of the means and variances of the individual variates.

Remark: We note that in this case $X - Y$ is also a normal variate with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

5.3.3 Standard Normal Variate

If X is normally distributed with mean μ and variance σ^2, generally written as $X \sim N(\mu, \sigma^2)$ and if we define $Z = \dfrac{X-\mu}{\sigma}$, then

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = 0 \quad \text{and} \quad \text{Var}(Z) = E(Z-\bar{Z})^2 = \frac{1}{\sigma^2}E(X-\bar{X})^2 = 1.$$

The variable Z defined so is called *standard normal variate* and its p.d.f. is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty. \qquad \ldots(5.16)$$

Obviously, the mean and variance of the standard variable are respectively zero and one, and it is denoted by $Z \sim N(0, 1)$.

The distribution function of a standard normal variate is given by

$$\Phi(z) = P\{Z < z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} dz. \qquad \ldots(5.17)$$

5.3.4 Area Property of a Normal Probability Curve

The probability of a normal variate lying between two values x_1 and x_2 is given by the area under the normal curve from x_1 to x_2, that is

$$P(x_1 \le X \le x_2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz, \left(z = \frac{x-\mu}{\sigma}\right)$$
$$= \frac{1}{\sqrt{2\pi}}\left[\int_0^{z_2} e^{-\frac{1}{2}z^2} dz - \int_0^{z_1} e^{-\frac{1}{2}z^2} dz\right]$$
$$= P(z_2) - P(z_1),$$

where the definite integral $P(z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_0^z e^{-\frac{1}{2}z^2} dz$ is known as *normal probability integral* and gives the *area under standard normal curve between the ordinates at Z = 0 and Z = z*. These areas have been tabulated for different values of z at intervals of 0.01 and are given at *Table I* (see p. 341, Appendix 1).

In particular, $P(\mu - \sigma < X < \mu + \sigma) = \displaystyle\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx$ can be evaluated as

$$P(-1 < Z < 1) = \int_{-1}^{1} \phi(z)dz = \frac{2}{\sqrt{2\pi}} \int_0^1 e^{-\frac{1}{2}z^2} dz = 2 \times 0.3413 = 0.6826, \text{ from } \textit{Table I}$$

Similarly,

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2) = \frac{2}{\sqrt{2\pi}} \int_0^2 e^{-\frac{1}{2}z^2} dz = 2(0.4772) = 0.9544$$

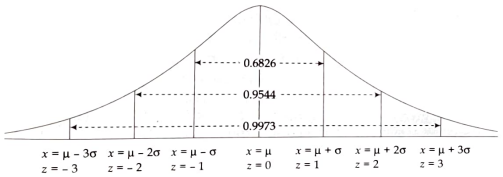

Fig. 5.2

and $P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3) = \frac{2}{\sqrt{2\pi}} \int_0^3 e^{-\frac{1}{2}z^2} dz = 2(0.49865) = 0.9973.$

Hence, the probability that a normal variate X lies outside the region $\mu \pm 3\sigma$ is given by

$$P(|x - \mu| > 3\sigma) = P(|z| > 3) = 1 - P(-3 \le z \le 3) = 1 - 0.9973 = 0.0027.$$

Thus, though theoretically normal variate ranges from $-\infty$ to ∞, yet in all probability we sho expect it to lie within the range $\mu \pm 3\sigma$, as shown in Fig. 5.2.

Remarks:

1. Since in *Table 1* we are given the areas under standard normal curve thus in numer problems we convert the variable in its standard form.
2. From the symmetry of the normal probability curve we have $P(Z > z) = P(Z < -z)$, as sho in Fig. 5.3.

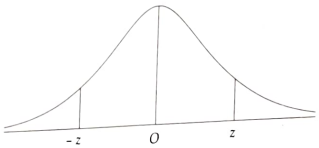

Fig. 5.3

3. From *Table I* we observe that $P(-1.96 < Z < 1.96) = 0.95$ and $P(-2.58 < Z < 2.58) = 0.99$, a these are two important area values to remember.

5.3.5 Chief Characteristics of the Normal Probability Distribution

The chief characteristics of the normal probability distribution with mean μ and variance σ^2 summarized as below:

1. The normal probability curve is bell shaped and symmetric about the line $x = \mu$; also t maximum probability occurs at $x = \mu$ given by $\frac{1}{\sigma\sqrt{2\pi}}$.
2. x-axis is an asymptote to the normal probability curve.
3. Mean, median and mode of the distribution coincide.
4. The points of inflexion of the curve are $x = \mu \pm \sigma$.
5. Mean deviation about the mean is $4\sigma/5$ (approx).
6. All odd order moments about mean are zero, that is, $\mu_{2n+1} = 0$, $n = 1, 2, \ldots$, and all even ord moments are given by $\mu_{2n} = (1)(3)(5) \ldots (2n-1)\sigma^{2n}$, $n = 1, 2, \ldots$ Hence $\beta_1 = 0$ and $\beta_2 =$ Thus the curve is symmetric and mesokurtic.
7. Sum and difference of two independent normal variates is again a normal variate.

8. Area property of the normal distribution is

$$P(\mu - \sigma < x < \mu + \sigma) = 0.6826,$$
$$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544,$$
$$P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973.$$

In addition to the properties mentioned above, most of the distributions occurring in practice tend to normal distribution for large samples. Because of all these characteristics normal distribution plays an important role in statistical theory and is applicable to various practical problems.

Example 5.3: For a normal distribution the first moment about 10 is 40 and the fourth moment about 50 is 48. Find the mean and standard deviation of the distribution

Solution: Let μ and σ be the mean and S.D. we have

$$\mu_1 = \mu - 10 = 40, \text{ which gives, } \mu = 50$$
$$\mu_4 = 48, \text{ gives } 3\sigma^4 = 48, \text{ that is, } \sigma = 2.$$

Example 5.4 Show that the mean deviation from the mean of a normal distribution is $4\sigma/5$ approximately.

Solution: The mean deviation from the mean μ, by definition is

$$E|x - \mu| = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x - \mu| e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-\frac{1}{2}z^2} dz, \quad z = \frac{x - \mu}{\sigma}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \left[-\int_{-\infty}^{0} z e^{-\frac{1}{2}z^2} dz + \int_{0}^{\infty} z e^{-\frac{1}{2}z^2} dz \right]$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_0^{\infty} z e^{-\frac{1}{2}z^2} dz = \sigma\sqrt{\frac{2}{\pi}} \left[-e^{-\frac{1}{2}z^2} \right]_0^{\infty}$$

$$= \sigma\sqrt{\frac{2}{\pi}} = 0.7979\,\sigma \simeq \frac{4}{5}\sigma$$

Example 5.5: If the amount of cosmic radiations to which a person is exposed while flying across a specific continent is a normal random variable with mean 4.35 units and S.D. 0.59 units. Find the probabilities that the amount of exposure during such a flight is

(a) between 4.00 and 5.00 units, (b) at least 5.50 units.

Solution: Let X be the amount of cosmic radiations exposed, we define

$$Z = \frac{X - 4.35}{0.59} \sim N(0, 1).$$

We have, $P(4 < X < 5) = P(-0.59 < Z < 1.10)$

$= P(1.10) + P(0.59)$

$= 0.3643 + 0.2224 = 0.5867$, from *Table I*.

The area is as shown in Fig. (5.4a).

(b) $P(X \geq 5.50) = P(Z \geq 1.95)$

$= 0.5 - P(1.95)$

$= 0.5 - 0.4744 = .0256$, from *Table I*.

The area is as shown in Fig. (5.4b).

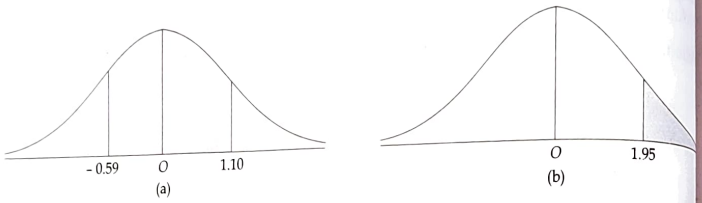

Fig. 5.4

Example 5.6: In a production of iron rods the diameter X can be approximated to be norma distributed with mean 2 inches and S.D. 0.008 inches.

(a) What percentage of defectives can we expect if we set the acceptance limits at $2 \pm$ inches?

(b) How should we set the acceptance limits to allow for 4% defectives?

Solution: (a) Define $Z = \frac{X - 2}{0.008} \sim N(0, 1)$. Then

$P(1.98 \leq X \leq 2.02) = P(-2.5 \leq Z \leq 2.5) = 2P(0 \leq Z \leq 2.5) = 2P(2.5)$

From Table *I*, $P(2.5) = 0.4938$, hence $P(1.98 \leq x \leq 2.02) = 0.9876$

This gives $P(|X - 2| > 0.02) = 1 - 0.9876 = .0124$.

Hence, the percentage of defectives expected is 1.24%.

(b) Let the acceptance limits be fixed at $2 \pm k$, then

$P(2 - k \leq X \leq 2 + k) = 0.96$

or, $P\left(\frac{-k}{0.008} < Z < \frac{k}{0.008}\right) = 0.96$, or $P\left(\frac{k}{0.008}\right) = 0.48$,

Again using *Table 1* it gives $\frac{k}{0.008} = 2.054$, or $k = .016432$.

The acceptance limits should be set at 2 ± 0.0164, that is, the interval [1.9836, 2.0164].

Example 5.7: A company has installed 10,000 electric lamps in a metro. If these lamps have an average life of 1,000 burning hours with a S.D. of 200 hours. Assuming normality, what number of lamps might be expected to fail.

(a) in the first 800 burning hours.

(b) between 800 and 1200 burning hours.

After what period of burning hours would you expect that

(c) 10% of the lamps would fail?

(d) 10% of the lamps would survive?

Solution: Let X be the life of a bulb in burning hours. Define $Z = \frac{X - 1000}{200}$, then $Z \sim N(0, 1)$.

(a) $P(X < 800) = P(Z < -1) = P(Z > 1) = 0.5 - P(0 < Z < 1)$

$= 0.5 - 0.3413 = 0.1587$, from Table *I*.

Therefore, out of 10,000 bulbs it is expected that 1587 will fail in the first 800 hours.

(b) $P(800 < X < 1200) = P(-1 < Z < 1) = 2P(0 < Z < 1) = 0.6826$.

Therefore, out of 10,000 bulbs it is expected that 6826 will burn between 800 and 1200 hours.

(c) If 10% of the bulbs fail after x_1 hours of burning, then x_1 be such that $P(X < x_1) = 0.10$. When $x = x_1$, then $z = (x_1 - 1000)/200 = -z_1$, say. Thus, we need to find z_1 such that

$$P(Z < -z_1) = 0.10, \text{ or } P(Z > z_1) = 0.10, \text{ or } P(0 < Z < z_1) = 0.40.$$

From *Table I*, $z_1 = 1.28$. Thus, $\frac{x_1 - 1000}{200} = -1.28$.

It gives, $x_1 = 1000 - 256 = 744$.

Therefore after 744 hours of burning life, 10% of the bulbs are likely to fail.

(d) If 10% of the bulbs are still burning after x_2 hours of burning, then x_2 be such that $P(X > x_2) = 0.10$. When $x = x_2$, then $z = (x_2 - 100)/200 = z_2$, say.

It gives, $P(Z > z_2) = 0.10$, or $P(0 < Z < z_2) = 0.40$.

From Table *I* $z_2 = 1.28$. Hence, $\frac{x_2 - 100}{200} = 1.28$, which gives, $x_2 = 1256$.

Thus, 10% of the bulbs are likely to burn after 1256 hours of the burning life.

5.3.6 Fitting of Normal Distribution

To fit normal distribution to the given data, we first calculate the mean μ and S.D. σ from the data. Then the normal probability curve to be fitted to the given data is given by

$$y = f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Example 5.8: Obtain the equation of the normal curve that may be fitted to the following data. Also obtain the expected normal frequencies.

Class	: 60–65	65–70	70–75	75–80	80–85	85–90	90–95	95–100
Frequency	: 3	21	150	335	326	135	26	4

Solution: First we calculate the mean and S.D. for the given data

Class	Frequency	Mid pt. (x)	$u = \frac{x - 77.5}{5}$	fu	fu^2
60 – 65	3	62.5	−3	−9	27
65 – 70	21	67.5	−2	−42	84
70 – 75	150	72.5	−1	−150	150
75 – 80	335	77.5	0	0	0
80 – 85	326	82.5	1	326	326
85 – 90	135	87.5	2	270	540
90 – 95	26	92.5	3	78	234
95 – 100	4	97.5	4	16	64
	1000			489	1425

We have, $\bar{u} = \frac{489}{1000} = 0.489, \quad \sigma_u^2 = \frac{1425}{1000} - (0.489)^2 = 1.425 - 0.239 = 1.186$

Thus, $\bar{x} = a + h\bar{u} = 77.5 + 5(0.489) = 77.5 + 2.445 = 79.945$

and, $\sigma_x = h\sigma_u = 5(1.089) = 5.445$

Hence, the equation of the normal curve to be fitted to the given data is given by

$$f(x) = \frac{1}{5.445\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-79.945}{5.445}\right)^2}.$$

To calculate the theoretical frequencies we calculate the area under this probability curve in interval (z_1, z_2) given by

$$\Delta z = \frac{1}{\sqrt{2\pi}} \int_0^{z_2} e^{-z^2/2} dz - \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-z^2/2} dz,$$

where $z = \frac{x - 79.945}{5.445}$.

Using *Table I*, we form the following table:

(x_1, x_2)	(z_1, z_2)	Area $\Delta z = P(z_2) - P(z_1)$	Expected Frequency = $N\Delta z$
$(-\infty, 60)$	$(-\infty, -3.663)$	0.00011	0.11 ≈ 0
(60, 65)	(−3.663, −2.745)	0.00291	2.91 ≈ 3
(65, 70)	(−2.745, −1.826)	0.03104	31.04 ≈ 31
(70, 75)	(−1.826, −0.908)	0.14787	147.87 ≈ 142
(75, 80)	(−0.908, −0.010)	0.32205	322.05 ≈ 322
(80, 85)	(−0.010, 0.928)	0.31930	319.30 ≈ 319
(85, 90)	(0.928, 1.487)	0.14407	144.07 ≈ 144
(90, 95)	(1.487, 2.675)	0.02979	29.79 ≈ 30
(95, 100)	(2.675, 3.683)	0.00273	2.73 ≈ 3
(100, ∞)	(3.683, ∞)	0.00011	0.11 ≈ 0
Total			1000

5.4 NORMAL DISTRIBUTION AS A LIMITING CASE OF BINOMIAL

Normal distribution can be derived as a limiting case of binomial distribution under the following conditions:

(i) n, the number of trials is sufficiently large

(ii) neither p nor q is very small.

We define a standard random variable z by

$$z = \frac{x - np}{\sqrt{npq}} \qquad \ldots(5.19)$$

where x is a binomial variate with mean np and standard deviation $\sqrt{npq}$. As x takes the value 0 to n, z takes the value $-np/\sqrt{npq}$ to $nq/\sqrt{npq}$ and these tend to $-\infty$ and $+\infty$ respectively under the conditions (i) and (ii) mentioned above. Also the increment in the value of z at each stage is $1/\sqrt{npq}$ and this tends to infinitesimally small as n tends to infinity, and let it be denoted by dz.

The probability function for the binomial variate is

$$p(x) = C_x^n p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}; \quad x = 0, 1, 2, \ldots.$$

We obtain the limiting form of $p(x)$ under the two conditions mentioned above. Using Stirling's approximation to $r!$ for large r, viz.

$$\lim_{r \to \infty} r! \simeq \sqrt{2\pi} e^{-r} r^{r+(1/2)}, \quad \text{we obtain}$$

$$\lim_{n \to \infty} p(x) = \lim \left[\frac{\sqrt{2\pi} e^{-n} n^{n+(1/2)} p^x q^{n-x}}{\sqrt{2\pi} e^{-x} x^{x+(1/2)} \sqrt{2\pi} e^{-(n-x)} (n-x)^{(n-x)+(1/2)}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \frac{(np)^{x+(1/2)} (nq)^{(n-x)+(1/2)}}{x^{x+(1/2)} (n-x)^{(n-x)+(1/2)}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}\sqrt{npq}} \left(\frac{np}{x} \right)^{x+(1/2)} \left(\frac{nq}{n-x} \right)^{n-x+(1/2)} \right]$$

$$= \frac{1}{\sqrt{2\pi}} \left(\lim \frac{1}{N} \right) dz,$$

where $N = \left(\frac{x}{np} \right)^{x+(1/2)} \left(\frac{n-x}{nq} \right)^{n-x+(1/2)}$

Using $x = np + z\sqrt{npq}$ from (5.19), we obtain

$$\ln N = \left(np + z\sqrt{npq} + \frac{1}{2} \right) \ln \left(1 + z\sqrt{\frac{q}{np}} \right) + \left(nq - z\sqrt{npq} + \frac{1}{2} \right) \ln \left(1 - z\sqrt{\frac{p}{nq}} \right)$$

$$= \left(np + z\sqrt{npq} + \frac{1}{2} \right) \left(z\sqrt{\frac{q}{np}} - \frac{z^2 q}{2np} + \ldots \right) + \left(nq - z\sqrt{npq} + \frac{1}{2} \right) \left(-z\sqrt{\frac{p}{nq}} - \frac{z^2 p}{2nq} - \ldots \right)$$

$$= \frac{z}{2\sqrt{n}} \left(\sqrt{\frac{q}{p}} - \sqrt{\frac{p}{q}} \right) + \frac{z^2}{2} - \frac{z^2}{4n} \left(\frac{q}{p} + \frac{p}{q} \right) + \text{terms with higher power of } \left(\frac{1}{n} \right)$$

Hence, when $n \to \infty$, $\ln N \to z^2/2$, that is, $N \to e^{z^2/2}$ or $\frac{1}{N} \to e^{-(1/2)z^2}$

Thus, from (5.20) the probability of z falling in the interval $\left(z - \frac{1}{2}dz, z + \frac{1}{2}dz \right)$ is

$$dP = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2} dz, \quad -\infty < z < \infty$$

Hence, the probability density function of continuous random variable z is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}, \quad -\infty < z < \infty$$

which is the p.d.f. of a standard normal variate, with mean zero and standard deviation one.

Example 5.9: A 20% of the memory chips made in a certain plant are defective. What are probabilities that in a lot of 100 randomly chosen for inspection

(a) at most 15 will be defective?

(b) exactly 15 will be defective?

Solution: Here mean $\mu = 100(0.20) = 20$ and S.D. $\sigma = \sqrt{100(0.20)(0.80)} = 4$, thus the binomial variable may be approximated by $X \sim N(20, 16)$, a normal variate with mean 20 and variance 16.

Since the variable under consideration is discrete, to spread its values over a continuous scale we represent each value k by the interval $(k - ½, k + ½)$. Thus, 15 is represented as 14.5 – 15.5.

(a) $P(X < 15.5) = P(Z < -1.13) = P(Z > 1.13)$

$= 0.5 - P(0 < Z < 1.13) = 0.5 - 0.3708 = 0.1292$, using Table *I*

(b) $P(14.5 < X < 15.5) = P(-1.38 < Z < -1.13) = P(1.13 < Z < 1.38)$

$= 0.4162 - 0.3708 = .0454$, using *Table I*.

Remark: In case calculations are worked out using binomial distribution, for part (a) we arrive at 0.1285 and for part (b) 0.0481, both in close with the respective approximations obtained. In general, the normal approximation to the binomial distribution is advisable when both np and $n(1 - p)$ are greater than 15.

Example 5.10: A multiple-choice quiz has 200 questions each with 4 possible answers of which only 1 is the correct answer. What is the probability that sheer guess-work yields from 25 to 30 correct answer for 80 of the 200 problems about which the student has no knowledge?

Solution: We have, $n = 80$, $p = \frac{1}{4}$, thus

$$\mu = np = 20 \quad \text{and} \quad \sigma = \sqrt{npq} = \sqrt{80 \times \frac{1}{4} \times \frac{3}{4}} = 3.873.$$

Let X be the normal approximation to the underlying binomial variate, then $X \sim N(20, 15)$, a normal variate with mean 20 and variance 15.

To spread the variable over the continuous scale we find the probability that X lies in the interval 24.5 – 30.5. We define $z = \frac{x - 20}{3.873}$. Thus,

$$P(24.5 < x < 30.5) = P(1.16 < z < 2.71) = 0.4966 - 0.3770 = 0.1196.$$

5.5 LOG-NORMAL DISTRIBUTION

A positive random variable X is said to have a log-normal distribution if $\ln X$ is normally distributed. The p.d.f. is given by

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\beta x} e^{-(\ln x - \alpha)^2/2\beta^2}, & x > 0, \beta > 0 \\ 0, & \text{otherwise} \end{cases} \quad \ldots(5.21)$$

where α and β^2 are the mean and variance of $\ln x$, the natural logarithm of x.

This distribution is positively skewed and arises in problems of economics, biology and risk-analysis.

5.5.1 Constants of Log-Normal Distribution

Let $Y = lnX$. This gives $X = e^Y$.

The r th moment about origin is.

$$\mu'_r = E(X^r) = E(e^{rY}) = M_Y(r); \quad \text{the } m.g.f. \text{ of } Y, r \text{ being the parameter}$$

$$= \exp\left[\alpha r + \frac{1}{2}\beta^2 r^2\right] \qquad ...(5.2$$

The *mean* is

$$\mu = E(X) = e^{\alpha + \frac{1}{2}\beta^2} \qquad ...(5.2$$

$$E(X^2) = e^{2\alpha + 2\beta^2}$$

The *variance* is

$$\sigma^2 = E(X^2) - [E(X)]^2$$

$$= e^{2\alpha + 2\beta^2} - e^{2\alpha + \beta^2}$$

$$= e^{2\alpha + \beta^2}\left(e^{\beta^2} - 1\right), \qquad ...(5.2$$

where α and β^2 are respectively the mean and variance of lnX.

Example 5.11: In a nuclear power plant, engineers model the strength 's' of steam generator supports in terms of their ability to withstand the peak acceleration caused by earthquakes. Experts suggest than $ln(s)$ is normally distributed with mean 4.0 and variance 0.09. Find the probability that supports will survive a peak acceleration of 33 units. Also find the mean and s.d. of s.

Solution: Here $ln(s) \sim N(4, 0.09)$.

The p.d.f. for the strength s is given by

$$f(s) = \frac{1}{\sqrt{2\pi}\beta s} e^{-(\ln(s) - \alpha)^2/2\beta^2}, \quad s > 0$$

where $\alpha = 4$, $\beta^2 = 0.09$.

Thus, $P(s > 33) = 1 - P(s \le 33)$

$$= 1 - \frac{1}{\sqrt{2\pi}\beta}\int_0^{33} e^{-\frac{1}{2}\left(\frac{(\ln(s) - \alpha}{\beta}\right)^2} \frac{ds}{s}$$

Put $t = ln(s)$, we obtain

$$P(s > 33) = 1 - \frac{1}{\sqrt{2\pi}\beta}\int_{-\infty}^{\ln 33} e^{-\frac{1}{2}\left(\frac{t - \alpha}{\beta}\right)^2} dt$$

$$= 1 - F\left(\frac{\ln(33) - \alpha}{\beta}\right)$$

$$= 1 - F\left(\frac{\ln 33 - 4.0}{0.30}\right)$$

$$= 1 - F(-1.68)$$

$$= 1 - 0.465 = 0.9535, \text{ using } \textit{Table 1}$$

Mean of s, $\mu = e^{\alpha + \frac{1}{2}\beta^2}$

$$= e^{4 + 0.045} = e^{4.045} = 57.11$$

Variance of s, $\sigma^2 = e^{2\alpha + \beta^2}\left[e^{\beta^2} - 1\right]$

$$= e^{8.09}[e^{0.09} - 1] = 307.17$$

Hence, S.D. $\sigma = 17.53$.

5.6 EXPONENTIAL DISTRIBUTION

A continuous random variable X with probability density function f(x) defined by

$$f(x) = \begin{cases} ae^{-ax}, & \text{if } x \ge 0 \\ 0, & \text{if } x < 0, \end{cases} \qquad ...(5.25)$$

for some a > 0, is called an exponential variate with parameter a and the distribution is said to be exponential distribution.

The function (5.25) defines a probability density function, since

$$\int_{-\infty}^{\infty} f(x)dx = a\int_0^{\infty} e^{-ax}dx = [-e^{-ax}]_0^{\infty} = 1.$$

The *distribution function F(x) of an exponential variate* is given by

$$F(x) = P(X \le x)$$

$$= \int_0^x ae^{-ax}dx = (1 - e^{-ax}), \quad x \ge 0. \qquad ...(5.26)$$

The exponential distribution arises as the distribution of amount of time until some specific event occurs. For example the amount of time starting from now a car comes for service at a service station, a new war breaks out, a patient comes at an emergency reception, etc. are all random variables that behave exponentially.

5.6.1 Constants of Exponential Distribution

The *mean* is $E(X) = \int_0^{\infty} x f(x)$

$$= a\int_0^\infty x\,e^{-ax}\,dx = a\left[-x\frac{e^{-ax}}{a} - \frac{e^{-ax}}{a^2}\right]_0^\infty = \frac{1}{a} \quad \ldots(5.2$$

Similarly, $$E(X^2) = a\int_0^\infty x^2 e^{-ax}\,dx = \frac{2}{a^2}$$

The *variance* is $$\sigma^2 = E(X^2) - [E(X)]^2$$

$$= \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2} \quad \ldots(5.2$$

Thus, the mean is reciprocal of the parameter a and the variance is equal to the square of th mean in case of an exponential variate, and hence for the exponential distribution $\sigma^2 \geqslant \mu$ for different values of the parameter a.

The *m.g.f. about origin* is

$$M_X(t) = E\left(e^{tX}\right)$$

$$= a\int_0^\infty e^{-ax}e^{tx}\,dx$$

$$= a\int_0^\infty e^{-(a-t)x}\,dx$$

$$= \frac{a}{a-t} = \left(1-\frac{t}{a}\right)^{-1} = \sum_{r=0}^{\infty}\left(\frac{t}{a}\right)^r,\ t < a \quad \ldots(5.29$$

Thus, $$\mu'_r = E(X^r)$$

$$= \text{coefficient of } \frac{t^r}{r!} \text{ in } M_x(t)$$

$$= \frac{r!}{a^r};\quad r = 1, 2, \ldots \quad \ldots(5.30$$

5.6.2 Lack of Memory

An important property of the exponential distribution is that it *lacks memory*, that is, if X has an exponential distribution, then

$$P(X > s + t \mid X > t) = P(X > s) \quad \ldots(5.31$$

for all $s, t > 0$.

Now, (5.31) can be written as

$$\frac{P(X > s + t \text{ and } X > t)}{P(X > t)} = P(x > s)x$$

or, $$P(X > s + t) = P(X > s)P(X > t),$$

which is satisfied when X has exponential distribution (5.25).

In case we interpret x as the lifetime of some equipment in hours, then (5.31) simply means that *the probability that the equipment survives for at least (s + t) hours given that it has survived t hours is the same as the initial probability that it survives for at least s hours.*

The memoryless property is further explained by the *failure rate function* (or, *hazard rate function*) of the exponential distribution.

Let X be a continuous random variable with distribution function F and density function f. The *failure rate function* $r(t)$ is defined by

$$r(t) = f(t)/\bar{F}(t) \quad \ldots(5.32)$$

The failure rate represents the conditional probability density that a t year old item will fail. If lifetime distribution is exponential, with parameter λ, then

$$f(t) = \lambda e^{-\lambda t} \quad \text{and} \quad \bar{F}(t) = 1 - F(t) = e^{-\lambda t}$$

and thus

$$r(t) = f(t)/\bar{F}(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda. \quad \ldots(5.33)$$

Thus, the failure rate in case of exponential distribution is constant, that is, a process with lifetime distribution as a memoryless random variable has a constant failure rate.

Example 5.12: A system contains a certain type of component whose lifetime X is exponentially distributed with mean of 5 years. If 8 such components are installed in different systems, then what is the probability that at least 3 are still working at the end of 7 years?

Solution: The p.d.f for the r.v. X is given by

$$f(x) = \frac{1}{5}e^{-x/5},\quad x \geq 0.$$

Thus, $$P(X > 7) = \frac{1}{5}\int_7^\infty e^{-x/5}dx = e^{-7/5} = 0.1827$$

If n represents the number of components out of 8 working after 7 years of instalment, then

$$P(n \geq 3) = \sum_{n=3}^{8} C_n^8 (0.1827)^n (0.8173)^{8-n}$$

$$= 1 - [C_0^8(0.8173)^8 + C_1^8(0.1827)(0.8173)^7 + C_2^8(0.1827)^2(0.8173)^6]$$

$$= 1 - [0.1991 + 0.3560 + 0.2786] = 0.1663.$$

Example 5.13: If on the average three trucks arrive per hour to be unloaded at a warehouse, using exponential distribution find the probabilities that the time between the arrival of successive trucks will be, (a) less than 5 minutes, (b) at least 45 minutes.

Solution: Let the r.v. t denote the time in hrs between arrival of successive trucks then its p.d.f is

$$f(t) = 3e^{-3t},\ 0 \leq t < \infty$$

(a) $$P(0 < t < 1/12) = \int_0^{1/12} 3e^{-3t}dt = 1 - e^{-1/4} = 0.221.$$

(b) $$P\left(\frac{3}{4} < t < \infty\right) = \int_{3/4}^{\infty} 3e^{-3t}dt = e^{-9/4} = 0.105.$$

Example 5.14: If X_i, $i = 1, 2, ..., n$ are independent exponential random variables with parameters λ_i, $i = 1, 2, ..., n$, then the smallest of them is also exponential with parameter $\sum_{i=1}^{n}\lambda_i$.

Solution: Let $X = \min\{X_1, X_2, ..., X_n\}$. Then

$$P\{\min X_i > x\} = P\{X_i > x,\ i = 1, 2, ..., n\}$$

$$= \prod_{i=1}^{n} P\{X_i > x\}, \quad \text{since } X_i\text{'s are independent}$$

$$= \prod_{i=1}^{n} e^{-\lambda_i x} = e^{-\left(\sum_{i=1}^{n}\lambda_i\right)x}$$

Thus, the distribution function $F(x)$ of X is given by

$$F(x) = P\{X < x\} = 1 - e^{-(\Sigma\lambda_i)x}, \quad x \geq 0.$$

Hence the p.d.f. $f(x)$ of X is

$$f(x) = \frac{d}{dx}F(x) = (\Sigma\lambda_i)e^{-(\Sigma\lambda_i)x}$$

Thus X is distributed like an exponential variate with parameter $\Sigma\lambda_i$.

5.7 WEIBULL DISTRIBUTION

Weibull distribution is closely related with exponential distribution, with probability density function given by

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & \text{for } x > 0, \alpha > 0, \beta > 0 \\ 0, & \text{elsewhere} \end{cases} \quad ...(5.34)$$

α, β are called parameters of the distribution. The function (5.34) defines a p.d.f. since

$$\int_0^{\infty} f(x)\,dx = \int_0^{\infty} \alpha\beta\, x^{\beta-1}\, e^{-\alpha x^{\beta}}\, dx$$

$$= \int_0^{\infty} \alpha e^{-\alpha y}\, dy; \quad \text{taking } y = x^{\beta}.$$

$$= 1$$

For $\beta = 1$, Weibull distribution (5.34) becomes exponential distribution (5.25) with parameter α. The graphs of Weibull density function $y = f(x)$ for $\alpha = 1$ and $\beta = 0.5, 1, 2$ and 3 are shown in Fig. 5.5.

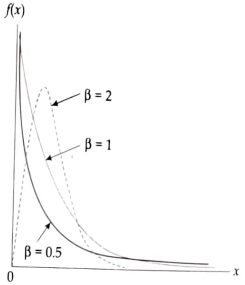

Fig. 5.5

5.7.1 Constants of Weibull Distribution

The r th moment about origin is given by

$$\mu'_r = E(X^r) = \int_0^{\infty} x^r\, \alpha\beta\, x^{\beta-1}\, e^{-\alpha x^{\beta}}\, dx$$

$$= \alpha^{-r/\beta} \int_0^{\infty} u^{r/\beta}\, e^{-u}\, du; \quad u = \alpha x^{\beta}$$

$$= \alpha^{-r/\beta}\,\Gamma\left(1 + \frac{r}{\beta}\right), \quad \text{since } \Gamma(l) = \int_0^{\infty} u^{l-1}\, e^{-u}\, du \quad ...(5.35)$$

For $r = 1$, the *mean* is $E(X) = \alpha^{-1/\beta}\,\Gamma\left(1 + \frac{1}{\beta}\right)$ ...(5.36)

For $r = 2$, $E(X^2) = \alpha^{-2/\beta}\,\Gamma\left(1 + \frac{2}{\beta}\right)$

Hence, the variance is, $\sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right\}$ ...(5.37)

Like exponential distribution, the Weibull distribution is also applicable to reliability and life-testing problems such as the time to failure or life length of a component measured from some specified time until it fails. In case of exponential life distribution the hazard rate λ is constant over

time while in case of Weibull distribution hazard rate either increases or decreases over t
depending upon its parameter β.

Example 5.15: Suppose that the lifetime (in hours) of an electronic tube is a random variable having Weibull distribution with $\alpha = 0.05$ and $\beta = 0.5$. Find

(a) the mean lifetime of these tubes

(b) the S.D. of the lifetime

(c) the probability that such a tube will last more than 1000 hrs.

Solution: (a) We have, $\alpha = 0.05$, $\beta = 0.5$.

Mean, $$\mu = \alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$$

$$= (0.05)^{-2}\,\Gamma(3) = (20)^2\,2! = 800 \text{ hrs.}$$

Variance, $$\sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}$$

$$= (0.05)^{-4}\,[\Gamma(5) - \{\Gamma(3)\}^2]$$

$$= (20)^4\,[4! - (2!)^2]$$

$$= 160000 \times 20 = 32 \times 10^5$$

Hence, the S.D. is $\sigma = 1789$ hrs. (approx.)

(c) $$P(X > 1000) = \int_{1000}^{\infty} (0.05)(0.5)\, x^{-0.5}\, e^{-0.05\, x^{0.5}}\, dx$$

$$= \int_{\sqrt{1000}}^{\infty} 0.05\, e^{-0.05y}\, dy, \qquad y = x^{0.5}$$

$$= \left(-e^{-0.05y}\right)_{\sqrt{1000}}^{\infty} = e^{-0.05\sqrt{1000}}$$

$$= 0.2057.$$

5.8 GAMMA DISTRIBUTION

A continuous random variable X with probability density function, for some $\alpha > 0$, $\beta > 0$, defined by

$$f(x) = \begin{cases} \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \ldots(5.3$$

is called gamma variate with parameters (α, β) and the distribution is called gamma distribution, whe

$\Gamma(\alpha)$ is the value of the gamma function with parameter $\alpha > 0$, given by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx$$

We have $\Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1)$ and when α is positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$

The function (5.38) defines a p.d.f., since

$$\int_{-\infty}^{\infty} f(x)dx = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-\beta x} dx = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} t^{\alpha-1} e^{-t} dt = 1, \quad t = \beta x.$$

We note that exponential p.d.f defined by (5.25) is a special case of gamma p.d.f (5.38) for $\alpha = 1$.

The relationship between the gamma and the exponential distribution allows the gamma function to find applications similar to that of exponential in particular in the field of queuing theory and reliability problems. In addition to this, gamma distribution is frequently used in *life-testing, the waiting time until death probability models* etc.

Taking $\beta = 1$ in (5.38), the p.d.f. given by

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, & x \geq 0,\ \alpha > 0 \\ 0, & \text{otherwise} \end{cases} \quad \ldots(5.39)$$

is defined as the p.d.f of the gamma variate x with *parameter* α.

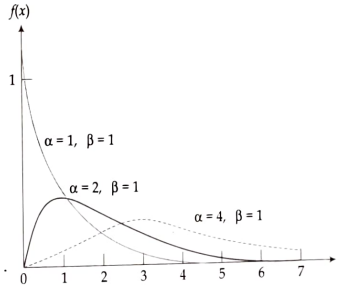

Fig. 5.6

The graphs of the gamma p.d.f for $\beta = 1$ and $\alpha = 1, 2$ and 4 are shown in Fig. 5.6.

5.8.1 Constants of Gamma Distribution

The *m.g.f. about origin is*

$$M_X(t) = E(e^{tX})$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x}\, dx$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-t)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y}\, dy\,; \quad y = (\beta - t)x$$

$$= \left(\frac{\beta}{\beta - t}\right)^\alpha, \quad t < \beta$$

$$= \left(1 - \frac{t}{\beta}\right)^{-\alpha} \qquad \ldots(5.40)$$

$$= 1 + \frac{\alpha t}{\beta} + \frac{\alpha(\alpha+1)}{2!}\frac{t^2}{\beta^2} + \ldots + \frac{\alpha(\alpha+1)\ldots(\alpha+r-1)}{r!}\frac{t^r}{\beta^r} + \ldots$$

Hence, μ'_r = coefficient of $\frac{t^r}{r!}$

$$= \frac{\alpha(\alpha+1)\ldots(\alpha+r-1)}{\beta^r}$$

Thus the *mean* is $\mu'_1 = \frac{\alpha}{\beta}$ ...(5.41)

Also, $\mu'_2 = \frac{\alpha(\alpha+1)}{\beta^2}$

The *variance* is $\sigma^2 = \mu'_2 - (\mu'_1)^2 = \frac{\alpha(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}$ ...(5.42)

Remarks:

(1) The *m.g.f.* of the gamma distribution defined by (5.39) is given by

$$M_X(t) = (1 - t)^{-\alpha} \qquad \ldots(5.43)$$

as obtained from (5.40) for $\beta = 1$. The mean and variance are respectively, givenby

$$\mu = \alpha \quad \text{and} \quad \sigma^2 = \alpha.$$

(2) Using (5.43), we can easily prove that *the sum of two independent gamma variates with parameters α_1 and α_2 is again a gamma variate with parameter $\alpha_1 + \alpha_2$.*

Example 5.16: The daily consumption of milk in a city, in excess of 20,000 litres, is approximately distributed as a gamma variate with parameters $\alpha = 2$ and $\beta = 1/10{,}000$. The city has a daily stock of 30,000 litres. What is the probability that the stock is insufficient on a particular day?

Solution: If X denotes the daily consumption in excess of 20,000 litres, then p.d.f of X is

$$f(x) = \frac{1}{(10{,}000)^2\,\Gamma(2)} x^{2-1} e^{-x/10{,}000}, \quad x > 0.$$

The stock of 30,000 litres will be insufficient on a particular day, if the excess consumption is more than 10,000 litres.

Thus, $$P(X > 10{,}000) = \int_{10{,}000}^\infty \frac{x e^{-x/10{,}000}}{(10{,}000)^2} dx$$

$$= \int_1^\infty t e^{-t} dt, \quad t = x/10{,}000$$

$$= [-te^{-t} - e^{-t}]_1^\infty$$

$$= 2/e = 0.736$$

5.9 BETA DISTRIBUTION

A continuous random variable X with probability density function, for some $\alpha > 0$, $\beta > 0$, defined by

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \qquad \ldots(5.44)$$

is called beta variable with parameters α and β and the distribution is called beta distribution, sometimes beta distribution of the first kind, where $B(\alpha, \beta)$ is the value of the beta function given by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

The relation between beta and gamma function is $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. Thus, for positive integral values of α and β, we have $B(\alpha, \beta) = (\alpha-1)!\,(\beta-1)!/(\alpha+\beta-1)!$.

Obviously, the function defined by (5.44) defines a p.d.f., since

$$\int_{-\infty}^\infty f(x)dx = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

$$= \frac{B(\alpha,\beta)}{B(\alpha,\beta)} = 1.$$

The graph of beta p.d.f (5.44) with $\alpha = 3$ and $\beta = 2$, is shown in Fig. 5.7.

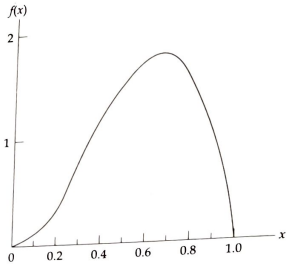

Fig. 5.7

5.9.1 Constants of Beta Distribution

The *r th moment about origin* is given by

$$\mu'_r = E(X^r) = \frac{1}{B(\alpha, \beta)} \int_0^1 x^{\alpha + r - 1}(1 - x)^{\beta - 1}\, dx$$

$$= \frac{B(\alpha + r, \beta)}{B(\alpha, \beta)}$$

$$= \frac{\Gamma(\alpha + r)\,\Gamma(\alpha + \beta)}{\Gamma(\alpha + r + \beta)\,\Gamma(\alpha)} \qquad ...(5.45)$$

The *mean* is

$$\mu'_1 = \frac{\Gamma(\alpha + 1)\,\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + 1)\,\Gamma\alpha} = \frac{\alpha}{\alpha + \beta} \qquad ...(5.46)$$

$$\mu'_2 = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}$$

The *variance* is $\mu_2 = \mu'_2 - (\mu'_1)^2$

$$= \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)} - \left(\frac{\alpha}{\alpha + \beta}\right)^2$$

$$= \frac{\alpha\beta}{(\alpha + \beta)^2\,(\alpha + \beta + 1)} \qquad ...(5.47)$$

Remark: There is another form of beta distribution called *beta distribution of second kind* with p.d.f. given by

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} \dfrac{x^{\alpha - 1}}{(1 + x)^{\alpha + \beta}}, & 0 < x < \infty,\ \alpha, \beta > 0 \\ 0, & \text{otherwise} \end{cases} \qquad ...(5.48)$$

This distribution transformed to beta distribution of first kind (5.44) by the transformation $1 + x = \frac{1}{y}$. Further we can easily find that in case of distribution defined by (5.48)

$$\text{mean} = \frac{\alpha}{(\beta - 1)},\ \beta > 1, \quad \text{and} \quad \text{variance} = \frac{\alpha(\alpha - \beta - 1)}{(\beta - 1)^2(\beta - 2)},\ \beta > 2.$$

Example 5.17: If the proportion of a brand of television set requiring service during the first year of operation is a random variable having a beta distribution with $\alpha = 3$ and $\beta = 2$, what is the probability that at least 80% of the new models sold this year of this brand will require service during the first year of operation?

Solution: If the r.v. X denotes the proportion of T.V. sets requiring service during the first year of operation, then its p.d.f. is

$$f(x) = \frac{1}{B(3, 2)} x^2(1 - x),\ 0 < x < 1.$$

Thus, $$P(x > 0.8) = \frac{1}{B(3, 2)} \int_{0.8}^1 x^2(1 - x)dx = 12\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_{0.8}^1 = 0.1808$$

REVIEW EXERCISES

1. Define uniform random variable of continuous type. Mention a specific situation under which this distribution is used. Find its *m.g.f.* and hence derive its mean and variance.
2. Define a normal variate. Show that in case of normal distribution mean, median and mode are equal.
3. Define a standard normal variate. What is its p.d.f.? Find its mean and variance.
4. Find the *m.g.f.* of normal distribution about its mean. Hence find its four central moments and the coefficients of skewness and kurtosis for the normal curve.
5. Show that in case of normal distribution

$$\mu_{2n+1} = 0 \quad \text{and} \quad \mu_{2n} = \frac{2n!}{2^n.n!}\sigma^{2n},\quad n = 0, 1, 2, \ldots.$$

6. Show that in case of normal distribution mean deviation from the mean is $4\sigma/5$ (approx).
7. Prove that sum and difference of two independent normal variate is again a normal variate.
8. Discuss the chief characteristics of a normal probability curve.
9. Discuss the area property of a normal probability curve.

10. Show that for the normal probability curve
    (a) the maximum probability occurs at the mean of the distribution.
    (b) the points of inflexion occur at a distance of ± σ from the mean, σ being the S.D.
11. Show that normal distribution is a limiting case of binomial.
12. Is there any limiting relation between the Poisson and the normal distribution? investigate the same?
13. Define a log-normal variate. What is its p.d.f.? Find its mean and variance. Mentio specific situation in which a log-normal variate is applicable.
14. If X is normal distributed with mean zero and variance σ^2, then find the p.d.f of $Y = e^X$
15. Define exponential distribution. Find its moment generating function and hence find mean and variance. What are the specific situations in which exponential variate applicable.
16. Show that exponential distribution lacks memory.
17. If X_i, $i = 1, 2, ..., n$, are n independent exponential random variables with parameter λ_i, $i = 1, 2, ..., n$, then find the distribution of $\min_i \{X_i\}$.
18. Define Weibull distribution. Find its mean and variance. Give a few of its spec applications. How is it related to exponential distribution?
19. Show that in case of exponential life distribution the failure rate is constant over time w it is not so in case of Weibull distribution.
20. Define a gamma variate, with parameter α. Find its *m.g.f.* Show that sum of independent gamma variates with parameters α_1 and α_2 is again a gamma variate parameter $\alpha_1 + \alpha_2$.
21. Define gamma variate with parameters (α, β). Find its mean and variance.
22. Define Beta variate of first kind with parameters (α, β). Find its rth moment about ori Hence find its mean and variance.
23. Define Beta variate of second kind. How is it transformed to Beta variate of first kind?

PROBLEM SET

1. A conference room can be reserved for no more than five hours. Assuming that duratio of a conference has a uniform distribution over the interval [0, 5]. What is the p.d.f.? W is the probability that any given conference lasts at most 4 hours?
2. The daily amount of coffee in litres dispensed by a machine at a plaza is uniform distributed between 7 litres and 10 litres. Find the probability that on a given day amount of coffee dispensed by this machine will be
   (a) at most 8.8 litres,
   (b) more than 7.4 litres but less than 9.5 litres,
   (c) at least 8.5 litres.
3. If X is normally distribution with mean 18 and S.D. 2.5, find
   (a) $P(X < 15)$, (b) $P(17 < X < 21)$
   (c) the value of k such that $P(X < k) = 0.2236$,
   (d) the value of k such that $P(X > k) = 0.1814$.

4. The actual amount of instant coffee that a filling machine put into '4-ounce' jar may be approximated as a normal random variable with S.D. 0.04 ounces. If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars?
5. If in a normal distribution 31% of the items are under 45 and 8% are over 64. Find the mean and S.D. of the distribution.
6. In a test on 200 electric bulbs, it was found that the life of a particular make was normally distributed with mean 2040 hours and S.D. 60 hours. Estimate the number of bulbs likely to burn for
   (a) more than 2150 hours,
   (b) less than 1950 hours, and
   (c) more than 1920 hours but less than 2160 hours.
7. The average life of an inverter is 10 years with a S.D. of 2 years. The manufacturer replaces free all inverters that fail while under guarantee. If he is willing to replace only 3% of the inverters that fail, how long a guarantee should he offer, assuming that the lifetime follows a normal distribution.
8. If the lifetime of a certain kind of automobile battery is normally distributed with a mean of 5 years and a S.D. of 1 year, and the manufacturer wishes to guarantee the battery for 4 years, what percentage of the batteries will he have to replace under the guarantee?
9. If the mathematics score of an entrance exam are normally distributed with mean 480 and S.D. 100 and if an institution sets 500 as the minimum score for new students, what per cent of students would not reach that score?
10. Cerebral blood flow (CBF) in the brains of healthy people is normally distributed with a mean of 74 and a S.D. of 16.
    (a) What percentage of healthy people will have CBF readings between 60 and 80?
    (b) If a person has a CBF reading below 40 he is classified 'at risk' for a stroke. What proportion of healthy people will mistakenly be diagnosed as 'at risk'?
11. Suppose that the amount of money spent by shoppers at a mall between 4 p.m. to 6 p.m. on Sunday is normally distributed with mean of Rs. 4250 and a S.D. of Rs. 500. A shopper is randomly selected on a Sunday between 4-6 p.m. and asked about his spending pattern
    (a) What is the probability that he has spent more than Rs. 4500 at the mall?
    (c) What is the probability that he has spent between Rs. 4500 and Rs. 5000 at the mall?
    (d) If two shoppers are randomly selected, what is the probability that both have spent more than Rs. 5000 at the mall?
12. Fit a normal distribution to the following data and calculate the expected frequencies

Class	1-3	3-5	5-7	7-9	9-11
Frequency	1	4	6	4	1.

13. The following table gives baseball throw distances by 303 first year students of a college
    (a) Fit a normal distribution and find the theoretical frequencies.
    (b) Find the expected number of students throwing baseballs at a distance exceeding 105 feet on the basis that the data fits a normal distribution.

Distance in feet	Number of students
15 – 25	1
25 – 35	2
35 – 45	7
45 – 55	25
55 – 65	33
65 – 75	53
75 – 85	64
85 – 95	44
95 – 105	31
105 – 115	27
115 – 125	11
125 – 135	4
135 – 145	1

14. A sample of 100 items is taken from a batch known to contain 40% defectives. Using norm approximations find the probability that sample contains
    (a) at least 44 defectives, (b) exactly 44 defectives.
15. A certain drug is effective in 72% of cases. Given 2,000 patients are treated with drug w is the probability that it will be effective for, (a) at least 1,400 patients, (b) less than 1, patients, (c) exactly 1,420 patients.
16. A process for manufacturing an electronic component is 1% defective. A quality cont plan is to select 100 items from the process and if none is defective the process continu Using normal approximation, find the probability that the process continues
    (a) for the sample plan described
    (b) even if the process has gone bad to produce 5% defective.
17. If C is the concentration of a certain pollutant, in parts per million, then it is observed $\ln C$ is normally distributed with mean 3.2 and variance 1. What is the probability t concentration exceeds 8 parts per million?
18. The logarithm of the ratio of the output to the input current, that is $\ln(I_o/I_i)$, is norma distributed with mean 2 and variance 0.01. Find (a) the mean and the variance of the r (I_o/I_i). (b) the probability that I_o/I_i will be between 6.1 and 8.2.
19. The amount of time that a surveillance camera will run without having to be rese random variable having exponential distribution with an average of 60 days. Find probability that such a camera will have to be reset, (a) in less than 60 days, (b) at leas days.
20. The length of time for one individual to be served at a canteen is a random variable ha an exponential distribution with mean of 4 minutes. What is the probability that a perso served in less than 3 minutes on at least 4 of the next 6 visits?
21. A continuous r.v X has the p.d.f. $f(x) = \begin{cases} Ae^{-x/5}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$

    Find the value of A and show that for any two positive numbers s and t,

    $$P[X > s + t \mid X > s] = P[X < t].$$

22. Following data gives the burning hours of 200 bulbs. Calculate the theoretical frequencies on the basis that burning hours of a bulb is exponentially distributed random variable.

Burning hrs	0-20	20-40	40-60	60-80	80-100
Number of bulbs	104	56	24	12	4

23. The response time of a certain computer system in seconds has an exponential distribution with a mean of 3 seconds.
    (a) What is the probability that response time exceeds 5 seconds?
    (b) What is the probability that response time is between 5-10 seconds?
24. Show that the failure rate function of the Weibull distribution with p.d.f. given by $f(t) = \alpha\beta\, t^{\beta-1} e^{-\alpha t^{\beta}}$; $t > 0$, $\alpha, \beta > 0$ is $\lambda(t) = \alpha\beta\, t^{\beta-1}$, $t > 0$.
25. The lifetime of a certain kind of an emergency backup battery (in hours) is a random variable X having the Weibull distribution with $\alpha = 0.1$ and $\beta = 0.5$. Find
    (a) the mean lifetime of these batteries.
    (b) the probability that such a battery will last more than 300 hours.
26. If a random variable has the gamma distribution with $\alpha = 2$, $\beta = 1$, find $P(1.8 < X < 2.4)$.
27. For a certain dose of the toxicant, a study on mice determines that the survival time, in weeks, has a gamma distribution with parameter $\alpha = 5$. What is the probability that a mouse survives no longer than 60 weeks?
28. The survival time in weeks of an animal when subjected to certain exposure of gamma radiation has a gamma distribution with $\alpha = 5$ and $\beta = 1/10$.
    (a) What is the mean survival time of a randomly selected animal of the type used in the experiment?
    (b) Find the S.D. of the survival time.
    (c) What is the probability that an animal survives more than 30 weeks?
29. Suppose that proportion of defectives, supplied by a vendor from lot to lot may be looked upon as a random variable having the beta distribution with $\alpha = 2$ and $\beta = 3$.
    (a) Find the average proportion of defectives in a lot from this vendor.
    (b) Find the probability that a lot from this vendor will contain 30% or more defectives.
30. In a certain city, the daily consumption of water (in million of litres) follows approximately a gamma distribution with $\alpha = 2$ and $\beta = 3$. If the daily consumption of that city is 9 million litres of water, what is the probability that on any given day the water is inadequate? Also find the mean and variance of the daily water consumption.

ANSWERS

1. 1/5, 4/5
2. (a) 0.6 (b) 0.7 (c) 0.5
3. (a) 0.1151 (b) 0.5403 (c) 16.1 (d) 20.275
4. 4.082 ounces
5. mean 50, S.D. 10,

6. (a) 67 (b) 184 (c) 1909
7. 6.24 yrs
8. 15.9%
9. 58%
10. (a) 0.4586 (b) 0.0526 (c) 0.0170
11. (a) 0.3085 (b) 0.2417 (c) 0.0045
12. 1, 4, 6, 4, 1, $y = \frac{\sqrt{2}}{4\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-6}{2}\right)^2}$
14. (a) 0.2376 (b) 0.0576
15. (a) 0.9782 (b) 0.0059 (c) 0.0121
16. (a) 0.3085 (b) 0.0197
17. 0.1314
18. (a) $\mu = 7.43$, $\sigma^2 = 0.55$ (b) 0.8139
19. (a) 0.3935 (b) 0.4346
20. 0.3968
25. (a) 200 hours (b) 0.177
26. 0.1545
27. 0.715
28. (a) mean 50, (b) S.D. 22.36 (c) 0.8155

6 CHAPTER Correlation and Regression

BB 56620

6.1 INTRODUCTION

In a bivariate distribution, we often come across situations in which movements in one variable are accompanied by movements in other variables. For example, the selling price of an apartment may vary with the covered area, or the extension resulted in a piece of wire vary with the force applied. In this chapter, we shall study this relationship between two random variables from statistical point of view. This problem is that of *correlation*. Another related aspect is predicting the value of the dependent variable (y) for a specific value of the independent variable (x). This problem is that of *regression* and we find a *regression line* that describes the dependence of y on x. The appropriate line to be fitted to the given data is obtained by the *method of least squares*. In case of more than two variables the problem becomes of *partial and multiple correlation and regression*.

In Section 6.2, we describe the method of least squares and curve fitting to a bivariate data. In Sections 6.3 and 6.4, we study the problem of correlation, and in Section 6.5, rank correlation is considered. The problem of regression and derivation of regression lines is considered in Section 6.6. The multiple and partial correlation has been discussed in Section 6.7. Chapter is concluded with a set of review exercises and a problem set.

6.2 METHOD OF LEAST SQUARES AND CURVE FITTING

Fitting of curve to a given bivariate data is important both from the point of view of theoretical and practical statistics. Theoretically, this is useful in the study of correlation and regression and practically, functional relationship between x and y enables us to predict the response y for a specific input x. The appropriate relationship to be fitted may be polynomial, algebraic, exponential or logarithmic depending upon the nature of the data. *Method of least squares* is an excellent technique for fitting an appropriate relationship to the given data.

7 Sampling Distributions and Large Sample Estimation

CHAPTER

7.1 INTRODUCTION

In the preceding chapters, we have learnt about special probability distributions when the values of the numerical descriptive measures, that is, parameters were known in advance. Sometimes we may be able to specify the type of probability distribution to be used as model, but the values of the parameters that specify the exact form of the distribution are unknown. In such a situation, we rely on the sample drawn to learn about these unknown parameters. Problems in which the form of the underlying distribution is specified up to a set of unknown parameters are called *parametric inference problems*, whereas those in which nothing is assumed about the underlying distribution are called *non-parametric inference problems*.

In any particular study, the number of observations recorded may be finite or infinite. For example, number of defective screws in a box of 1000 results in a finite number of observations while if we could toss a pair of dice indefinitely and record the total obtained, then we obtain an infinite set of observations. A *population* consists of the totality of the observations under study. In case the number of observations are finite, population is called *finite population* otherwise *infinite population*. When it is not desirable to take into account all the observations, then we take a finite subset of the population called *sample*. The main *advantages of sampling* over complete enumeration are of reduced cost, greater speed and scope, and sometimes, even better precision. When testing is destructive in nature, then sampling becomes necessary.

In this chapter, we consider parametric inference problems and focus on sampling from population and study sample statistics like, sample mean ($\bar{x}$), sample variance (s^2), etc. and see how the information drawn from the sample is utilized to draw some conclusion about the population parameters like, mean (μ), variance (σ^2), etc. In Section 7.2, we consider the various sampling plans for selecting a sample. The concept of statistics and sampling distribution is explained in Section 7.3. In Section 7.4, we discuss the central limit theorem, one of the most important result in statistical theory. Sections 7.5 and 7.6 deal respectively with the sampling distribution of the sample mean and sample proportion. Tests of significance and related aspects have been discussed in Section 7.7. In Section 7.8, we discuss large sample testing, leading to simple sampling of attributes in Section 7.9, and sampling of variables in Section 7.10. In the end, a set of review exercises and a problem set is given.

7.2 SAMPLING PLANS

The way a sample is selected from the population under study is called the *sampling plan* and determines the quantity of information in the sample. Some of the commonly employed sampling plans are:

(a) *Random sampling* If each unit of the population has the same chance of being selected in the sample, then sampling is said to be random sampling. Suppose we take a sample of size n from a population of finite size N, then in case of random sampling each of the ${}^{N}C_{n}$ samples has the same probability, that is, $1/{}^{N}C_{n}$ of being selected.

The simplest method of drawing a random sample is the lottery method, assigning numbers 1 to N to each unit of the population; writing these numbers on N identical slips; putting these slips in a box and then drawing n slips one by one from this well-shuffled lot. The n units corresponding to the numbers on the slips drawn constitute the random sample. For example, if we need to select a sample of size $n = 2$ from a population containing $N = 4$ elements, say denoted by x_1, x_2, x_3 and x_4, then there are six distinct pairs, listed as x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4, that could be selected. These six samples have the same probability, each equal to 1/6 of selection and the resulting sample is called a *random sample*.

The above method of selecting a random sample is not very practical, particularly when the population is large. A simpler and more reliable method is the use of random numbers. *Random numbers* are the digits generated so that values 0 to 9 occur randomly with equal frequency. These numbers can be generated by a computer, or even by a scientific calculator. Alternatively, these are available from *random numbers tables*. One such table is given as Table VII (*See p. 348 Appendix 1*). The chance mechanism that generates the random number table ensures that each of single digits 0, 1, 2, ..., 9 has the same chance of occurence, all pairs 00, 01, ..., 99 have the same chance of occurence, and so on. Also any collection of digits is random in nature.

To illustrate the use of random numbers table, suppose we are to take a random sample of 8 blood samples out of the 80 investigated for reinvestigation. We number the samples investigated from 1 to 80. Since the population size $N = 80$ is a two-digit number, the digits must be selected two at a time. We begin by arbitrarily selecting a row and a column. Say, we select row 11 and column 5 in the Table VII. The relevant portion is

11	8135	5004	7299	8981	4689	1950	2271	2201	8344	3852
12	4414	6855	0127	5489	5157	6386	7492	3736	7164	0498
13	3727	7959	5056	5983	8021	0204	7616	4325	7454	5039
14	5434	7342	0314	7525	0067	2800	6292	4706	3454	6881
15	7195	8828	9869	2785	3186	8375	7417	7232	0401	2483
16	2705	8245	6251	9611	1077	0641	0195	7024	6202	3899
17	1547	8981	4972	1280	4286	5678	0338	8096	8284	7010
18	3424	1435	1354	7631	7260	7361	0151	8903	9056	8684
19	8969	7551	3695	4915	7921	2913	3840	9031	9747	9735
20	5225	8720	8898	2478	3342	9200	8836	7269	2992	6284
21	6432	9861	1516	2849	2539	2208	4595	8616	6170	5865
22	3085	5903	8319	2744	0814	7318	8619	7614	3265	5999
23	0264	1246	3687	9759	6995	6565	3949	1012	0179	0059

Reading the digits in columns 5, 6 and proceeding downward, we obtain

50, 68, 79, 73, 88, 82, 89, 14, 75, 87, 98, 59, 12

Ignoring the number greater than 80 and ignoring any when it appears second time, continue reading until eight different numbers in the range 1 to 80 are selected. The sample obtained is

50, 68, 79, 73, 14, 75, 59, 12

So the blood-samples numbered above will be re-investigated.

(b) *Simple random sampling* Simple random sampling is random sampling in which each of the population has an equal probability p of being included in the sample and probability is independent of the previous drawings. Thus, random sampling becomes simple, if either the units are drawn with replacement, or when the population is infinite. simple sample of size n from a population may be identified with a series of n Bernoulli trials with constant probability p of success for each trial. A simple random sample can drawn using random numbers Table VII in the same was as explained above only with exception that here we don't ignore the repeated numbers.

(c) *Stratified sampling* In case the population is heterogeneous, then it is divided into homogeneous *stratas* (groups) of various sizes. Then units are sampled at random from each of these stratas according to the relative importance of the stratas in the population. The sample drawn thus will be more representative than the simple random sample, since each strata will be represented in the sample drawn proportion to its size.

(d) *Systematic sampling* Let there be population with $N = nk$ ordered units from 1 to N. systematic sampling if we are to draw a sample of size n, then we draw a unit at random from the first k ordered units and, then every kth unit is drawn to form the sample.

(e) *Purposive sampling* Here the units are selected with definite purpose in view. Usually, selected units do not form a representative sample of the population and yield results which are generally biased.

In general, not all sampling plans involve random selection but any sampling plan used for drawing inferences must involve randomization.

7.3 STATISTICS AND SAMPLING DISTRIBUTIONS

The numerical descriptive measures calculated from the sample are called *statistics* and the numerical descriptive measures of the population, (generally unknown), are called *parameters*. The statistics vary for each different random sample selected and hence they are random variables. The probability distributions for statistics, e.g., sample mean, sample variance, etc. are called *sampling distributions*.

For example, consider a population consisting of 5 numbers 3, 5, 7, 9, 11. If a random sample of size $n = 2$ is selected without replacement, then we can find the sampling distribution of the sample mean $\bar{x}$ as follows.

There are 10 possible equally likely random samples each of size $n = 2$. The values of $\bar{x}$ in random sampling when $n = 2$ and $N = 5$ are tabulated below.

Sample	Sample units	Sample mean, $\bar{x}$
1	3, 5	4
2	3, 7	5
3	3, 9	6
4	3, 11	7
5	5, 7	6
6	5, 9	7
7	5, 11	8
8	7, 9	8
9	7, 11	9
10	9, 11	10

Hence, sampling distribution of the sample mean $\bar{x}$ is

$\bar{x}$ =	4	5	6	7	8	9	10
$f(\bar{x})$ =	1	1	2	2	2	1	1
$p(\bar{x})$ =	0.1	0.1	0.2	0.2	0.2	0.1	0.1

$$\text{Population mean, } \mu = \frac{3+5+7+9+11}{5} = 7$$

$$\text{Population variance, } \sigma^2 = \frac{(-4)^2+(-2)^2+0+(2)^2+(4)^2}{5} = 8$$

Mean of 'sample means' $= 0.4 + 0.5 + 1.2 + 1.4 + 1.6 + 0.9 + 1.0 = 7.0$

Variance of 'sample means' $= (-3)^2(0.1) + (-2)^2(0.1) + (-1)^2(0.2) + 0 + (1)^2(0.2) + (2)^2(0.1) + (3)^2(0.1)$

$= 0.9 + 0.4 + 0.2 + 0.2 + 0.4 + 0.9 = 3.0$

We observe that mean of the sample means is the same as the population mean but variance of the sample means is not the same as the population variance. The standard deviation of sampling distribution of a statistic is called its *standard error* (S.E.). In this case S.E. is $\sqrt{3}$. Normally, we employ statistical theory to derive sampling distribution of a statistic or use simulation to derive the sampling distribution empirically.

An important statistical theorem which describes the sampling distribution of a statistic which are sum or averages of the sample observations is the *central limit theorem* as presented in the next section.

7.4 THE CENTRAL LIMIT THEOREM

The central limit theorem is one of the most remarkable result in probability theory which asserts that the sum of a large number of independent random variables is approximately distributed as a normal variate. This provides a simple method for computing approximate probabilities for sum of independent random variables and also explains the fact why empirical frequencies of so many natural populations exhibit a normal curve.

In its simplest form the theorem is stated as follows.

Theorem 7.1 (Central Limit Theorem): *If $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables each with mean μ and variance σ^2, then the sum $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically normal with mean $n\mu$ and variance $n\sigma^2$. That is, for large n the variable*

$$Z = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

is distributed like a standard normal variate.

Proof Let $M(t)$ be the moment generating function of each of the deviation $(X_i - \mu)$ and $M_Z(t)$ the m.g.f. of the standard variable $Z = (S_n - n\mu)/\sigma\sqrt{n}$.

Since $E(X_i - \mu) = 0$ and $E(X_i - \mu)^2 = \sigma^2$, thus

$$M(t) = 1 + \mu_1' t + \mu_2' \frac{t^2}{2!} + \mu_3' \frac{t^3}{3!} + \ldots$$

$$= 1 + \sigma^2 \frac{t^2}{2!} + \text{terms with } t^3 \text{ and higher powers of } t$$

Next, we have

$$Z = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{(X_1 + X_2 + \ldots + X_n) - n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma\sqrt{n}}\right)$$

Since X_i's are independent, thus

$$M_Z(t) = M_{\sum_{i=1}^{n}(X_i - \mu)/\sigma\sqrt{n}}(t)$$

$$= M_{\sum_{i=1}^{n}(X_i - \mu)}\left(t/\sigma\sqrt{n}\right)$$

$$= \prod_{i=1}^{n} M_{(X_i - \mu)}\left(t/\sigma\sqrt{n}\right)$$

$$= \left[M\left(t/\sigma\sqrt{n}\right)\right]^n$$

$$= \left[1 + \frac{t^2}{2n} + \text{terms with } n^{-3/2} \text{and lower powers of } n\right]^n, \quad \text{using (7.1)}$$

For every fixed t, terms with $n^{-3/2}$ and lower powers of t tends to zero as $n \to \infty$. Therefore $n \to \infty$, we obtain

$$\lim_{n\to\infty} M_z(t) = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n}\right]^n = \exp(t^2/2),$$

which is the m.g.f. of a standard normal variate. Thus,

$$Z = \frac{S_n - n\mu}{\sigma\sqrt{n}} \sim N(0, 1)$$

or

$$S_n \sim N(n\mu, n\sigma^2).$$

This completes the proof.

Remarks

1. The form given above is a particular case of the more general form of the central limit theorem stated as follows:
   If X_i $(i = 1, 2, \ldots, n)$ are independent random variables with mean and variance μ_i and σ_i^2 respectively for each i, then the random variable $S_n = \sum_{i=1}^{n} X_i$ is asymptotically normal with mean and variance $\sum_{i=1}^{n}\mu_i$ and $\sum_{i=1}^{n}\sigma_i^2$ respectively.
2. Another important form of the central limit theorem is in the context with the binomial random variable. The result is stated as follows:
   If
   $$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q \end{cases}$$
   where $p + q = 1$, then $S_n = \sum_{i=1}^{n} X_i$, where X_i's are independent, is asymptotically normal with mean np and variance npq.
   The proof for this is on the same lines as given in *Theorem 7.1* above.
3. The central limit theorem has important contribution in statistical inference since many estimates that are used to make inference about population parameters are sum or averages of the sample observations, and when the sample size n is large, then these estimators can be approximated as normal variates. In case the population itself is normal, then sampling distribution of $\bar{x}$ is always normal, irrelevant of the size n of the sample selected. But when the population is skewed then the sample size n must be large, say $n > 30$ to approximate the distribution of $\bar{x}$ as normal.

Example 7.1: Suppose that the amount of weight W (in '000 pounds) that a certain span of a bridge can withstand without resulting in structural damage, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (in '000 pounds) of a car is random variable with mean 3 and standard deviation 0.3. How many cars would have to be on the bridge span for the probability of structural damage to exceed 0.1?

Solution: If P_n denotes the probability of structural damage when there are n cars on the bridge, then

$$P_n = P\{X_1 + X_2 + \ldots + X_n \geq W\}$$
$$= P\{X_1 + X_2 + \ldots + X_n - W \geq 0\},$$

where X_i is the weight of the ith car, $i = 1, 2, \ldots, n$. From central limit theorem, $\sum_{i=1}^{n} X_i$ asymptotically normal with mean $3n$ and variance $0.09n$. Further, since W is independent of X_i and is also normal, thus $\sum_{i=1}^{n} X_i - W$ is approximately normal with mean equal to $3n - 400$ and variance $0.09 + 1600$.

Let
$$Z = \frac{\left(\sum_{i=1}^{n} X_i - W\right) - (3n - 400)}{\sqrt{0.09n + 1600}}$$

then
$$P_n = P\left(Z \geq \frac{-(3n - 400)}{\sqrt{0.09n + 1600}}\right),$$

where $Z \sim N(0, 1)$.

From *Table I*, $P(Z \geq 1.28) \simeq 0.1$. Thus, if the number of cars n is such that

$$\frac{400 - 3n}{\sqrt{0.09n + 1600}} \leq 1.28,$$

or, $n^2 - 266.54\, n + 17486.51 \leq 0$

or, $(n - 149.84)(n - 116.66) < 0$

that is, if $n \geq 117$, only then there is at least 1 chance in 10 that structural damage will occur.

Example 7.2: The number of tourists which can be adjusted comfortably in a coach is 50. The owner, knowing from its past experience that on the average only 80% of those booked seats will actually join the tour, book 60 tourists. Compute the probability that more than 50 tourists will join the tour.

Solution: Let X be the number of tourists who join the tour, then under the assumption that each tourist will act independently. Using central limit theorem, it follows that X is a binomial random variate with parameters

$$n = 60 \quad \text{and} \quad p = 0.8$$

Since binomial is a discrete distribution; applying continuity correction, we calculate $P(X = i)$ as $P(i - 0.5 < X < i + 0.5)$, and thus, $P(X > 50)$ as

$$P(X > 50.5) = P\left(\frac{X - 60(0.8)}{\sqrt{60(0.8)(0.2)}} \geq \frac{50.5 - 60(0.8)}{\sqrt{60(0.8)(0.2)}}\right)$$

$$= P(z > 0.807) = 0.209, \quad \text{using } \textit{Table I}$$

Thus, about 21% of the time more than 50 of the first 60 tourists booked will actually join the tour.

7.5 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

If the population mean μ is unknown, then the statistic sample mean $\bar{x}$, in general, is chosen as the natural estimate of the population mean. The following theorem gives sampling distribution of the sample mean $\bar{x}$.

Theorem 7.2 (Sampling Distribution of the Sample Mean): *If a random sample of size n is selected from a population with mean μ and S.D. σ, then the sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard error (S.E.) $\sigma/\sqrt{n}$.*

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

We have,
$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \qquad \ldots(7.2)$$

and,
$$\text{var}(\bar{x}) = \text{var}\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n^2}[\text{var}(x_1) + \text{var}(x_2) + \ldots + \text{var}(x_n)],$$

since x_i's are independent, thus the covariances terms are absent. This gives

$$\text{var}(\bar{x}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad \ldots(7.3)$$

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

Hence, $\bar{x}$ is distributed with mean μ and S.E. $\sigma/\sqrt{n}$.

In case the sampled population is normal, the distribution of $\bar{x}$ will be exactly normal irrespective of the size n, however, if the population is non-normal, then the distribution of $\bar{x}$ will be approximately normal for large n by central limit theorem. Thus, the statistic z, defined by

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

Remark: A statistic θ is called an *unbiased estimator* of a polulation parameter γ, if $E(\theta) = \gamma$. Since, from (7.2), $E(\bar{x}) = \mu$, thus *sample mean $\bar{x}$ is an unbiased estimate of the population mean* μ. But sample variance, $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ is not an unbiased estimate of the population variance as shown below.

We have,
$$E(s^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) - E(\bar{x}^2) \qquad \ldots(7.4)$$

Now, $E(x_i^2) = \text{var}(x_i) + [E(x_i)]^2 = \sigma^2 + \mu^2$,

and, $E(\bar{x}^2) = \text{var}(\bar{x}) + [E(\bar{x})]^2 = \dfrac{\sigma^2}{n} + \mu^2$, using (7.2) and (7.3).

Using these in (7.4), we obtain

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

However, if we define $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then

$$E(S^2) = E\left(\frac{n}{n-1}s^2\right) = \frac{n}{n-1}E(s^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

...(7.5)

Hence, $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ *is an unbiased estimate of the population variance* σ^2.

We shall discuss the sampling distribution of S^2 in *Section 8.4.*

7.6 THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

In some practical situations, we need to estimate the proportion p of people in the population who have a specified characteristic say smoking, computer literacy, etc. If x out of the n sampled people have this characteristic, then the sample proportion $\hat{p} = x/n$ can be taken as an estimate of the population proportion p. We observe that the distribution of the random variable x is binomial with mean np and S.D. $\sqrt{npq}$, and thus $\hat{p} = \dfrac{x}{n}$ will also be distributed like a binomial variate with mean as

$$E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = p,$$

...(7.6)

variance as

$$\text{var}(\hat{p}) = \text{var}\left(\frac{x}{n}\right) = \frac{1}{n^2}\text{var}(x) = \frac{pq}{n}$$

...(7.7)

and, hence the standard error, as

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{pq}{n}},$$

...(7.8)

where, $q = 1 - p$.

Further, since the binomial distribution can be approximated to normal for large n, thus the statistic z given by

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

will be a standard normal variate for large n, that is, $z \sim N(0, 1)$.

...(7.9)

Example 7.3: An electrical firm manufactures light bulbs that have burning life normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average burning life of less than 775 hours.

Solution: Let $\bar{x}$ denote the average burning life of 16 bulbs, then $\bar{x}$ is a normal variate with mean 800 hrs and S.E. $= 40/\sqrt{16} = 10$. Thus,

$$z = \frac{\bar{x} - 800}{10} \sim N(0, 1).$$

Hence, $P(\bar{x} < 775) = P(z < -2.5) = P(z > 2.5) = 0.5 - P(0 < z < 2.5) = 0.5 - 0.4938 = 0.0062$, using *Table 1.*

Example 7.4: The duration of Alzheimer's disease from the appearance of symptoms until death, is distributed with an average of 9 years and a S.D. of 4 years. The medical records of 36 randomly selected deceased patients from a large medical database has been taken. Find the approximate probability that average duration lies within 7 and 11 years.

Solution: Let $\bar{x}$ be the average duration of survival after the appearance of symptoms. Since the sample size n is 36, $\bar{x}$ can be approximated as a normal variate with mean $\mu = 9$ and standard error $= \sigma/\sqrt{n} = 4/\sqrt{36} = 2/3$. Thus

$$z = \frac{\bar{x} - 9}{2/3} \sim N(0, 1).$$

Hence, $P(7 < \bar{x} < 11) = P(-3 < z < 3) = 2P(0 < z < 3) = 2(0.4987) = 0.9974$, using *Table I.*

Example 7.5: A random sample of 100 students was taken from a campus and 12 were found to be smokers. Estimate the proportion of smokers in the campus as well as the S.E. of the estimate. Find the almost certain limits to the percentage of smokers in the campus.

Solution: The proportion of smokers in the sample of size $n = 100$, is

$$\hat{p} = \frac{12}{100} = 0.12, \quad \hat{q} = 0.88$$

Thus, $$\text{S.E.}(\hat{p}) = \sqrt{\frac{(0.12)(0.88)}{100}} = 0.0325,$$

Hence, the proportion of smokers lies certainly between

$$\hat{p} \pm 3(\text{S.E.}) = 0.12 \pm 3(0.0325) = 0.12 \pm 0.0975,$$

that is, between 0.0225 and 0.2175. Therefore, the percentage of smokers almost certainly lies between 2.25 and 21.75.

7.7 TESTS OF SIGNIFICANCE

An important aspect of sampling theory is to make decision about the parameter value. The tests hypothesis enable us to decide on the basis of the statistic obtained, that whether the deviation between the observed and the theoretical value is significant or might be attributed to fluctuations of sampling. Since for large n the sampling distribution of the statistic under study can approximated to normal, so for large sample testing normal distribution is applied. However, case of small sample testing we employ specific variates like t, χ^2, F, etc.

7.7.1 Null and Alternative Hypotheses

A *statistical hypothesis* is an assertion concerning one or more populations. A hypothesis we wish test, is called the *null hypothesis* and is denoted by H_0; and any hypothesis, complementary to null hypothesis, is called an *alternative hypothesis* and is usually denoted by H_1.

The *null hypothesis* H_0 *is usually a hypothesis of no-difference and is tested for possible rejection under the assumption that it is true.*

For example, if we want to test that average daily wages of workers in a construction compa is different from Rs. 170, the national average, then we can set up the null hypothesis as

$$H_0 : \mu = 170$$

and, the alternative hypothesis could be any of

(a) $H_1 : \mu \neq 170$ (b) $H_1 : \mu < 170$ (c) $H_1 : \mu > 170$.

The alternative hypothesis, (a) is known as *two-tailed alternative;* (b) is known as *left-tailed alternative;* and (c) as *right-tailed alternative.*

The hypothesis (a) is *composite alternative*, while hypotheses (b) and (c) are *simple alternatives*.

7.7.2 Acceptance and Rejection Regions

The decision to reject or accept the null hypothesis is based on the information contained in sample drawn from the population under study. On the basis of the data in the sample, a test *statistic* is formulated and using this test statistic a probability value is calculated (generally obtained from the statistical tables available). On the basis of these measures obtained, the hypothesis H_0 is rejected or accepted. Now the important question arises: *How to decide whether reject or accept* H_0? This is answered as follows.

Since, our decision is based on the value of the test statistic obtained, thus entire set of values that the test statistic may attain is divided into two regions, the acceptance region and the rejection region.

The region consisting of the values which support the null hypothesis, leading to acceptance H_0, is called the *acceptance region*, and the region consisting of values which support the alternative hypothesis, leading to rejection of H_0, is called the *rejection region* or the *critical region*. The value that separate the acceptance and rejection region is (are) called the *critical value* (*s*).

For example, in case we wish to test the null hypothesis $H_0 : \mu = 170$ against the alternative $H_1 : \mu \neq 170$, then the acceptance and rejection regions are as shown in Fig. 7.1a.

In case the alternative is $H_1 : \mu < 170$, or $H_1 : \mu > 170$, then acceptance and rejection regions are as shown in Figs. 7.1b and 7.1c, respectively.

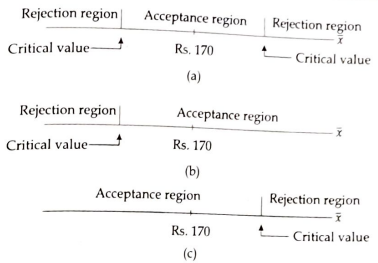

Fig. 7.1

The type of test in case of (a) is called *two-tailed test* and, in case of (b) and (c) is called *left-tailed test* and *right-tailed test*, (jointly as *single-tailed tests*), respectively.

7.7.3 Types of Errors and Level of Significance

The decision procedure described above can lead to either of the following two types of errors:

Type I error : *Rejection of the null hypothesis* H_0 *when it is true (Rejection error).*

Type II error : *Acceptance of the null hypothesis* H_0 *when it is false (Acceptance error).*

The probability of Type I error, that is,

$$P\{\text{Rejection of } H_0 \text{ when it is true}\} = P\{\text{Reject } H_0 \mid H_0\},$$

is called the *level of significance*, or *size of the test* and is denoted by α.

The probability of Type II error, that is,

$$P\{\text{Acceptance of } H_0 \text{ when it is false}\} = P\{\text{Accept } H_0 \mid H_1\}$$

is denoted by β. The factor $1 - \beta$, the probability of rejecting H_0, when a specific alternative H_1 is true, is called the *power of a test*.

For a fixed sample size, a decrease in the probability of one type error will usually result in an increase in the probability of the other type of error. Both types of errors can be reduced only by increasing the sample size n. For applying the test of significance, the level of significance, that is, the size of Type I error is kept fixed normally at 5% or 1% and the sampling is so designed that for given α, the size of the Type II error β is minimum.

Remark In statistical quality control 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, thus α and β are often referred to as *producer's risk* and *consumer's risk*, respectively.

7.8 LARGE SAMPLES TESTING

In case of large sampling the test statistic z, say $z = \dfrac{\bar{x} - E(\bar{x})}{S.E.(\bar{x})}$, is approximated to $N(0, 1)$. If z_α is the critical value of the test statistic z at level of significance α, then for a *two-tailed test* it is given by

$P(|z| > z_\alpha) = \alpha$, that is, z_α is the value so that the total area of the critical region on both tails is α, the standard normal probability curve is symmetrical about its mean $z = 0$, thus

$$P(z > z_\alpha) = P(z < -z_\alpha) = \alpha/2,$$

as shown in Fig. 7.2a.

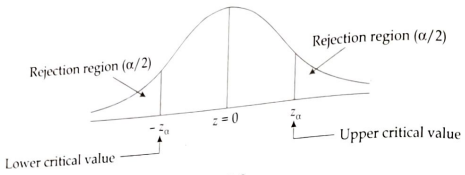

Fig. 7.2a

In case of left-tailed test or right-tailed test the total area to the left of $-z_\alpha$, or to the right of z_α the case is, is α. That is, for the left-tailed test $P(z < -z_\alpha) = \alpha$ and for the right-tailed test $P(z > z_\alpha)$ as shown in Figs. 7.2b and 7.2c, respectively.

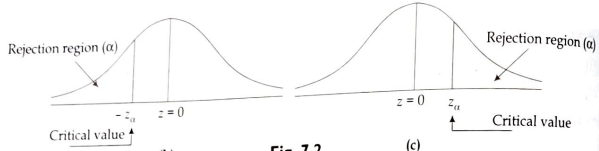

Fig. 7.2

In case of large sample testing, the critical value z_α of z for two-tailed test at level of significance α is numerically same as the critical value for a single-tailed test at level of significance $\alpha/2$, but this does not hold good when the sample size is small.

When test statistic is approximated to standard normal distribution, for two-tailed test critical values at 1% and 5% level of significance are 2.58 and 1.96, respectively. For left-tailed tests, these values are −2.33 and −1.645, respectively and, for right-tailed tests, the values are 2.33 and 1.645 respectively, refer to *Table I*.

Procedure for testing We set up the null hypothesis H_0 and compute the test-statistic z under the assumption that H_0 is true. If $|z| > 3$, H_0 is rejected outright. In case $|z| \le 3$, we test its significance at a specified level usually at 5% or 1% level of significance.

For a two-tailed test, if $|z| > 1.96$, H_0 is rejected at 5% level of significance and if $|z| > 2.58$ also it is rejected even at 1% level of significance also. In case $|z| < 2.58$, H_0 may be accepted at 1% level of significance, and if $|z| < 1.96$, then H_0 may be accepted at 5% level of significance also.

For a single-tailed test, the test-statistic $|z|$ is compared with 1.645 at 5% level and with 2.33 at 1% level and H_0 is accepted and rejected accordingly.

Remark In case the sample size n is small, say <30, and sampled population is not normal then distribution of the test-statistic cannot be approximated to normal and thus these critical values don't hold good. In such cases the values based on the exact sampling distribution of the test statistic are used e.g. t, χ^2, F, etc. to be studied in *Chapter 8*.

7.9 SIMPLE SAMPLING OF ATTRIBUTES. TESTS FOR SINGLE PROPORTION AND DIFFERENCE BETWEEN TWO PROPORTIONS

The theory of sampling can be studied under two heads: *sampling of attributes* and *sampling of variables*. In this section, we discuss the sampling of attributes.

In the sampling of attributes, we are concerned only with the possession or non-possession of some specified attribute or, characteristic say, smoking, swimming, inoculated against a disease, etc. by the individuals selected in the sample. The possession of the specified attribute by the individual selected in the sample is termed as *success* while the non-possession as *failure*.

In this case, simple sampling of n observation may be identified with that of a series of n independent Bernoulli trials with constant probability p of success for each trial and so the mean number of successes is np and the S.E. of the number of successes is $\sqrt{npq}$, where $q = (1 - p)$.

In case we consider the proportion of successes $\hat{p} = \frac{x}{n}$, then mean and S.E. of proportion of successes are given by respectively

$$\text{Mean}\,(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{np}{n} = p$$

and,

$$\text{S.E.}\,(\hat{p}) = \sqrt{\text{var}\left(\frac{x}{n}\right)} = \frac{1}{n}\text{S.E.}\,(x) = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

The factor $\sqrt{n}/\sqrt{pq}$ is called the *precision of the proportion of successes* and varies as $\sqrt{n}$, since $\sqrt{pq}$ is constant for the specific population under study.

7.9.1 Test for Single Proportion

If x is the number of successes in n independent trials and suppose we wish to test the hypothesis H_0 that proportion of success in each trial is p, then under H_0

$$E(x) = np \text{ and S.E.}\,(x) = \sqrt{npq}$$

and for large n, the test statistic z given by

$$|z| = \frac{|x - np|}{\sqrt{npq}} \qquad \ldots(7.10)$$

is a standard normal variate, that is, $z \sim N(0, 1)$ and so we can apply the normal test.

Similarly, for large n the test statistic z for proportion of successes $\hat{p}$ given by

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}} \qquad \ldots(7.11)$$

is a standard normal variate, that is, $z \sim N(0, 1)$.

Further, the limits

$$x = np \pm 1.96\sqrt{npq}$$

are called the *95% confidence limits*, and the limits

$$x = np \pm 2.58\sqrt{npq}$$

are called the *99% confidence limits* for the number of successes x.

Similarly, we can write the 95% and 99% confidence limits for the proportion of successes.

Example 7.6: A dice is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Can the dice be regarded as unbiased? Also find the limits between which the probability of a throw of 3 or 4 is most likely to lie.

Solution: Let the null hypothesis H_0 be that dice is unbiased. Under H_0, if p is the probability of getting a throw of 3 or 4, then we test

$$H_0: p = 1/3, \text{ against the alternative } H_1: p \neq 1/3.$$

Here $n = 9000$, $x = 3240$, $np = 3000$, $q = 2/3$. Thus, $\sqrt{npq} = \sqrt{9000 \times 1/3 \times 2/3} = 44.72$

Under H_0 the test statistic z, given by

$$|z| = \frac{|x - np|}{\sqrt{npq}} = \frac{240}{44.72} = 5.37 > 3$$

is highly significant and hence the hypothesis H_0 is rejected and we regard that dice is almost certainly biased.

Since, the dice is not unbiased, the most likely limits in which the probability of a throw of 3 or 4 lie are given by

$$\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}}, \text{ where } \hat{p} = \frac{x}{n} = \frac{3240}{9000} = 0.36, \text{ and } \hat{q} = 0.64$$

Hence, the limits are $0.36 \pm 3\sqrt{\frac{(0.36)(0.64)}{9000}} = 0.345$ and 0.375.

Example 7.7: During testing in a sample of 300 chips 10 have been found to be defective. Can the manufacturer's claim that 2% of the chips are defective may be accepted?

Solution: Let the null hypothesis H_0 be that 2% of the chips are defective, thus we test

$$H_0: p = .02 \text{ against } H_1: p \neq .02$$

We have, $n = 300$, $np = 6$, $\sqrt{npq} = \sqrt{300(.02)(.98)} = 2.42$

Under H_0, the test statistic z, given by

$$|z| = \frac{|x - np|}{\sqrt{npq}} = \frac{|10 - 6|}{2.42} = 1.65 < 1.96$$

is not significant and hence H_0 is accepted at 5% level of significance; manufacturer's claim may be accepted.

Example 7.8: Long-term database indicates that 5% of the components produced at a certain manufacturing facility are defective. A training programme for the workforce employed has been conducted with the aim to reduce the percentage of defective produced. After this if a random sample of 500 items consists of 16 defectives, can we conclude that training was effective?

Solution: Let the null hypothesis H_0 be that the training was not effective in reducing the proportion p of defectives. Thus, we test

$$H_0: p = 0.05 \text{ against the left-tailed alternative } H_1: p < 0.05$$

We have, $n = 500$, $np = 25$, $\sqrt{npq} = \sqrt{500(0.5)(0.95)} = 4.87$.

Under H_0, the test statistic z, given by

$$|z| = \frac{|x - np|}{\sqrt{npq}} = \frac{|16 - 25|}{4.87} = 1.848 > 1.645$$

is significant and hence, using the left-tailed test, hypothesis is rejected at 5% level of significance. However, since $|z| = 1.848 < 2.33$, by left-tailed test the hypothesis may be accepted at 1% level of significance.

Example 7.9: Out of the twenty persons who were reported to be attacked by brain fever only eighteen survived. Using the large sample test, test the hypothesis at 5% level that if, attacked by brain fever survival rate is 85% against the alternative that it is more.

Solution: Let the null hypothesis be that survival rate is 85%, that is, p = 0.85. Thus, we test

$$H_0: p = 0.85, \text{ against right-tailed alternative } H_1: p > 0.85.$$

We have, $n = 20$, $x = 18$, $\hat{p} = x/n = 0.9$ and $\sqrt{pq/n} = \sqrt{(0.85)(0.15)/20} = 0.0798$

Under H_0, the test-statistic is

$$|z| = \frac{|\hat{p} - p|}{\sqrt{pq/n}} = \frac{|0.90 - .85|}{.0798} = 0.627,$$

which is less than 1.645. Using the right-tailed test it is not significant at 5% level and hence the hypothesis may be accepted at 5% level of significance.

Example 7.10: In a random sample of 525 families owning television set in the region of New Delhi, it is found that 370 subscribe to Star Plus. Find a 95% confidence interval for the actual proportion of such families in New Delhi which subscribe to Star Plus.

Solution: We have, $\hat{p} = x/n = 370/525 = 0.705$

Therefore the 95% confidence limits for the actual proportion p are

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.705 \pm 1.96\sqrt{\frac{(0.705)(0.295)}{525}},$$

that is, between 0.666 and 0.744. Hence, the 95% confidence interval for the actual proportion p is $0.666 < p < 0.744$.

7.9.2 Test for Difference Between Two Proportions

Suppose we want to compare two distinct populations with respect to the prevalence of a specific attribute. For example, we may be interested in comparing the prevalence of lung cancer among smokers (population I) and non-smokers (population II). Let x_1, x_2 be the number of persons with this attribute in random samples of size n_1 and n_2 selected from population I and population II respectively. Then sample proportions are

$$\hat{p}_1 = x_1/n \quad \text{and} \quad \hat{p}_2 = x_2/n.$$

If p_1 and p_2 are proportion for the two populations, then

$$E(\hat{p}_1) = p_1, \quad E(\hat{p}_2) = p_2, \quad \text{and} \quad \text{var}(\hat{p}_1) = \frac{p_1 q_1}{n_1}, \quad \text{var}(\hat{p}_2) = \frac{p_2 q_2}{n_2}.$$

Since for large samples $\hat{p}_1$ and $\hat{p}_2$ are each approximately normally distributed with means p_1 and p_2 and variances p_1q_1/n_1 and p_2q_2/n_2 respectively, and also the samples being independent, thus $\hat{p}_1 - \hat{p}_2$ is also normally distributed with mean

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2,$$

and,

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Thus,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0, 1) \qquad \ldots(7.12)$$

Let H_0: There is no difference between the population proportions, that is, $p_1 = p_2 = p$, say.

Under H_0, $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0$,

and,

$$\text{var}(\hat{p}_1 - \hat{p}_2) = \frac{pq}{n_1} + \frac{pq}{n_1} = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Hence under H_0 the test statistic (7.12) becomes

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots(7.13)$$

Normally p is unknown so we use $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$, an unbiased estimate of p, in place of p. Thus, in this case the required test- statistic z under H_0 is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1). \qquad \ldots(7.14)$$

Example 7.11: Suppose that a testing procedure A results in 20 unacceptable transistors out of 100 produced, whereas another testing procedure B results in 12 unacceptable transistors out of 100 produced. Can we conclude at 5% level that the two methods are equivalents.?

Solution: Let p_1 and p_2 be the true proportions of unacceptable and let the null hypothesis be that procedure is I and procedure is II are equivalent, that is, we test the null hypothesis

$$H_0: p_1 = p_2 \text{ against the alternative } H_1: p_1 \neq p_2$$

We have,

$$\hat{p}_1 = \frac{20}{100} = 0.20, \quad \hat{p}_2 = \frac{12}{100} = 0.12$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{20 + 12}{100 + 100} = 0.16.$$

Under H_0 the test statistic z given by

$$|z| = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|0.20 - 0.12|}{\sqrt{(0.16)(0.84)\left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{0.08}{0.052} = 1.538 < 1.96.$$

Thus, it is not significant at 5% level and hence the null hypothesis may be accepted.

Example 7.12: An alternate manufacturing mechanism is being tested. Samples are taken using both the existing and the alternate mechanism so as to determine if the alternate mechanism results in an improvement. If 50 of 1000 items from the existing mechanism and 60 of 1500 items from the alternate mechanism were found to be defective, find a 90% confidence interval for the true difference of defectives between the two mechanisms. Can you conclude that alternate mechanism decreases the proportion of defectives significantly?

Solution: Let p_1 and p_2 be the true proportions of defectives in the existing and alternate mechanism respectively.

We have, $\hat{p}_1 = 50/1000 = 0.05$ and $\hat{p}_2 = 60/1500 = 0.04$. Thus

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01.$$

Also,

$$|z| = \frac{|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)|}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0, 1), \text{ refer to (7.12).}$$

From Table I, $z_{0.05} = 1.645$, therefore 90% confidence limits for $p_1 - p_2$ are

$$(\hat{p}_1 - \hat{p}_2) \pm 1.645\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

Since p_1, p_2 are unknown, using $\hat{p}_1$ and $\hat{p}_2$ unbiased estimators for p_1 and p_2, the limits obtained are

$$0.01 \pm 1.645\sqrt{\frac{(0.05)(.95)}{1000} + \frac{(0.04)(.96)}{1500}},$$

or, $0.01 \pm .0141$, that is, -0.041 to $.0241$.

Hence, the 90% confidence interval for true difference in the fraction of defectives between two mechanisms is, $-0.041 < p_1 - p_2 < .0241$. Since the interval contains the value zero, so we cannot conclude that alternate mechanism decreases the proportion of defectives being produced by existing mechanism significantly.

Example 7.13: A tea company claims that its premium tea brand outsells its normal brand by 10%. If it is found that 46 out of a sample of 200 tea-users prefer premium brand and 19 out of another independent sample of 100 tea-users prefer normal brand, test the validity of the claim made by the company.

Solution: Let p_1 and p_2 be the true proportions of the premium and normal brands, and let we set up the null hypothesis such that company's claim is valid one, that is,

$H_0 : p_1 - p_2 = 0.1$ against the alternative $H_1 : p_1 - p_2 \neq 0.1$.

We have, $n_1 = 200$, $x_1 = 46$, $\hat{p}_1 = x_1/n_1 = 46/200 = 0.23$

$n_2 = 100$, $x_2 = 19$, $\hat{p}_2 = x_2/n_2 = 19/100 = 0.19$.

Under H_0, the test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1),$$

where, $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \dfrac{46 + 19}{200 + 100} = \dfrac{65}{300} = 0.217$, and $\hat{q} = 1 - \hat{p} = 0.783$.

Thus, $$|z| = \frac{|(0.04) - (0.1)|}{\sqrt{(0.217)(0.783)\left(\frac{1}{200} + \frac{1}{100}\right)}} = \frac{0.06}{0.0505} = 1.18.$$

Since $|z| = 1.18 < 1.96$, it is not significant at 5% level of significance and hence null hypothesis may be accepted and thus company's claim may be considered to be valid one.

Example 7.14: The percentage of officials in two big PSU's with computer knowledge is 30 and 25, respectively. Is this difference likely to be hidden in samples of 1000 and 800 officials respectively from the two PSU's?

Solution: Let p_1 and p_2 be the true proportions of the officials in the two PSU's and let the null hypothesis be that difference is likely to be hidden, that is,

$H_0 : \hat{p}_1 = \hat{p}_2$ against the alternative $H_1 : \hat{p}_1 \neq \hat{p}_2$

We have, $n_1 = 1000$, $n_2 = 800$, $p_1 = 0.30$, $p_2 = 0.25$.

Under H_0, the test-statistic z given by

$$z = \frac{|p_1 - p_2|}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} = \frac{0.30 - 0.25}{\sqrt{\frac{(0.3)(0.7)}{1000} + \frac{(0.25)(0.75)}{800}}} = \frac{0.05}{0.021} = 2.37 > 1.96,$$

is significant at 5% level and hence the hypothesis is rejected and thus the difference is likely to be revealed in the samples drawn at 5% level of significance.

We observe that, since $|z| = 2.37 < 2.58$ is not significant at 1% level and so that hypothesis may be accepted at 1% and hence the samples are unlikely to reveal the difference at 1% level of significance.

7.10 SAMPLING OF VARIABLES. TESTS FOR SINGLE MEAN AND DIFFERENCE BETWEEN TWO MEANS

In this section, we discuss the sampling of the values of a variable such as height, weight, marks obtained in a test, weekly wages, etc. Each unit of the population provides a specific value (measurement) of the variable under study and the aggregate of these values forms the frequency distribution of the population. From this population, (that is, the aggregate of the values), a random sample of size n is selected to estimate and draw the conclusions about the population parameters, generally unknown.

7.10.1 Test for Single Mean

For large sample size n, sample mean is distributed normally with its mean as sampled population mean μ and S.E. as $\sigma/\sqrt{n}$, where σ is the S.D. of the sampled population. Thus, under the null hypothesis H_0, that the sample has been drawn from a population with mean μ and S.D. σ, the test statistic z given by

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

is a standard normal variate with mean zero and S.E. one, when n is large.

If $|z| < 1.96$ the deviation is not significant at 5% and the hypothesis is accepted, otherwise, it is rejected.

The limits $\bar{x} \pm 1.96\left(\sigma/\sqrt{n}\right)$ are the *95% confidence limits for the population mean μ*; and

$$\bar{x} - 1.96\left(\sigma/\sqrt{n}\right) < \mu < \bar{x} + 1.96\left(\sigma/\sqrt{n}\right)$$

is the *95% confidence interval.*

Similarly, $\bar{x} \pm 2.58\left(\sigma/\sqrt{n}\right)$ are the *99% confidence limits for the population mean μ.*

Remark: In case population S.D. σ is unknown we use s, the sample S.E. as its estimate, for $s^2 = \dfrac{n-1}{n}S^2 = \left(1 - \dfrac{1}{n}\right)S^2$ and thus for large n, $s^2 \to S^2$ and further $E(S^2) = \sigma^2$ justifies s^2 as an estimate of σ^2 in case latter is unknown.

Example 7.15: Sugar is packed in bags by an automatic machine with mean contents of bags as 1.000 kg. A random sample of 36 bags is selected and mean mass has been found to be 1.003 kg. If a S.D. of 0.01 kg, is acceptable on all the bags being packed, determine on the basis of sample test whether the machine requires adjustment.

Solution: Let the null hypothesis H_0 be that the machine does not require any adjustment, that is,

$$H_0: \mu = 1.000 \text{ kg. against } H_1: \mu \neq 1.000 \text{ kg.}$$

Under H_0, the statistic z given by

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.003 - 1.000}{.01/\sqrt{36}} = 1.8 < 1.96$$

Thus it is not significant at 5% level and hence H_0 may be accepted, that is, machine does not require any adjustment.

Example 7.16: The daily collection of milk at a plant has averaged 850 kilolitres for the last several years. An observer wants to know whether the average has changed in recent months. He randomly selects 40 days from the database and finds the average collection as $\bar{x} = 840$ kilolitres with a S.D. $s = 18$ kilolitres. Test the appropriate hypothesis at $\alpha = 0.05$.

Solution: We test the null hypothesis $H_0: \mu = 850$, against $H_1: \mu \neq 850$.

Under H_0 the test-statistic z, is given by

$$|z| = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} = \frac{|840 - 850|}{18/\sqrt{40}} = 3.51 > 1.96.$$

Thus, it is significant at 5% (even it is significant at 1% level) and hence hypothesis is rejected, that is, daily average collection of milk has changed.

Example 7.17: If e is the permissible error for estimating the population parameter μ, then prove that the minimum sample size n required for estimating μ with 95% confidence is given by $n = (1.96\sigma/e)^2$, where σ^2 is the population variance.

Solution: For large n, the test-statistic z is given by

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

Thus, $$P\left(|x - \mu| \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

We need n such that $P(|x - \mu| < e) > 0.95$.
Comparing these two, we obtain

$$\text{min. } e = \frac{1.96\sigma}{\sqrt{n}}, \text{ thus } \frac{1.96\sigma}{\sqrt{n}} \leq e, \text{ which gives, } n \geq \left(\frac{1.96\sigma}{e}\right)^2.$$

Hence, $$\text{min. } n = \left(\frac{1.96\sigma}{e}\right)^2.$$

Remark. For 99% confidence, min. $n = \left(\frac{2.58\sigma}{e}\right)^2$.

Example 7.18: The average zinc concentration recovered from a sample of zinc measurements in 40 different locations is found to be 2.54 gm per millilitre. Find the 95% confidence intervals for the mean zinc concentration in the river assuming the population S.D. to be 0.32 gm. Find the minimum sample size required at 95% confidence if the permissible error is 0.05 gm.

Solution: We have, $\bar{x} = 2.54$, $n = 40$, $\sigma = 0.32$

For large n the statistic z is given by

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

where μ is the population mean.

Thus, 95% confidence interval for μ is

$$\bar{x} - (1.96)\,\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\,\sigma/\sqrt{n}$$

or, $$2.54 - (1.96)\frac{0.32}{\sqrt{40}} < \mu < 2.54 + 1.96\frac{0.32}{\sqrt{40}},$$

which simplifies to $2.44 < \mu < 2.63$.

The minimum sample size n required at 95% confidence is

$$n = \left(\frac{1.96\sigma}{e}\right)^2 = \left(\frac{(1.96)(0.32)}{0.05}\right)^2 = 157.35 \simeq 158.$$

Example 7.19: The average monthly earnings for women in executive positions is Rs. 33,500. A random sample of $n = 40$ men in the executive positions showed average monthly earning $\bar{x}$ = Rs. 36,250, with S.E. s = Rs. 5100. Do men in the same position have average monthly earnings higher than those for women?

Solution: We test $H_0: \mu = 33{,}500$ against the right-tailed alternate $H_1: \mu > 33{,}500$.

Under H_0, the test statistic z, is given by

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim N(0, 1).$$

We have, $$z = \frac{36250 - 33500}{5100/\sqrt{40}} = 3.41 > 1.645,$$

the value of z from Table *I* at $\alpha = .05$, for right-tailed test.

Hence, calculated z is significant and thus hypothesis is rejected. Thus men in the same positions have higher salary than their females counterparts.

7.10.2 Test for Difference Between Two Means

In many situations we are concerned with the comparison of two population means. For example, our problem of concern may be the comparison of the lead levels in drinking water in two different locations of a city.

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 sample of size n_2 from a popution with mean μ_2 and variance σ_2^2. Since the two samples are independent, for large values of n_1 and n_2, $\bar{x}_1 - \bar{x}_2$ can be approximated to a normal variate with mean and variance respectively as

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2;$$

and,
$$\text{var}(\bar{x}_1 - \bar{x}_2) = \text{var}(\bar{x}_1) + \text{var}(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Thus, the statistic z given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \ldots(7.15)$$

is a standard normal variate with mean zero and S.E. one, that is, $z \sim N(0, 1)$.

Under the null hypothesis H_0: $\mu_1 = \mu_2$, that is, there is no significant difference between the two population means, the statistic (7.15) becomes

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1), \quad \ldots(7.16)$$

and hence, can be tested accordingly.

Remarks.

1. In case σ_1^2 and σ_2^2 are not known, then $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is used as an estimate of the S.E. of $(\bar{x}_1 - \bar{x}_2)$ for calculating the test statistic z; and so z becomes

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad \ldots(7.17)$$

2. If the samples have been drawn from the two populations with common variance σ^2, then under H_0 the test statistic (7.16) becomes

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \ldots(7.18)$$

When σ is unknown, then $\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$ is taken as an estimate of σ.

Example 7.20: Two random samples of 100 male students each of car-owners and non-owners of cars were drawn from a college. The grade point average for the non-owners of cars had an average equal to 2.82 with S.E. 0.63, while for the car-owners average equal to 2.43 with S.E. 0.65. Do the data present sufficient evidence to indicate a difference in the average achievements between car-owners and non-owners?

Solution: Let μ_1 and μ_2 be the true averages for non-owners of cars and car-owners respectively. We set up the null hypothesis of no-difference, that is, H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 \neq 0$. Under H_0 the test statistic z given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2.82 - 2.43}{\sqrt{\frac{(0.63)^2}{100} + \frac{(0.65)^2}{100}}} = \frac{0.39}{0.0905} = 4.30 > 3.$$

Thus, z is highly significant and hence the hypothesis is rejected. Hence, there is difference in the average achievements between car-owners and non-owners students.

Example 7.21: Two types of engines A and B were compared in respect of mileage in miles per litre of the petrol under identical conditions. The average of 50 trials in case of engine A was 34 miles and the average of 60 trials in case of engine B was 42 miles per litre. If μ_A and μ_B are population mean mileages for engines A and B respectively, find the 95% confidence interval for $\mu_B - \mu_A$ assuming that population S.D. for A and B engines are respectively 6 and 8 miles. What do you conclude about the difference in the population mean mileages from the confidence interval?

Solution: The statistic z is given by

$$z = \frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0, 1)$$

for large n_A and n_B.

Hence, the 95% confidence interval for $\mu_B - \mu_A$ is

$$(\bar{x}_B - \bar{x}_A) - 1.96\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} < \mu_B - \mu_A < (\bar{x}_B - \bar{x}_A) + 1.96\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

We have, $n_A = 50$, $\bar{x}_A = 34$, $n_B = 60$, $\bar{x}_B = 42$, $\sigma_A = 6$ and $\sigma_B = 8$.

Using these values, 95% confidence interval becomes

$$8 - 1.96\sqrt{\frac{36}{50} + \frac{64}{60}} < \mu_B - \mu_A < 8 + 1.96\sqrt{\frac{36}{50} + \frac{64}{60}}$$

or,
$$5.38 < \mu_B - \mu_A < 10.62.$$

Since the interval does not include zero, $\mu_B - \mu_A > 0$ throughout the interval so we can conclude that, at $\alpha = .05$ there is difference in the population mean mileage of the engines A and B.

Example 7.22: The mean heights in two large samples of 1000 and 2000 men are 67.5 inches and 68.0 inches, respectively. Can the two samples be regarded as drawn from the same population with S.D. 2.5 inches?

Solution: We have, $n_1 = 1000$, $\bar{x}_1 = 67.5$, $n_2 = 2000$, $\bar{x}_2 = 68.0$. Let the null hypothesis be that the samples have been drawn from the same population with S.D. 2.5 inches, that is, we test

$H_0 : \mu_1 = \mu_2$ and $\sigma = 2.5$ inches against $H_1 : \mu_1 \neq \mu_2$.

Under H_0, the test statistic z is given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1).$$

We have,

$$z = \frac{67.5 - 68.0}{2.5\sqrt{\frac{1}{2000} + \frac{1}{1000}}} = \frac{-0.5}{.0968} = -5.165.$$

Since, $|z| = 5.165 > 3$, the value is highly significant and so the null hypothesis is rejected. The two samples cannot be regarded as drawn from the same population with S.D. 2.5 inches.

Example 7.23: A random sample of 500 coins has the mean weight 28.57 gm with S.D. 1.25 gm. Another random sample of 400 coins has the mean weight 29.62 gm with S.D. of 1.42 gm. Can the samples be considered to be drawn from the same population?

Solution: We have, $n_1 = 500$, $\bar{x}_1 = 28.57$, $s_1 = 1.25$

$n_2 = 400$, $\bar{x}_2 = 29.62$, $s_2 = 1.42$.

Let the null hypothesis be that the samples have been drawn from the same population with S.D. σ, that is, we test

$H_0 : \mu_1 = \mu_2$ with S.D. σ against the alternative $H_1 : \mu_1 \neq \mu_2$.

Under H_0, the test statistic z is given by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Since, σ^2 is not given, we use

$$\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} = \frac{500(1.25)^2 + 400(1.42)^2}{500 + 400} = 1.764$$

as an estimate to σ^2.

Thus, we have

$$z = \frac{28.57 - 29.62}{1.328\sqrt{\frac{1}{500} + \frac{1}{400}}} = \frac{-1.05}{0.0891} = -11.78.$$

Since, $|z| = 11.78 > 3$, the value is highly significant, and hence the null hypothesis is rejected. Samples cannot be assumed to be drawn from the same population.

REVIEW EXERCISES

1. What is a sample? A population? Why do we go for sampling?
2. What are the various sampling plans? Explain.
3. What is a parameter? A statistic? Explain.
4. If we have several samples from the same population do they have the same sample distribution function? The same mean and variance?
5. Explain the following terms:
   (i) Null hypothesis and alternative hypothesis
   (ii) Type I and Type II errors
   (iii) Critical region
   (iv) Level of significance and power of a test.
6. Explain the terms 'standard error' and 'sampling distribution'. Show that in a series of n independent trials with constant probability p of success, the standard error of the probability of successes is $\sqrt{pq/n}$, where $q = 1 - p$.
7. If n individuals fall into one or the other two categories with probabilities p and q $(= 1 - p)$ and if the number in the two categories are x_1 and x_2 $(x_1 + x_2 = n)$, then show that covariance between x_1 and x_2 is $-npq$. Also obtain the variance of the difference $[(x_1/n) - (x_2/n)]$, between the two proportions.
8. What is hypothesis testing? Why do we test? What are the two types of errors that arise in testing?
9. What are one-tailed and two-tailed tests? Give specific examples.
10. Derive the expressions for the standard error of
    (i) The mean of a random sample of size n, and
    (ii) The difference of the means of two independent random samples of sizes n_1 and n_2.
11. State and prove Central Limit Theorem. Explain its importance in statistical inference.
12. Show that the sample mean is an unbaised estimate of the population mean but sample variance is a baised estimate of the population variance. What is the unbaised estimate of the population variance?

PROBLEM SET

1. A soft drink machine is being regulated so that the amount of drink dispensed averages 240 ml. with a S.D. 15 ml. The machine is checked periodically by taking a sample of 40 drinks and if the mean amount $\bar{x}$ for the sample taken lies within $E(\bar{x}) \pm 2$ S.E. $(\bar{x})$, the machine is certified O.K., otherwise is rectified. An apprentice from the company found the mean of 40 drinks to be 236 ml and certified O.K. Was that a reasonable decision? Justify your answer.

2. A random sample of 500 fuses was taken from a large consignment and 65 were found to be defective. Show that the percentage of defectives in the consignment almost certainly lies between 8.5 and 17.5.
3. A coin is tossed 1000 times and the head comes out 550 times. Can the deviation from expected value be due to fluctuations of sampling?
4. A large batch of electric bulbs have a mean time to failure of 800 hours and the S.D. of 60 hours. For a random sample of 64 electric bulbs, determine the probability that mean time to failure will be
   (a) less than 785 hours, (b) more than 820 hours.
5. The contents of a consignment of 1200 tins of a product have a mean mass of 5040 gm with a S.D. of 2.3 gm. Find the probability that a random sample of 40 tins drawn from the consignment will have a combined mass of
   (a) less than 20.13 kg, (b) between 20.13 kg and 20.17 kg, and (c) more than 20.17 kg.
6. It has been observed that almost 75% of customers visiting a textile exhibition prefer natural fabrics in comparison to man-made fabrics. A random sample of 200 customers has been selected and the number who like natural fabrics is recorded.
   (a) What is the approximate sampling distribution for the sample proportion $\hat{p}$?
   (b) What is the probability that the sample proportion is greater than 80%?
   (c) Within what limits the sample proportion are expected to lie about 95% of the time?
7. An insurance company has 2500 automobiles policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 83.0 lacs.
8. The ideal size of a first year student class at a particular institute is 150 students. From past experience it is known that on the average only 30% of those accepted for admission will actually attend, the institute uses a policy of approving the application of 450 students. Compute the probability that more than 150 first year students attend this institute.
9. In a survey of 100 adults over 40 years old, a total of 15 people were found to be participating in a fitness activity at least twice a week. Test the hypothesis at $\alpha = .05$ that the participation rate for adult over 40 years of age is not less than the 20% figure.
10. A sleep inducing tablet when administered to 50 insomniacs was found to be effective on 37 patients. Test the hypothesis at $\alpha = 0.05$ that tablet was effective in at least 80% cases.
11. A random sample of 500 fuses was taken from a large consignment and out of these 60 were found to be defective. Obtain the 98% confidence limits for the percentage of defective fuses in the consignment.
12. 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 daily earning of Rs. 250 or less. Estimate the almost certain limits within which such families are likely to lie.
13. In a city A, 20% of a random sample of 900 Sr. Sec. school boys had computer knowledge, while in another city B, 18.5% of a random sample of 1600 Sr. Sec. school boys had computer knowledge. Test the hypothesis that there is no difference in proportions of boys with computer knowledge among Sr. Sec. students in the two cities.

14. In two large cities out of the houses with cable connections 30% and 25%, respectively have Discovery channel. Is this difference likely to be hidden in samples of 1200 and 900 such houses from the two cities?
15. A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustment before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line. At the 0.01 level of significance, does this support the claim that the second production line does superior work?
16. An airline claims that only 6% of all lost luggage is never found. If in a random sample, 17 of 200 pieces of lost luggage are not found, test the null hypothesis $p = 0.06$ against the alternative hypothesis $p > 0.06$ at the 0.05 level of significance.
17. To compare two different types of paints, eighteen specimens are painted using type A and the drying time in hours is recorded on each. Then the same is done with type B. If $\bar{x}_A$ and $\bar{x}_B$ are the mean drying times in hours for types A and B, respectively, find $p[\bar{x}_A - \bar{x}_B > 1.0]$ under the assumption that population mean drying times for two types A and B are same with S.D. of 1.0.
18. An insurance agent claims that the average age of policy-holders who insure through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy-holders insured through him gave the following age distribution.

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

Test his claim at the 5% level on the basis of the data obtained.

19. A survey is proposed to be conducted to estimate the monthly income of the alumni of a technical institution. How large should the sample be taken in order to estimate the annual earning within plus and minus Rs. 10,000 at 95% confidence level, assuming the S.D. of the annual earnings of the entire alumni is known to be Rs. 30,000?
20. A taxi company is to decide whether to purchase brand A or brand B tires for its fleet of taxis. To estimate the difference in the two brands, an experiment is conducted using 30 tires of each brand. The tires are run until they wear out. The results are

$$\bar{x}_A = 36{,}300 \text{ kilometre} \quad \bar{x}_B = 38{,}100 \text{ kilometre}$$
$$s_A = 5{,}000 \text{ kilometre} \quad s_B = 6{,}100 \text{ kilometre}$$

Compute a 95% confidence interval for $\mu_B - \mu_A$ assuming the populations to be normal. What do you conclude from confidence interval obtained?

21. A random sample of 100 pieces was immersed in a bath for 24 hrs yielding an average of 12.2 millimetres of metal removed and a sample S.D. of 1.1 millimetre. A second sample of 200 pieces was exposed to some treatment, followed by the 24 hours immersion in the bath, resulting in an average removal of 9.1 millimetre with a sample S.D. of 0.9 millimetre. Compute a 98% confidence interval for the difference between the population means. Does the treatment appear to reduce the mean amount of metal removed?
22. The mean breaking strength of cables supplied by a manufacturer is 1800 with a S.D. 100. By a new technique in the manufacturing process, it is claimed that the breaking strength has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at $\alpha = 0.01$?

23. The electric light tubes of type *A* have a lifetime of 1400 hrs with a S.D. of 200 hrs, while of type *B* have mean lifetime of 1200 hrs with a S.D. of 100 hrs. If random samples of 125 tubes of each batch are tested, what is the probability that the type *A* tubes will have a mean lifetime which is at least, (a) 160 hrs. more than the type *B* tubes, and (b) 250 hrs. more than the type *B* tubes?

ANSWERS

1. Yes
3. No
4. (a) 0.023 (b) 0.0038 (c) 0.242.
5. 0.0179 (b) 0.740
6. (a) Approximately normal with mean 0.75 and S.E. 0.0306
   (b) 0.0516 (c) 0.69 to 0.81.
7. Only 2.3 chances out of 10,000 that total yearly claims will exceed 83 lac.
8. Only 6% of the time do more than 150 of the first 450 accepted actually attend.
9. $z = -1.25$, rejected
10. $z = -1.06$, rejected
11. (8.61, 15.38)
12. 20% to 29%
13. $z = 0.37$, accepted
14. $z = 2.5$, unlikely to be hidden
15. $z = 2.38$, second production line does superior work
16. $z = 1.489$, accepted
17. 0.0013
18. $z = -2.681$, claim may be considered to be valid
19. $n = 35$
20. (−1022, 4622) interval contains zero, cannot conclude that type *B* is superior to type *A*
21. (2.80, 3.40), yes
22. $z = 3.535$, yes
23. (a) 0.9772 (b) 0.0062.

8 Exact Sampling Distributions and Small Sample Testing

CHAPTER

8.1 INTRODUCTION

In the preceding chapter, we have discussed the large samples testing. All these tests were based on central limit theorem to justify the normality of the test statistic derived. In case the process of collecting the data is very expensive or very time-consuming, then we are unable to collect a large sample and thus the test procedures described so far in Chapter 7 are of no use. In this chapter, we introduce some equivalent statistical procedures which can be employed when the sample size is small. However, in all the statistical procedures studied here we shall assume that sample has been selected randomly and the sampled population is normally distribution.

Through Sections 8.2 to 8.5, we introduce χ^2 – variate, derive its its distribution, discuss its various properties and applications in statistical hypothesis testing, including χ^2-test of *goodness-of-fit* and χ^2-test of independence in contingency tables. The *t*-variate, its distribution, various properties and applications have been discussed in Sections 8.6 to 8.9, and in Sections 8.10 to 8.12 we consider *F*-variate its properties and applications. Finally the chapter concludes with a set of review exercises and a problem set on the topics studied.

8.2 THE CHI-SQUARE DISTRIBUTION

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 the variate

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \qquad \ldots(8.1)$$

is defined as a χ^2 (pronounced as chi-square) variate with n degrees of freedom (d.f.).

8.2.1 Derivation of the Chi-square Distribution

We apply the method of moment generating function (m.g.f.) to derive the distribution of χ^2 – variate. Rewriting (8.1) as

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

Since X_i's are independent normal variates with means μ_i and variances σ_i^2, $(i = 1, 2, \ldots, n)$, thus Z_i's are independent standard normal variates. Therefore the m.g.f. of the variate χ^2 is given by

$$M_{\chi^2}(t) = M_{\sum Z_i^2}(t) = \prod_{i=1}^{n} M_{Z_i^2}(t) = \left[M_{Z_i^2}(t)\right]^n, \qquad \ldots(8.2)$$

since Z_i's, (hence Z_i^2), are independently and identically distributed standard normal variates.

Next, by definition

$$M_{Z_i^2}(t) = E\left[e^{tZ_i^2}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz_i^2} e^{-z_i^2/2}\, dz_i, \quad \text{since } z_i \sim N(0, 1)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)z_i^2}\, dz_i$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\sqrt{\frac{1-2t}{2}}} = (1-2t)^{-1/2}. \quad \text{since} \int_{-\infty}^{\infty} e^{-a^2z^2}\, dz = \frac{\sqrt{\pi}}{a}$$

Using this in (8.2), we obtain

$$M_{\chi^2}(t) = (1-2t)^{-n/2}, \quad |2t| < 1 \qquad \ldots(8.3)$$

which is the m.g.f. of a Gamma variate with parameters $n/2$ and $1/2$, refer to (5.40), and hence, the probability that χ^2 lies in the interval $d\chi^2$ is given by

$$dP = \frac{1}{2^{n/2}\Gamma(n/2)} \left[\exp\left(-\frac{1}{2}\chi^2\right)\right] (\chi^2)^{\frac{n}{2}-1}\, d\chi^2, \quad 0 \le \chi^2 < \infty, \qquad \ldots(8.4)$$

refer to (5.38).

The distribution (8.4) is called the *χ^2-distribution with n degrees of freedom*. The number of independent variates is called the number of *degrees of freedom*. In general, if X is a χ^2-variate with n degrees of freedom, then X is written as $X \sim \chi_n^2$.

The shape of the χ^2-density curve for some specific values of $n = 1, 3$, and 10 is given in Fig. 8.1. We observe that χ^2-axis is asymptote to the curve and for $n \ge 1$, the curves are positively skewed.

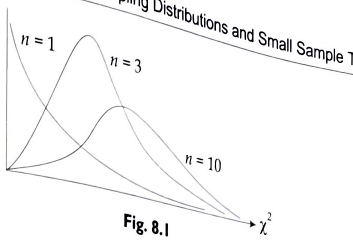

Fig. 8.1

8.3 PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

In this section, we consider some properties of the χ^2-distribution.

8.3.1 The *m.g.f.* of χ^2 Distribution

Let X be a χ^2-variate with n d.f., then its moment generating function about origin $M_0(t)$ is given by

$$M_0(t) = E\,(e^{tx})$$

$$= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{\infty} e^{tx} e^{-x/2} x^{(n/2)-1}\, dx$$

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

$$= \frac{1}{2^{n/2}\Gamma(n/2)} \frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} = (1-2t)^{-n/2}, \quad |2t| < 1 \qquad \ldots(8.5)$$

which is in confirmity with the result obtained at (8.3).

8.3.2 The *r* th ordinary moment μ_r'

It is given by

$$\mu_r' = \text{coefficient of } \frac{t^r}{r!} \text{ in expansion of } M_0(t)$$

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

$$= n\,(n+2)\,(n+4) + \ldots\,(n+2r-2). \qquad \ldots(8.6)$$

This gives mean, variance and other central moments as

$$\mu = \mu_1' = n$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = n(n+2) - n^2 = 2n$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= n\,(n+2)\,(n+4) - 3n\,(n+2)\,(n) + 2n^3$$

$$= 8n$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$
$$= n(n+2)(n+4)(n+6) - 4n(n+2)(n+4) + 6n(n+2)n^2 - 3n^4$$
$$= 48n + 12n^2.$$

Hence, the coefficients of skewness and kurtosis are

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n}, \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3 \qquad \ldots(8.7)$$

We note that as n tends to large $\beta_1 \to 0$ and $\beta_2 \to 3$, and, hence in limiting form distribution of χ^2 tends to that of normal.

In general, since $\beta_1 > 0$ and $\beta_2 > 3$ thus *χ^2-distribution is positively skewed and leptokurtic.*

8.3.3 Asymptotic Behaviour of Chi-Square Variate

The chi-square distribution tends to normal distribution as n tends to infinity

Let $X \sim \chi^2_n$, then $Z = \dfrac{X - n}{\sqrt{2n}}$ is a standard χ^2-variate.

The m.g.f. of Z is given by

$$M_Z(t) = e^{-nt/\sqrt{2n}}\left(1 - \frac{2t}{\sqrt{2n}}\right)^{-n/2}$$

This implies that

$$\ln M_z(t) = -t\sqrt{\frac{n}{2}} - \frac{n}{2}\ln\left[1 - t\sqrt{\frac{2}{n}}\right]$$

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

$$= \frac{t^2}{2} + o(n^{-1/2}) \text{ which tends to } t^2/2 \quad \text{as } n \text{ tends to } \infty.$$

Hence, $M_Z(t) \to e^{t^2/2}$ as $n \to \infty$ which is the m.g.f. of a standard normal variate, refer to (5.13). Thus, χ^2 tends to normal distribution as $n \to \infty$.

8.3.4 Additivity Property of χ^2-Variates

The sum of two independent χ^2-variates is again a χ^2-variate with d.f. as the sum of the d.f.'s of individual variates.

Let X_1 and X_2 be two independent χ^2-variates with d.f. respectively n_1 and n_2. Then their m.g.f.s are given respectively by

$$M_{X_1}(t) = (1-2t)^{-n_1/2} \quad \text{and} \quad M_{X_2}(t) = (1-2t)^{-n_2/2}; \quad |2t| < 1.$$

The m.g.f. of the sum $X_1 + X_2$ is given by

$$M_{X_1+X_2}(t) = M_{X_1}(t)\, M_{X_2}(t), \quad \text{since } X_1 \text{ and } X_2 \text{ are independent}$$
$$= (1-2t)^{-n_1/2}(1-2t)^{-n_2/2}$$

$$= (1-2t)^{-(n_1+n_2)/2},$$

which is the m.g.f. of a χ^2-variate with $(n_1 + n_2)$ d.f.

This proves the result

Remarks

1. This result can be extended to the case of n independent variates also.
2. A useful implication of this result is that, *if X_1 and X_2 are two independent non-negative variates such that $X_1 + X_2$ follows a chi-square distribution with $n_1 + n_2$ d.f. and if one of them say X_1 is a chi-square variate with n_1 d.f., then the second variable X_2 is also a chi-square variate with n_2 d.f.*
3. Since for random sampling from a normal population with mean μ and variance σ^2, $\bar{x}$ is distributed normally with mean μ and variance σ^2/n, thus $\left[\sqrt{n}\dfrac{(\bar{x}-\mu)}{\sigma}\right]^2$ is a χ^2 variate with 1 d.f.

The shape of the χ^2-density curve for some specific values of $n = 1, 3$, and 10 is given in Fig. 8.1. We observe that χ^2-axis is asymptote to the curve and for $n \geq 1$, the curves are positively skewed.

8.3.5 Ratio of Two Independent Chi-Square Variates

If X_1 and X_2 are two independent χ^2-variates with n_1 and n_2 d.f. respectively, then the ratio X_1/X_2 is a beta variate of second kind with parameters $(n_1/2, n_2/2)$.

Since X_1 and X_2 are two independent χ^2-variates with n_1 and n_2 d.f. respectively, then their joint probability differential given by the multiplication law of probability, is

$$dP(x_1, x_2) = \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \exp\{-(x_1+x_2)/2\}\,(x_1)^{\frac{n_1}{2}-1}(x_2)^{\frac{n_2}{2}-1}\,dx_1\,dx_2,\ 0 \leq x_1, x_2 < \infty$$

Introducing the transformation

$$u = x_1/x_2 \quad \text{and} \quad v = x_2$$

This gives $x_1 = uv$ and $x_2 = v$, and the Jacobian J as

$$J = \frac{\partial(x_1, x_2)}{\partial(u_1, u_2)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v.$$

Thus, the joint probability differential for the random variables $U = X_1/X_2$ and $V = X_2$ becomes

$$dP(u, v) = \frac{1}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \exp\{-(1+u)v/2\}\,(uv)^{\frac{n_1}{2}-1}(v)^{\frac{n_2}{2}-1}v\,du\,dv,\ 0 \leq u, v < \infty$$

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

Integrating it w.r.t. v over the range 0 to ∞, the marginal probability differential of the random variable U is given by

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

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

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

which is the probability distribution of a beta variate of second kind with parameters $n_1/2$ and $n_2/2$.

refer to (5.48) and hence, $U = \frac{X_1}{X_2}$ is distributed like a beta variate of second kind with parameters

$n_1/2$ and $n_2/2$.

8.4 SAMPLING DISTRIBUTION OF THE SAMPLE VARIANCE

An important result which we are in a position to prove now is about the sampling distribution of

sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ of the sample of size n drawn from a normal population

with mean μ and variance σ^2. In Section 7.5, we had considered the sampling distribution of the sample mean $\bar{x}$. Here we not only obtain the distribution of S^2 but will also establish that for a normal population, distribution of $\bar{x}$ and S^2 are independent. In fact, we prove the following theorem.

Theorem 8.1 (Joint Distribution of $\bar{x}$ and S^2) *If $x_1, x_2, \ldots, x_n$ is a sample from a normal population with mean μ and variance σ^2, then the statistics $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ are independent, $\bar{x}$ being normal with mean μ and variance σ^2/n, and $(n-1)S^2/\sigma^2$ being chi-square with $(n-1)$ degrees of freedom.*

Proof: We have,

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

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Dividing both sides by σ^2, the sample variance, we obtain

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

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{\sigma^2} + \left[\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}}\right]^2 \qquad \ldots(8.8)$$

or,

The left side of Eq. (8.8) is sum of square of n independent standard normal variates and thus is a chi-square variate with n d.f. Now, on the right hand side of (8.8), $(\bar{x} - \mu)/(\sigma/\sqrt{n})$ is a standard normal variate, so $[(\bar{x} - \mu)/(\sigma/\sqrt{n})]^2$ is a chi-square vairiate with one d.f. Thus, Eq. (8.8) equates a chi-square variate with n d.f. to the sum of two variates, one of which is a chi-square variate with one d.f. But, since the sum of two independent chi-square variates is also a chi-square variate with d.f. equal to the sum of the degrees of freedom of individual variates, thus, the two terms on the right hand side of Eq.(8.8) can be considered to be independent, with the term $\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma^2$, [$= (n-1)S^2/\sigma^2$], having a chi-square distribution with $(n-1)$ d.f., and this prove the desired result.

The sampling distribution of the sample variance finds applications in statistical hypothesis testing as we shall observe in the applications of chi-square variate.

8.5 APPLICATIONS OF CHI-SQUARE DISTRIBUTION

Chi-square distribution has wide application in statistical hypothesis testing. In this section, we study some testing procedures based on the use of chi-square distribution.

8.5.1 Chi-Square Test for Population Variance

In some practical situations the knowledge of the variance of the sampled population may be more important than the population mean. For example, our concern may be to know the precision of a measuring instrument being used, or we may be much concerned about the variation of the water level at different points during a flood.

Suppose we want to test, whether a random sample $x_1, x_2, \ldots, x_n$ has been drawn from a normal population with a specified variance σ^2. Then under the null hypothesis that the population variance is σ^2, the statistic defined by,

$$\chi^2 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{(n-1)S^2}{\sigma^2}, \qquad \ldots(8.9)$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, refer to Theorem 8.1, follows a chi-square distribution with $n-1$ (= ν, say) degrees of freedom. Thus its density function, refer to (8.4), is given by

$$\frac{1}{2^{\nu/2}\Gamma(\nu/2)}\left[\exp\left(-\frac{1}{2}\chi^2\right)\right](\chi^2)^{\frac{\nu}{2}-1} \quad 0 < \chi^2 < \infty. \qquad \ldots(8.10)$$

The distribution defined by (8.10) is called χ^2-probability distribution with $\nu = n - 1$ degrees of freedom with probability curve as shown in Fig. 8.2. The curve is skewed towards right and its shape varies with the degrees of freedom $\nu = n - 1$.

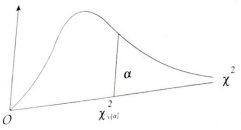

Fig. 8.2

Critical values and test of significance. Let $\chi^2_{\nu[\alpha]}$ denote the value of chi-square variate for ν d.f. such that the area to the right of this point is α, that is, $P[\chi^2 > \chi^2_{\nu[\alpha]}] = \alpha$, as shown in Fig. 8.2. The *Table* II(see, p. 342, Appendix I) gives the critical values or significant values of $\chi^2_{\nu[\alpha]}$ for the right-tailed test for different degrees of freedom ν and significant level α. We observe that value of $\chi^2_{\nu[\alpha]}$ increases with increase in ν and decrease in α. At a specific level α and df ν, the null hypothesis H_0: $\sigma^2 = \sigma_0^2$ is rejected against the alternate hypothesis

(i) $H_1: \sigma^2 > \sigma_0^2$, if calculated $\chi^2 > \chi^2_{\nu[\alpha]}$, refer to Fig. 8.2

(ii) $H_1: \sigma^2 < \sigma_0^2$, if calculated $\chi^2 < \chi^2_{\nu[1-\alpha]}$, refer to Fig. 8.3

(iii) $H_1: \sigma^2 \neq \sigma_0^2$, if calculated $\chi^2 > \chi^2_{\nu[\alpha/2]}$ or $\chi^2 < \chi^2_{\nu[1-\alpha/2]}$, refer to Fig. 8.4

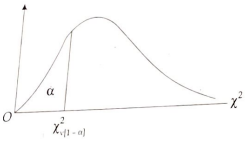

Fig. 8.3

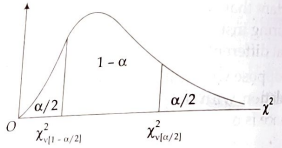

Fig. 8.4

Equal tails are used for the two-tailed χ^2 test as a matter of mathematical convenience only otherwise, chi-square distribution is not symmetric. However, normally in practice right-tailed test is applicable.

Example 8.1: A manufacturer of car batteries claims that the life of the batteries produced is approximately normally distributed with a S.D. of 0.9 years. If a random sample of 10 of these batteries has a S.D. of 1.1 years, do you think $\sigma > 0.9$ years at $\alpha = 0.05$?

Solution: We test $H_0: \sigma^2 = 0.81$ against the right-tailed alternative $H_1: \sigma^2 > 0.81$.

We have, $n = 10$, $\sigma^2 = 0.81$, $s^2 = (1.1)^2 = 1.21$

Under H_0, the test statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{ns^2}{\sigma^2} = \frac{10(1.21)}{0.81} = 14.94$$

follows χ^2 distribution with d.f. $v = n - 1 = 10 - 1 = 9$.

From Table II, $\chi^2_{9[0.5]} = 16.92$. Since χ^2 calculated is less than χ^2 tabulated so value is not significant and hence hypothesis H_0 may be accepted at 5% level of significance.

Example 8.2: Following data give the 11 measurements of the same object on the same instrument:

2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7, 2.5.

At 1% level, test the hypothesis that the variance of the instrument is no more than 0.16.

Solution: We test the null hypothesis $H_0: \sigma^2 = 0.16$ against alternative $H_1: \sigma^2 > 0.16$.

For the given data $\bar{x} = \frac{27.6}{11} = 2.51$, $\Sigma (x - \bar{x})^2 = 0.1891$.

Under H_0, the test statistic χ^2, given by

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{\Sigma (x-\bar{x})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182$$

follows χ^2 distribution with degrees of freedom $\nu = n - 1 = 11 - 1 = 10$.

From Table II, $\chi^2_{10[.01]} = 23.2$, and since the χ^2 calculated is less than the χ^2 tabulated, so hypothesis may be accepted at 1% level of significance.

8.5.2 Chi-Square Test of Goodness-of-Fit

In many random experiments, we are interested to know whether a particular probabilistic model is appropriate or not. For example, we may hypothesize that number of industrial accidents occurring monthly at a particular industrial plant follows Poisson distribution. This hypothesis can be tested by observing the number of accidents over a sequence of months; finding the theoretical or expected number of accidents on the basis of the hypothesis made and then testing whether the deviations between the observed and the expected number of accidents in each category can be attributed to fluctuations of sampling or not.

The statistical tests that determine whether a given theoretical probability distribution is appropriate in case of the random phenomena under study are called *goodness-of-fit* tests.

Suppose that $O_1, O_2 \ldots, O_k$ are the observed frequencies and $E_1, E_2, \ldots, E_k$ are the corresponding expected frequencies in the k categories on the basis of the hypothesis made. If hypothesis is correct, the observed cell frequency O_i should not be much different from the expected frequency E_i. The larger the difference, the more likely it is that the hypothesis is incorrect. The chi-square statistic to test the *goodness-of-fit* between the observed and the expected frequencies is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad \ldots(8.11)$$

where $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$, the total cell frequency.

When n is large the statistic χ^2 defined by (8.11) follows chi-square probability distribution with $\nu = k - m$ degrees of freedom, where k is the number of categories and m is the number of constraints applied to the observed data to calculate the expected frequencies.

If the theoretical cell frequencies are correct, then χ^2 is close to zero but if theoretical cell frequencies are incorrect, then χ^2 is large and thus we use right-tailed statistical test to find the significant value of χ^2 for the specified degrees of freedom ν and level of significance α.

Also to apply χ^2-test of goodness-of-fit, we pool some of the data so that no expected frequency is less than 5 and we change the degrees of freedom accordingly. This is done to avoid irregularities due to discontinuity since distribution of χ^2 is continuous but distribution of frequencies by nature is discontinuous.

Example 8.3: Suppose that a dice is tossed 120 times and the recorded data is as follow:

Face :	1	2	3	4	5	6
Observed frequency :	20	22	17	18	19	24

Test the hypothesis that the dice is unbiased at $\alpha = 0.05$.

Solution: On the basis of the null hypothesis that dice is unbiased the probability p_i for the face is $1/6$. So we test the hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_6 = 1/6.$$

Thus, expected frequencies E_i for the face i is $np_i = 120 \times 1/6 = 20$, $i = 1, 2, \ldots, 6$

Under the hypothesis H_0, the statistic $\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ follows χ^2 distribution with degrees of freedom $\nu = k - m = 6 - 1 = 5$.

We have, $$\chi^2 = \frac{0 + 4 + 9 + 4 + 1 + 16}{20} = \frac{34}{20} = 1.7.$$

From *Table* II, $\chi^2_{5[.05]} = 11.07$. Since χ^2 calculated is less than the χ^2 tabulated the hypothesis may be accepted, that is, dice may be considered to be unbiased.

Example 8.4: The proportion of blood phenotypes A, B, AB and O in a population are expected to be 0.41, 0.10, 0.04, and 0.45, respectively. To determine whether or not the actual proportions fit this set of probabilities, a random sample of size 200 is selected from this population and blood phenotypes of the units selected are recorded. The observed data is given as follows:

Phenotypes :	*A*	*B*	*AB*	*O*
No. of units :	89	18	12	81

Test the 'goodness-of-fit' of these blood phenotype proportions at $\alpha = .05$.

Solution: The hypothesis to be tested is

$H_0 : p_A = 0.41, \quad p_B = 0.10, \quad p_{AB} = 0.04, \quad p_O = 0.45.$

Under H_0, the expected cell frequencies are

$$E(A) = 200\ (0.41) = 82, \quad E(B) = 200\ (0.10) = 20$$
$$E(AB) = 200\ (0.04) = 8, \quad E(O) = 200\ (0.45) = 90$$

Thus, $$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{49}{82} + \frac{4}{20} + \frac{16}{8} + \frac{81}{90} = 3.70$$

Number of degrees of freedom = 4 – 1 = 3.

From *Table II*, $\chi^2_{3[.05]} = 7.82$. Since χ^2 calculated is less than χ^2 tabulated, thus null hypothesis H_0 may be accepted at 5% level of significance.

Example 8.5: During 400 five-minute interval the air traffic control of an airport received 0, 1, 2, ..., or 13 radio messages with respective frequencies of 3, 15, 47, 76, 68, 74, 46, 39, 15, 9, 5, 2, 0 and 1. Test at $\alpha = 0.05$, the hypothesis that the number of radio messages received during a 5 minute interval follows Poisson distribution with $\lambda = 4.6$.

Solution: Let the random variable X be the number of radio messages received during a 5-minute interval and p_x be the probability of receiving x messages. Set the null hypothesis that X follows a Poisson distribution with parameter 4.6. Thus, we test the hypothesis.

$$H_0 : p_x = e^{-4.6} \frac{(4.6)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Form the following table:

No. of radio messages (x)	*Observed frequencies (O)*	*Poisson probabilities (p_x)*	*Expected frequencies (E = 400 p_x)*
0	3	0.010	4.0
1	15	0.046	18.4
2	47	0.107	42.8
3	76	0.163	65.2
4	68	0.187	74.8
5	74	0.173	69.2
6	46	0.132	52.8
7	39	0.087	34.8
8	15	0.050	20.0
9	9	0.025	10.0
10	5	0.012	4.8
11	2	0.005	2.0
12	0	0.002	0.8
13	1	0.001	0.4

To apply χ^2-test of goodness-of-fit, since no expected frequency should be less than 5, so we pool the first two expected frequencies and the last four expected frequencies. The modified frequencies are:

Observed (O) :	18	47	76	68	74	46	39	15	9	8
Expected (E) :	22.4	42.8	65.2	74.8	69.2	52.8	34.8	20.0	10.0	8.0

Thus, $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(4.4)^2}{22.4} + \frac{(4.2)^2}{42.8} + \frac{(10.8)^2}{65.2} + \frac{(6.8)^2}{74.8} + \frac{(4.8)^2}{69.2} + \frac{(5.8)^2}{52.8} + \frac{(4.2)^2}{34.8} + \frac{(5)^2}{20} + \frac{(1)^2}{10} + \frac{(0)^2}{8}$$

$$= 6.749$$

Number of degrees of freedom = 10 – 1 = 9.

From *Table II*, $\chi^2_{9;[.05]} = 16.919$. Since χ^2 calculated is less than the χ^2 tabulated so null hypothesis may be accepted at 5% level of significance.

8.5.3 Chi-Square Test of Independence in Contingency Tables

Sometimes experimental units are classified according to two characteristics generating a bivariate data. The resulting observations are displayed in the form of a two-way table, called a *contingency table*, consisting of finite numbers of rows and columns. One characteristic varies along the rows and the second characteristic varies along the columns.

For example, a random sample of 500 employees of a PSU are classified whether they are in low, medium, or high income bracket and whether or not they favour the new salary structure announced. The data can be presented in the form of the following 2×3 contingency table.

Salary structure	Income level			Total
	Low	*Medium*	*High*	
For	91	106	98	295
Against	80	72	53	205
Total	171	178	151	500

A contingency table with r rows and c columns is referred to as an $r \times c$ table. The row and column totals are called the *marginal frequencies*.

In two categorical variable data, our interest may be to know whether or not the two characteristics are independent. The chi-square test procedure can be used to test the hypothesis for independence of two characteristics of classification. We test the null hypothesis

H_0: The two characteristics of classification are independent,

against the alternative

H_1: The two characteristics of classification are dependent.

Let O_{ij} be the observed cell frequency in row i and column j of the contingency table and if we know E_{ij} the expected cell frequency under H_0, then we can use χ^2 to compare the observed and expected frequencies.

Let p_{ij} be the probability of falling a specific observation in the ith row and jth column and if n is the total number of observations, then

$$E_{ij} = np_{ij} = n\, p_i\, q_j$$

since under hypothesis of independence, $p_{ij} = p_i q_j$, where p_i and q_j are the marginal probabilities of falling observations in the ith row and ith column respectively.

We approximate p_i with $\hat{p}_i = \frac{n_i}{n}$ and q_j with $\hat{q}_j = \frac{m_j}{n}$, where n_i, m_j respectively are the ith row and jth column totals. Thus, estimated expected cell frequencies under H_0 become

$$E_{ij} = n \frac{n_i}{n} \frac{m_j}{n} = \frac{n_i m_j}{n},$$

and the statistic χ^2 is given by

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \qquad ...(8.12)$$

This test statistic χ^2 can be shown to have an approximate χ^2 probability distribution with $(r-1)(c-1)$ degrees of freedom and the hypothesis H_0 is tested accordingly using the right-tailed test.

Example 8.6: A company operates four machines on three separate shifts daily. The following table presents the data for machine breakdowns resulted during a 6-month time period.

Shift	Machine				Total
	A	*B*	*C*	*D*	
1	10	12	6	7	35
2	10	24	9	10	53
3	13	20	7	10	50
Total	33	56	22	27	138

Test the hypothesis that for an arbitrary breakdown the machine causing the breakdown and the shift on which the breakdown occurred are independent.

Solution: Let, H_0: For an arbitrary breakdown the machine and the shift are independent.

Then the twelve expected frequencies are given by

$$E_{11} = \frac{35 \times 33}{138} = 8.37, \quad E_{12} = \frac{35 \times 56}{138} = 14.20, \quad E_{13} = \frac{35 \times 22}{138} = 5.58,$$

$$E_{14} = \frac{35 \times 27}{138} = 6.85 \quad E_{21} = \frac{53 \times 33}{138} = 12.67, \quad E_{22} = \frac{53 \times 56}{138} = 21.50,$$

$$E_{23} = \frac{53 \times 22}{138} = 8.45, \quad E_{24} = \frac{53 \times 27}{138} = 10.37, \quad E_{31} = \frac{50 \times 33}{138} = 11.96,$$

$$E_{32} = \frac{50 \times 56}{138} = 20.29, \quad E_{33} = \frac{50 \times 22}{138} = 7.79, \quad E_{34} = \frac{50 \times 27}{138} = 9.78.$$

Under H_0, the statistic χ^2 is given by

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Thus, $$\chi^2 = \frac{(10-8.37)^2}{8.37} + \frac{(12-14.20)^2}{14.20} + \frac{(6-5.58)^2}{5.58} + \frac{(7-6.85)^2}{6.85}$$

$$+ \frac{(10-12.67)^2}{12.67} + \frac{(24-21.50)^2}{21.50} + \frac{(9-8.45)^2}{8.45} + \frac{(10-10.37)^2}{10.37}$$

$$+ \frac{(13-11.96)^2}{11.96} + \frac{(20-20.29)^2}{20.29} + \frac{(7-7.97)^2}{7.97} + \frac{(10-9.78)^2}{9.78} = 1.78$$

The d.f. are $(2-1)(4-1) = 3$.

From *Table II*, $\chi^2_{3\,[.05]} = 7.8$. Since χ^2 calculated is less than χ^2 tabulated for 3 d.f. at $\alpha = .05$, hence null hypothesis may be accepted.

Example 8.7: To know about the student's response over the proposed evaluation system it was decided to select 200 1st year, 150 2nd year and 150 3rd year students from a college and record whether they are for, against or undecided for the proposed evaluation system. The observed responses are tabulated as below.

	Students			
Evaluation System	*1st year*	*2nd year*	*3rd year*	*Total*
For	82	70	62	214
Against	93	62	67	222
Undecided	25	18	21	64
Total	200	150	150	500

Test the hypothesis that the three categories of the students are homogeneous with respect to their opinions on the proposed evaluation system.

Solution: Let, H_0: Students are homogeneous with respect to their opinions on the proposed evaluation system.

Assuming homogeneity, the expected cell frequencies are:

$$E_{11} = \frac{200\times214}{500} = 85.6, \quad E_{12} = \frac{150\times214}{500} = 64.2, \quad E_{13} = \frac{150\times214}{500} = 64.2,$$

$$E_{21} = \frac{200\times222}{500} = 88.8, \quad E_{22} = \frac{150\times222}{500} = 66.6, \quad E_{23} = \frac{150\times222}{500} = 66.6,$$

$$E_{31} = \frac{200\times64}{500} = 25.6, \quad E_{32} = \frac{150\times64}{500} = 19.2, \quad E_{33} = \frac{150\times64}{500} = 19.2.$$

Thus, $$\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(3.6)^2}{85.6} + \frac{(5.8)^2}{64.2} + \frac{(2.2)^2}{64.2} + \frac{(4.2)^2}{88.8} + \frac{(4.6)^2}{66.6} + \frac{(0.4)^2}{66.6} + \frac{(0.6)^2}{25.6} + \frac{(1.2)^2}{19.2} + \frac{(1.8)^2}{19.2}$$

$= 1.53.$

Degrees of freedom $\nu = (r-1)(c-1) = (3-1)(3-1) = 4$

From *Table II*, $\chi^2_{4[0.5]} = 9.488$. Since χ^2 calculated is less than the χ^2 tabulated, hypothesis may be accepted at 5% level of significance.

Example 8.8: Two different sampling techniques were adopted while investigating the same group of students to find the number of students falling in different intelligence level. The results are tabulated as follows.

	No. of Studerits				
Techniques	*Below average*	*Average*	*Above average*	*Genius*	*Total*
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Are the sampling techniques adopted significantly different?

Solution: Let H_0: Data obtained is independent of the sampling techniques adopted.

Under H_0, the expected frequencies are

$$E_{11} = \frac{126\times200}{300} = 84, \quad E_{12} = \frac{93\times200}{300} = 62, \quad E_{13} = \frac{69\times200}{300} = 46, \quad E_{14} = \frac{12\times200}{300} = 8,$$

$$E_{21} = \frac{126\times100}{300} = 42, \quad E_{22} = \frac{93\times100}{300} = 31, \quad E_{23} = \frac{69\times100}{300} = 23, \quad E_{24} = \frac{12\times100}{300} = 4.$$

Since to apply χ^2 test no expected cell frequency should be less than 5 but here $E_{24} = 4$, so we pool E_{24} with E_{23} (or, with E_{14}) and accordingly O_{24} with O_{23} (or with O_{14}). We have,

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(86-84)^2}{84} + \frac{(60-62)^2}{62} + \frac{(44-46)^2}{46} + \frac{(10-8)^2}{8} + \frac{(40-42)^2}{42} + \frac{(33-31)^2}{31} + \frac{[(25+2)-(23+4)]^2}{23+4}$$

$= 0.92.$

Degrees of freedom $\nu = (2-1)(4-1) - 1 = 2$.

From *Table II*, $\chi^2_{2[.05]} = 5.991$. Since χ^2 calculated is less than χ^2 tabulated, thus the null hypothesis H_0 is accepted at 5% level. Hence, there is no significant difference in the sampling techniques adopted.

8.5.4 Yate's Correction for Continuity

The 2×2 contingency tables are of wide practical importance. However, in a 2×2 table there is only one degree of freedom and the frequency of only one cell can be assigned arbitrarily, but in case any of the expected frequency is less than 5, then pooling that cell frequency results in χ^2 with zero degree of freedom which is meaningless. In this case we apply a correction due to *F.* Yates known as *Yate's correction for continuity.* It consists in adding 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. The χ^2 test is applied in the resultant table without making any further correction.

Example 8.9: Two batches each of 12 experimental animals 'inoculated' and the other, 'not inoculated', were exposed to the infection of a disease. The following frequencies of dead and surviving animals were noted in the two cases, can the inoculation be regarded as effective against the disease?

Animals	*Dead*	*Survived*	*Total*
Inoculated	2	10	12
Not inoculated	8	4	12
Total	10	14	24

Solution: Let, H_0: Inoculation is not effective against the disease.

Since the cell frequencies are less than 5, applying Yate's correction for continuity the corrected table is

Animals	*Dead*	*Survived*	*Total*
Inoculated	2.5	9.5	12
Not inoculated	7.5	4.5	12
Total	10	14	24

Under H_0 the expected frequencies are

$$E_{11} = \frac{10 \times 12}{24} = 5, \qquad E_{12} = \frac{14 \times 12}{24} = 7, \qquad E_{21} = \frac{10 \times 12}{24} = 5, \qquad E_{22} = \frac{14 \times 12}{24} = 7.$$

$$\text{Hence, } \chi^2 = \frac{(2.5)^2}{5} + \frac{(2.5)^2}{7} + \frac{(2.5)^2}{5} + \frac{(2.5)^2}{7} = 4.286$$

Degrees of freedom, $\nu = (2 - 1)(2 - 1) = 1$.

From *Table II*, $\chi^2_{1[.05]} = 3.841$. Since χ^2 calculated is greater than χ^2 tabulated, thus null hypothesis is rejected at 5% level of significance, that is, inoculation may be considered to be effective against the disease.

8.6 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. In this case the statistic, $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, has normal distribution

for any sample size n, small or large. In case the population S.D. σ is unknown and the sample size n is small, the statistic $(\bar{x} - \mu)/(S/\sqrt{n})$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is no longer distributed as a normal variate and is defined as *Student's t-variate*, that is,

$$t = \frac{(\bar{x} - \mu)}{S/\sqrt{n}}. \qquad \ldots(8.13)$$

The statistic t defined as in (8.13) and its distribution, was mathematically derived by W.S. Gosset in 1908. He published his work under the pen name 'Student' and hence the distribution is known as *Student's t-distribution.*

8.6.1 Derivation of Student's *t*-Distribution

Writing (8.13) as

$$t^2 = \frac{n(\bar{x} - \mu)^2}{S^2}$$

This gives, $$\frac{t^2}{n-1} = \frac{(\bar{x} - \mu)^2}{\sigma^2/n} \frac{1}{(n-1)S^2/\sigma^2} = \frac{(\bar{x} - \mu)^2/(\sigma^2/n)}{(n-1)S^2/\sigma^2} \qquad \ldots(8.14)$$

Since 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, thus $\bar{x} \sim N(\mu, \sigma^2/n)$ and hence, $(\bar{x} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.

Further, $(\bar{x} - \mu)^2/(\sigma^2/n)$ being the square of a standard normal variable is a chi-square variate with 1 d.f. and $(n - 1)S^2/\sigma^2$ is a chi-square variate with $(n - 1)$ d.f., with $\bar{x}$ and S^2 being independently distributed. Hence, $\frac{t^2}{n-1}$, as given by (8.14), being the ratio of two independent chi-square variates with 1 and $(n - 1)$ degrees of freedom respectively, is a beta variate of second kind with parameters $\frac{1}{2}$ and $\frac{n-1}{2}$, refer to *Section 8.3.5*, and hence, its distribution is given by

$$dP = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \frac{(t^2/\nu)^{\frac{1}{2}-1}}{(1 + t^2/\nu)^{(\nu+1)/2}} d(t^2/\nu), \quad 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} dt, \quad -\infty < t < \infty \qquad \ldots(8.15)$$

where $\nu = n - 1$; and the factor 2 disappearing since the integral ranges from $-\infty < t < \infty$ should sum to unity. The distribution defined by (8.15), is known as *t-distribution with ν d.f.*

The probability density function of the *t*-variate (8.13) is given by

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}, \quad -\infty < t < \infty, \qquad \ldots(8.16)$$

where $\nu = n - 1$.

In general, if X is a *t*-variate with ν d.f., then we denote $X \sim t_\nu$.

The probability curves for *t*-variate for some specific values of $\nu = n - 1 = 2, 5$ are shown in Fig. 8.5

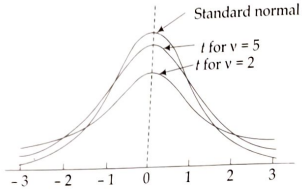

Fig. 8.5

8.7 SOME PROPERTIES OF *t*-DISTRIBUTION

In this section we consider some of the properties of *t*-distribution.

8.7.1 The *r* th Order Ordinary Moment μ'_r

Since the p.d.f. $f(t)$ of the *t*-variate as given by (8.16) contains only even power of *t*, and thus, $f(t)$ is symmetric about the line $t = 0$. Hence, *all the moments of odd order about origin are zeros*, that is,

$$\mu'_{2r+1} = 0, \quad r = 1, 2, \ldots$$

In particular, μ'_1 = mean = zero and hence, *all odd order central moments are zeros*, that is

$$\mu_{2r+1} = 0, \quad r = 1, 2, \ldots$$

Next, the moments of even order are given by

$$\mu_{2r} = \mu'_{2r} = \int_{-\infty}^{\infty} t^{2r} f(t)\,dt = 2\int_{0}^{\infty} t^{2r} f(t)\,dt$$

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

$$= \frac{\nu^r}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \int_0^\infty \frac{\left(\frac{t^2}{\nu}\right)^r \left(\frac{t^2}{\nu}\right)^{-1/2} d\left(\frac{t^2}{\nu}\right)}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}$$

$$= \frac{\nu^r}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \int_0^\infty \frac{\left(\frac{t^2}{\nu}\right)^{r+\frac{1}{2}-1}}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} d\left(\frac{t^2}{\nu}\right)$$

$$= \frac{\nu^r}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} B\left(r + \frac{1}{2}, \frac{\nu}{2} - r\right), \quad n > 2r$$

$$= \nu^r \frac{(2r-1)(2r-3)\ldots 3.1}{(\nu-2)(\nu-4)\ldots(\nu-2r)}, \quad \nu > 2r \qquad \ldots(8.17)$$

In particular

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

Hence, the *coefficients of skewness and kurtosis* are

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

As $\nu \to \infty$, $\beta_2 \to 3$, that is, *distribution of t-variate tends to normal.*

8.7.2 The Limiting Form of the *t*-Distribution

The t-distribution tends to standard normal distribution as ν, the d.f. tends to ∞. The p.d.f. for *t*-variate with ν d.f. is

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Taking limit as $\nu \to \infty$, we have

$$\lim_{\nu \to \infty} f(t) = \lim_{\nu \to \infty} \frac{1}{\sqrt{\nu}} \frac{\Gamma(\nu+1)/2}{\Gamma(1/2)\Gamma(\nu/2)} \left[\left(1 + \frac{t^2}{\nu}\right)^{\nu}\right]^{\frac{1}{2}} \left[1 + \frac{t^2}{\nu}\right]^{-\frac{1}{2}}$$

$$= \lim_{\nu\to\infty}\left[\frac{1}{\sqrt{\nu}}\frac{\Gamma(\nu+1)/2}{\Gamma(1/2)\Gamma(\nu/2)}\right].\lim_{\nu\to\infty}\left[\left(1+\frac{t^2}{\nu}\right)^{\nu}\right]^{-\frac{1}{2}}.\lim_{\nu\to\infty}\left[1+\frac{t^2}{\nu}\right]^{-\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2\pi}}\exp(-t^2/2), \quad -\infty < t < \infty, \qquad ...(8.18)$$

using the facts that large ν, $\frac{\Gamma(\nu+k)}{\Gamma(\nu)} = \nu^k$ and $\lim_{\nu\to\infty}\left(1+\frac{a}{\nu}\right)^{\nu} = e^a$ and also $\Gamma(1/2) = \sqrt{\pi}$

Since, (8.18) is the p.g.f. of a standard normal variate, and hence, the desired result follows.

8.7.3 Chief Characteristics of the Probability Curve for *t*-Statistic

The chief characteristics of the t-statistic are symmerised as follows.

1. It is mound shape and symmetric about $t = 0$, but is more flat on the top than the normal curve.

2. From Fig. 8.5, we observe that the t curve does not approach the horizontal axis as fast as standard normal curve approaches, in fact for small n, $p(|t| \geq t_0) \geq p(|z| \geq t_0)$, where $z \sim N(0, 1)$.

3. The shape of the t curve depends on the sample size n and when $n \to \infty$ the distribution of t tends to standard normal.

4. Since $f(t)$ is symmetrical about the line $t = 0$, all the odd order moments of about $t = 0$ are zeros, that is, $\mu'_{2r+1} = 0$. In particular μ, the mean is zero. Hence, the central moments, that is, the moments about mean coincide with moments about zero and so $\mu_{2r+1} = 0$.

The degrees of freedom $\nu = n - 1$ is the quantity by which $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is divided in order to obtain an unbiased estimate of σ^2 and refers to the amount of information available in the data used for estimating σ^2. For each possible value of d.f. $\nu = n - 1$, there is a different t distribution and as ν approaches large, the t-variate tends to z-variate, the standard normal variate.

In particular, for $\nu = 1$, that is, for $n = 2$, (8.16) gives

$$f(t) = \frac{1}{\beta\left(\frac{1}{2},\frac{1}{2}\right)}\frac{1}{(1+t^2)}; \quad -\infty < t < \infty$$

$$= \frac{1}{\pi(1+t^2)}; \quad -\infty < t < \infty, \qquad ...(8.19)$$

the p.d.f. of a standard *Cauchy variate.*

The t-variate has vital applications in statistical inference. Although distribution of t is asymptotically normal for large n, but for small n it considerably differs from normal. The discovery of t variate have led to many important contributions in the development of small sample theory.

Significant values of *t*: The significant values of t at level of significance α and degrees of freedom ν for two-tailed test, denoted by $t_{\nu[\alpha]}$, are given by

$$P[|t| > t_{\nu[\alpha]}] = \alpha \qquad ...(8.20)$$

or,

$$P[|t| < t_{\nu[\alpha]}] = 1 - \alpha. \qquad ...(8.21)$$

The *Table III*, (see p. 343, Appendix I), gives the values $t_{\nu[\alpha]}$ of t-distribution (two-tailed areas) for different values of α and ν, as shown in Fig. 8.6.

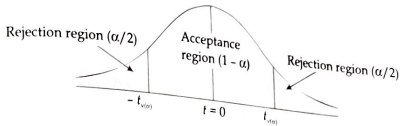

Fig. 8.6

Since the distribution is symmetrical about $t = 0$, (8.20) gives

$$P[t > t_{\nu[\alpha]}] + P[t < -t_{\nu[\alpha]}] = \alpha$$

which implies, $2P[t > t_{\nu[\alpha]}] = \alpha$, or $P[t > t_{\nu[\alpha]}] = \frac{\alpha}{2}$, or $P[t > t_{\nu[2\alpha]}] = \alpha$.

Thus, the significant values of t at level of significance α for single-tailed tests (left or right) are those of two-tailed test at level of significance 2α.

For example, from *Table III*

$t_{10[0.05]}$ for single-tailed test = $t_{10[.10]}$ for two-tailed test = 1.81

We should note that the significant values of t lead us to reliable inferences about the sampled population if the sample drawn meets the following requirements.

1. The sample must be randomly selected.
2. The sampled population must be normally distributed.

8.8 APPLICATIONS OF *t*-DISTRIBUTION

Like chi-square distribution, t-distribution is also widely applied in statistical hypothesis testing. We study some of the situations where t-variate is used.

8.8.1 Test for Single Mean

Given a random sample $x_1, x_2, \ldots x_n$ from a normal population. We need to test the hypothesis that the mean of the sampled population is μ. Under the null hypothesis H_0, the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}, \qquad ...(8.22)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ follows t-distribution with $\nu = n - 1$ *df*.

We compare the calculated value of t with the tabulated value at the desired level of significance α. If calculated $|t|$ is greater than the tabulated t, null hypothesis is rejected and if calculated $|t|$ is less than the tabulated t, the null hypothesis is accepted at the level of significance α.

If $t_{\nu[0.05]}$ is the tabulated value of t for $\nu = n - 1$ *df* at 5% level of significance, then *95% confidence limits for the population mean* μ are given by $\bar{x} \pm t_{\nu[0.05]}\left(S/\sqrt{n}\right)$. Similarly, *99% confidence limits for the population mean* μ are given by $\bar{x} \pm t_{\nu[0.01]}\left(S/\sqrt{n}\right)$.

Example 8.10: Ten individuals were chosen at random from a normal population and their heights were found to be in inches as 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. Test the hypothesis that the mean height of the population is 66 inches. Also find the 95% confidence limits for the true population mean μ.

Solution: We have, $\bar{x} = 67.8$ inches

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

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{(\Sigma d_i)^2}{n}\right], \quad d_i = x_i - 68$$

$$= \frac{1}{9}\left[82 - \frac{4}{10}\right] = 9.067$$

or, $S = 3.011$ inches.

We test the null hypotheses, H_0: $\mu = 66$ against two-tailed alternative H_1: $\mu \neq 66$.

Under H_0, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_\nu$, the t-variate with $\nu = n - 1 = 9$ df.

Here, $t = \dfrac{67.8 - 66}{3.011/\sqrt{10}} = 1.89$.

From *Table III*, $t_{9[.05]} = 2.26$. Since t calculated is less than the t tabulated hence the null hypotheses H_0 may be accepted at 5% level of significance.

Also the 95% confidence limits for the population mean μ are:

$$\mu = \bar{x} \pm t_{\nu[0.05]}\left(S/\sqrt{n}\right)$$

$$= 67.8 \pm (2.26)\left(3.011/\sqrt{10}\right) = 67.8 \pm 2.53.$$

Example 8.11: The mean weekly sales of TVs of a particular brand in company's showrooms was 14.6 TV per showroom. After announcing a few incentives the mean weekly sales in 22 stores for a typical week increased to 15.4 with S.D. of 1.7. Were the incentives announced effective in boosting the sale?

Solution: We have, $n = 22$, $\bar{x} = 15.4$, $s = 1.7$. Let the null hypothesis H_0 be that incentive announced were not effective. Thus, we test

H_0 : $\mu = 14.6$ against the right-tailed alternative H_1 : $\mu > 14.6$.

Under H_0, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_\nu$.

We have, $t = \dfrac{15.4 - 14.6}{1.7/\sqrt{21}} = \dfrac{\sqrt{21}(0.8)}{1.7} = 2.16$.

From *Table III*, $t_{21[0.05]}$ for single-tailed test = $t_{21[0.10]}$ for two-tailed test = 1.72.

Since calculated value of t is greater than the tabulated value so hypothesis H_0 is rejected at 5% level of significance, that is, incentives announced were effective.

Example 8.12: The specifications for a certain kind of ribbon call for a mean breaking strength of 180 pounds. If five randomly selected pieces of the ribbon have a mean breaking strength of 169.5 pounds with a S.D. of 5.7 pound, test the null hypothesis $\mu = 180$ against the alternative hypothesis $\mu < 180$ at 5% level of significance.

Solution: We have, $n = 5$, $\bar{x} = 169.5$, $s = 5.7$. Let test the null hypothesis be H_0 : $\mu = 180$ against the left-tailed alternative H_1 : $\mu < 180$.

Under H_0, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_\nu$.

We have, $|t| = \dfrac{|169.4 - 180|}{5.7/\sqrt{n-1}} = |-3.72| = 3.72$.

From *Table III*, $t_{4[0.05]}$ for single-tailed test = $t_{4[0.1]}$ for double-tailed test = 2.13.

Since t calculated is greater than the t tabulated, the null hypothesis H_0 is rejected at 5% level of significance.

8.8.2 Test for Difference Between Two Means

We shall consider tests of significance for following different cases:

Case I: Given two independent random samples $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ with means $\bar{x}$ and $\bar{y}$ and standard deviations s_x and s_y from normal populations with means μ_x and μ_y and with the same variances. We need to test the hypothesis that the population means are the same, that is, samples have been drawn from the same normal population.

Under the null hypothesis $H_0: \mu_1 = \mu_2$, the test statistic t given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{...(8.23)}$$

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

and, $S^2 = \frac{1}{n_1+n_2-2}[(n_1-1)S_1^2 + (n_2-1)S_2^2] = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{i=1}^{n_2}(y_i-\bar{y})^2\right]$

follows t-distribution with $\nu = (n_1 + n_2 - 2)$ degrees of freedom and thus the hypothesis H_0 can be tested accordingly.

Remark: We use this test under the assumption that the sampled populations have the same variance and use S^2 as an unbiased estimate of the common variance σ^2. In case this assumption is not justified the test becomes invalid.

Case II: When the observations are paired like (x_i, y_i), $i = 1, 2, ..., n$, refer to *Example 8.15*, then sample sizes n_1 and n_2 become same and two samples are not independent, then in this case under the null hypothesis H_0, the test statistic t given by

$$t = \frac{\bar{d}}{S/\sqrt{n}}, \quad \text{...(8.24)}$$

where, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$; $d_i = (x_i - y_i)$

follows t-distribution with $(n - 1)$ degrees of freedom.

Example 8.13: The following random samples are measurements of the heat producing capacity in millions of calories per ton of specimens of coal from two mines:

Mine I: 8,260 8,130 8,350 8,070 8,340

Mine II: 7,950 7,890 7,900 8,140 7,920 7,840.

Test at 5% level of significance whether the difference between the means of these two samples is significant.

Solution: Let μ_1, μ_2 be the two population means for Mine I and Mine II, respectively. We test the null hypothesis, $H_0: \mu_1 = \mu_2$ against the alternative $H_1: \mu_1 \neq \mu_2$.

Here,

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{41150}{5} = 8{,}230, \quad \bar{y} = \frac{1}{6}\sum_{i=1}^{6} y_i = \frac{47640}{6} = 7{,}940$$

$$S_1^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{63{,}000}{4} = 15{,}750, \quad S_2^2 = \frac{1}{5}\sum_{i=1}^{6}(y_i - \bar{y})^2 = \frac{54{,}600}{5} = 10{,}920$$

Under H_0, $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_\nu$; $\nu = n_1 + n_2 - 2$

Also, $S^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2 + \sum_{i=1}^{6}(y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{63000}{9} + \frac{54600}{9} = 13066.67$

Thus, $t = \frac{8230 - 7940}{114.31\sqrt{\frac{1}{5} + \frac{1}{6}}} = 4.19.$

From *Table III*, $t_{9[0.05]} = 2.26$. Since t calculated is greater than the t tabulated, hence hypothesis is rejected at 5% level of significance.

Example 8.14: Samples of two types of electric light bulbs were tested for length of life and following data were obtained

	Type I	Type II
Sample sizes :	$n_1 = 8$	$n_2 = 7$
Sample means :	$\bar{x}_1 = 1234$ hrs	$\bar{x}_2 = 1036$ hrs
Sample S.D. :	$s_1 = 36$ hrs	$s_2 = 40$ hrs.

Does the data support the hypothesis that Type I is superior to Type II regarding length of life ?

Solution: Let μ_1, μ_2 be the population means, we test the null hypothesis,

$H_0: \mu_1 = \mu_2$ against right-tailed alternative $H_1: \mu_1 > \mu_2$

Under H_0, $t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_\nu$; $\nu = n_1 + n_2 - 2$

Here $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{1}{13}[8(36)^2 + 7(40)^2] = \frac{1}{13}[10368 + 11200] = 1659.08.$

Thus $t = \frac{1234 - 1036}{40.73\left(\sqrt{\frac{1}{8} + \frac{1}{7}}\right)} = \frac{198}{21.08} = 9.39.$

From *Table III*, $t_{13[0.05]}$ for single-tailed test = $t_{13[0.10]}$ for two-tailed test = 1.77.

Since t calculated is greater than t tabulated, the hypothesis H_0 is rejected at 5% level of significance.

Example 8.15: The yields of two types Type I and Type II of grains in pounds per acre in 6 replications are given below. Give your comments on the difference in the mean yields.

Replication	*Type I*	*Type II*
1	205	248
2	246	263
3	230	282
4	300	308
5	304	300
6	238	220

Solution: Let the null hypothesis H_0 be that there is no difference in the mean yields of Type I and Type II. We test the null hypothesis

$H_0 : \mu_1 = \mu_2$ against the two-tailed alternative $H_1 : \mu_1 \neq \mu_2$.

Under H_0, the test statistic t in case of paired observation is,

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_\nu;\ \nu = n - 1$$

Here, $$\bar{d} = \frac{1}{6}\sum_{i=1}^{6}(x_i - y_i) = \frac{-43 - 17 - 52 - 8 + 4 + 18}{6} = -16.3$$

and, $$S^2 = \frac{1}{5}\sum_{i=1}^{6}(d_i - \bar{d})^2$$

$$= \frac{(-26.7)^2 + (0.49)^2 + (-35.7)^2 + (8.3)^2 + (20.3)^2 + (34.3)^2}{5} = 729.07.$$

Thus, $$|t| = \frac{|\bar{d}|}{S/\sqrt{n}} = \frac{|(-16.3)\sqrt{6}|}{27} = 1.48.$$

From *Table III*, $t_{5[0.05]} = 2.57$. Since t calculated is less than t tabulated hypothesis may be accepted at 5% level of significance.

Case III: Given two random samples $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ from normal populations with the same variance, we need to test the hypothesis that the population means are μ_x and μ_y respectively.

In this case under H_0, the test statistic t given by

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad ...(8.25)$$

follows t distribution with $\nu = (n_1 + n_2 - 2)$ degrees of freedom and thus we test the hypothesis accordingly.

Example 8.16: To test the claim that the resistance of electric wire can be reduced by at least 0.05 ohm by alloying, 25 measurements obtained for each alloyed wire and standard wire produced the following results:

	Mean	S.D.
Alloyed wire (x) :	0.083 ohm	0.003 ohm
Standard wire (y) :	0.136 ohm	0.002 ohm

Test the claim at 5% level of significance.

Solution: Let the claim be valid, so we test the null hypothesis

$H_0 : \mu_x - \mu_y \geq 0.05$ against the left-tailed alternative $H_1 : \mu_x - \mu_y < 0.05$.

Under H_0, $$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_\nu;\ \nu = n_1 + n_2 - 2$$

Here $$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{(25)(.003)^2 + 25(.002)^2}{48} = .0000067.$$

Therefore, $$|t| = \frac{|(0.083 - 0.136) - 0.05|}{(0.00260)\sqrt{\left(\frac{1}{25} + \frac{1}{25}\right)}} = \frac{|-.103|}{.00073} = |-144.09| = 144.09.$$

From *Table III*, $t_{48[.05]}$ for single-tailed test = $t_{48[.1]}$ for double-tailed test = 1.65.

Since $|t|$ calculated is greater than t tabulated, hypothesis is rejected at 5% level of significance.

8.8.3 Testing the Significance of an Observed Correlation Coefficient

Let r be the observed correlation coefficient in random sample of n observations (x_i, y_i) from a bivariate normal population; we need to test the hypothesis H_0 that the sampled population correlation coefficient ρ is zero.

We can show that under H_0, the test statistic t given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}} \quad ...(8.26)$$

is a t variate with $\nu = (n - 2)$ degrees of freedom and thus we test the hypothesis accordingly.

Example 8.17: A random sample of fifteen paired observations from a bivariate normal population gives a correlation coefficient of –0.5. Does this signify the existence of correlation in the sampled population?

Solution: Let the null hypothesis H_0 be that sampled population is uncorrelated. Thus, we test

$H_0 : \rho = 0$ against the alternate $H_1 : \rho \neq 0$.

Under H_0, $\quad t = r\sqrt{n-2}\Big/\sqrt{1-r^2} \sim t_\nu;\ \nu = n-2$

Here, $\quad n = 15,\ \ r = -0.5$, thus

$$|t| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.5\sqrt{13}}{\sqrt{0.75}} = 2.08$$

From *Table III*, $t_{13[.05]} = 2.16$. Since t calculated is less than the t tabulated, hence null hypothesis may be accepted at 5% level of significance.

8.9 SNEDECOR'S *F*-DISTRIBUTION

Let $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ be the two independent samples drawn from the same normal population with variance σ^2 and let

$$S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \quad \text{and} \quad S_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$$

The statistic F defined as the *variance ratio* is given by

$$F = \frac{S_x^2}{S_y^2} \qquad \ldots(8.27)$$

Writing (8.27) as

$$\frac{\nu_1 F}{\nu_2} = \frac{(n_1 - 1)S_x^2/\sigma^2}{(n_2 - 1)S_y^2/\sigma^2} \qquad \ldots(8.28)$$

where $\nu_1 = n - 1$ and $\nu_2 = n_2 - 1$.

The numerator and denominator in (8.28) are independent χ^2-variates with ν_1 and ν_2 d.f respectively, hence $\nu_1 F/\nu_2$ is distributed like beta variate of second kind with parameters $\frac{\nu_1}{2}$ and $\frac{\nu_2}{2}$, refer to *Section 8.3.5*. Thus, F follows a sampling distribution with probability density function given by

$$\frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \frac{F^{(\nu_1/2)-1}}{\left(1 + \frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}, \quad 0 < F < \infty, \qquad \ldots(8.29)$$

where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

We note that distribution of F depends upon ν_1 and ν_2, and is independent of σ^2, the population variance. The distribution defined by (8.29) is called *Snedecor's F-distribution with (ν_1, ν_2) degrees of freedom* and the variate F is denoted by $F_{(\nu_1, \nu_2)}$. Generally the greater of the two variances S_x^2 and S_y^2 is taken as numerator and ν_1 corresponds to the variance in the numerator. The probability curve for the F-distribution is shown in Fig. 8.7. The F-axis is asymptotic to the probability curve.

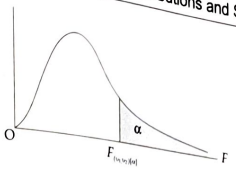

Fig. 8.7

8.10 PROPERTIES OF *F*-DISTRIBUTION

The rth order moment about origin is given by

$$\mu'_r = E(F^r)$$

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

Substituting $\frac{\nu_1 F}{\nu_2} = y$, so that $dF = \frac{\nu_2}{\nu_1}dy$, we obtain

$$\mu'_r = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2 y}{\nu_1}\right)^{r+(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy$$

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

$$= (\nu_2/\nu_1)^r \frac{B\left(\frac{\nu_1}{2} + r, \frac{\nu_2}{2} - r\right)}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}, \quad \nu_2 > 2r \qquad \ldots(8.30)$$

In particular

$$\mu'_1 = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2$$

$$\mu'_2 = \frac{(\nu_1 + 2)\nu_2^2}{(\nu_2 - 2)(\nu_2 - 4)\nu_1}, \quad \nu_2 > 4$$

Therefore,

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = \frac{(\nu_1 + 2)\nu_2^2}{(\nu_2 - 2)(\nu_2 - 4)\nu_1} - \frac{\nu_2^2}{(\nu_2 - 2)^2} = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \nu_2 > 4$$

Similarly, we can find the higher order moments. Further, for large degrees of freedom, ν_1 and ν_2, the distribution of F also tends to that of normal.

Also the mode of F-distribution is $\frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)}$, $\nu_1 > 2$. We note that in case of F-distribution while mean is always greater than 1 but mode is less than 1 and hence the distribution is positively skewed.

Another interesting property of F-variate which follows from (8.27) is that, *if F represents an F-variate with (ν_1, ν_2) d.f., then $\frac{1}{F}$ is distributed as an F-variate with (ν_2, ν_1) d.f.* That is, for an F-variate *the reciprocal relation*

$$F_{(\nu_1, \nu_2)[\alpha]} = \frac{1}{F_{(\nu_2, \nu_1)[1-\alpha]}}$$

holds.

8.11 *F*-TEST FOR THE EQUALITY OF TWO POPULATION VARIANCES

Sometimes we need to compare two population variances. For example, we may be interested to compare the precisions of the two measuring instruments, or we may be interested in the stability of measurement on a manufactured product from two assembly lines, specifically suppose we want to test whether the two independent samples $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ have been drawn from the normal populations with the same variance σ^2. Under the null hypothesis that the population variances σ_x^2 and σ_y^2 are the same, that is, $\sigma_x^2 = \sigma_y^2 = \sigma^2$, the *variance ratio statistic* F, defined by

$$F = S_x^2 / S_y^2, \qquad \ldots(8.31)$$

where, $S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$ and $S_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$

are unbiased estimates of the common population variance σ^2, follows Snedecor's F distribution with (ν_1, ν_2) d.f., where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$, with p.d.f. given by (8.29).

Critical values and test of significance. Let $F_{(\nu_1, \nu_2)[\alpha]}$ denote the value of F for (ν_1, ν_2) degrees of freedom such that the area to the right of this point is α, that is, $P[F > F_{(\nu_1, \nu_2)[\alpha]}] = \alpha$, as shown in Fig. 8.8. The *Tables IV A & B, (see p. 344-45, Appendix 1)*, give critical values or significant values of $F_{(\nu_1, \nu_2)[\alpha]}$ for the right-tailed test for different degrees of freedom (ν_1, ν_2) at significant levels $\alpha = 0.05$ and 0.01, respectively.

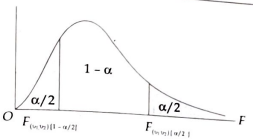

Fig. 8.8

At a specific level α and degrees of freedom (ν_1, ν_2) the null hypothesis H_0: $\sigma_x^2 = \sigma_y^2$ is tested against one of the alternate hypotheses is given by

(i) $H_1 : \sigma_x^2 > \sigma_y^2$ and $F = \frac{S_x^2}{S_y^2}$, if $F > F_{(\nu_1, \nu_2)[\alpha]}$

(ii) $H_1 : \sigma_x^2 < \sigma_y^2$ and $F = \frac{S_y^2}{S_x^2}$, if $F > F_{(\nu_2, \nu_1)[\alpha]}$

(iii) $H_1 : \sigma_x^2 \neq \sigma_y^2$ and $F = \frac{S_x^2}{S_y^2}$, if $F > F_{(\nu_1, \nu_2)[\alpha/2]}$

We must note that F-distribution is not symmetric but as in chi-square test, here also equal tails are used in two-tailed test as a matter of mathematical convenience only.

Example 8.18: There are two different choices to stimulate a certain chemical process. To test whether the variance of the yield is the same no matter which catalyst is used, a sample of 10 batches is produced using the first catalyst, and of 12 using the second. If the resulting data is $S_1^2 = 0.14$ and $S_2^2 = 0.28$, test the hypothesis of equal variance at 2% level.

Solution: We have, $n_1 = 10$, $n_2 = 12$, $S_1^2 = 0.14$, $S_2^2 = 0.28$.

We test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against two-tailed alternative H_1: $\sigma_1^2 \neq \sigma_2^2$.

Under H_0, the test statistic F is given by

$$F = \frac{S_2^2}{S_1^2} = \frac{0.28}{0.14} = 2$$

The statistic F follows F-distribution with (11, 9) degress of freedom.

From *Table IVB*,

$F_{(11, 9)[.02]}$ two-tail alternative = $F_{(11, 9)[.01]}$ at right-tailed alternative = 5.02.

Since, F calculated is less than F-tabulated, thus it is not significant and hence the null hypothesis may be accepted at 2% level of significance.

Example 8.19: With reference to the data given in *Example 8.13*, test at 10% level of significance whether it is reasonable to assume that the two population variances are the same against the alternative that they are different.

Solution: We have, $n_1 = 5,\ n_2 = 6,\ S_1^2 = 15{,}750,\ S_2^2 = 10{,}920.$

We test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against two-tailed alternative H_1: $\sigma_1^2 \neq \sigma_2^2$.

Under H_0, the variance ratio statistic F, is

$$F = \frac{S_1^2}{S_2^2} = \frac{15750}{10920} = 1.44 \sim F_{(4,5)}$$

From *Table IVA*,

$F_{(4,5)[.10]}$ for two-tailed test = $F_{(4,5)[.05]}$ for single-tailed test = 5.19.

Since, the calculated value of F is less than the tabulated value of F, thus the null hypothesis may be accepted at 10% level of significance.

Example 8.20: 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 sample come from the same normal population at 5% level of significance.

Solution: Since a normal population is specified by two parameters, mean μ and variance σ^2, thus to test that two independent samples have been drawn from the same population, we need to test (i) the equality of population means using t-test (ii) the equality of population variances using F-test.

Since t-test is applied under the assumption that population variances are the same, so first we shall test for the equality of population variances.

Here, we have

$n_1 = 10,\ n_2 = 12,\ \bar{x} = 15,\ \bar{y} = 14,\ \Sigma (x_i - \bar{x})^2 = 90,\ \Sigma(y_i - \bar{y})^2 = 108,$

$$S_1^2 = \frac{90}{9} = 10, \quad S_2^2 = \frac{108}{11} = 9.82, \text{ and } S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma (x_i - \bar{x})^2 + \Sigma (y_i - \bar{y})^2] = \frac{90 + 108}{20} = 9.9$$

We test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against the right-tailed alternative H_1: $\sigma_1^2 > \sigma_2^2$.

Under H_0, the statistic F given by

$$F = \frac{S_1^2}{S_2^2} = \frac{10}{9.82} = 1.018 \sim F_{(9,11)}$$

From *Table IV A*, $F_{(9,11)[.05]} = 2.90$. Since F calculated is less than the F tabulated, hence H_0 is accepted.

Next, since H_0: $\sigma_1^2 = \sigma_2^2$ is established, we can now apply t test for testing the null hypothesis H_0: $\mu_1 = \mu_2$ against the alternative H_1: $\mu_1 \neq \mu_2$.

Under H_0, the statistic t given by

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{15 - 14}{3.15\sqrt{\frac{1}{10} + \frac{1}{12}}} = \frac{1}{1.349} = 0.74 \sim t_{20}$$

From *Table III*, $t_{20[.05]} = 2.086$. Since t calculated is less than the t-tabulated hence the hypothesis H_0: $\mu_1 = \mu_2$ may be accepted.

Since both the null hypotheses H_0: $\sigma_1^2 = \sigma_2^2$ and H_0: $\mu_1 = \mu_2$ are accepted so samples may be considered to come from the same normal population.

REVIEW EXERCISES

1. Define χ^2-variate. Derive its p.d.f. with v degrees of freedom.
2. If X has chi-square distribution with v d.f., then find $M_X(t)$.
3. If the sum of two independent positive variables is a χ^2-variate with $(m + n)$ d.f. and if one of them is a χ^2-variate with m d.f., then show that other is a χ^2 variate n d.f.
4. Find the m.g.f. of a standard χ^2 with variate and obtain its limiting form as $\nu \to \infty$.
5. If X and Y are two independent χ^2-variates with ν_1 and ν_2 d.f. respectively, then find the distribution for
   (i) $U = X + Y$ (ii) $V = X/Y$
6. Prove that $(n - 1)S^2/\sigma^2$ is distributed like a χ^2-variate with $(n - 1)$ d.f., where
   $(n-1)S^2 = \sum_{i=1}^{n}(x_1 - \bar{x})^2$.
7. Obtain the first four central moments of χ^2-distribution. Show that χ^2-distribution is positively skewed and leptokurtic.
8. Describe χ^2-test for testing the null hypothesis that a normal population has a specified variance σ^2.
9. Describe χ^2-test of 'goodness-of-fit'.
10. What are contingency tables? Describe the χ^2-test of independence in contingency tables.
11. Write a note on Yate's correction for continuity.
12. Define Student's t-variate. Derive its distribution.
13. Show that as d.f. tends to infinity, the distribution of t tends to normal.
14. Find the coefficients of skewness and kurtosis for the t-distribution.
15. Describe the chief characteristics of the p.d.f. of a t-variate.
16. Describe t-test for (i) single mean, (ii) difference of means (iii) testing the significance of an observed sample correlation coefficient.
17. Describe paired t-test for difference of means.
18. Define Snedecor's F-variate and derive its distribution.
19. Find mean and variance of F-variate. Derive its mode, and show that the distribution is positively skewed.

20. Describe F-test for equality of two population variances.

21. Prove that $P\left[F_{(\nu_1,\nu_2)} \geq c\right] = P\left[F_{(\nu_2,\nu_1)} \leq \frac{1}{c}\right]$, where $c > 0$ is a constant.

22. If a statistic t follows Student's t-distribution with ν d.f., then show that t^2 follows F-distribution with $(1, \nu)$ d.f.

23. If in $F(\nu_1, \nu_2)$ distribution, we let $\nu_2 \to \infty$, then show that the variate $\chi^2 = \nu_1 F$ follows χ^2 distribution with ν_1 d.f.

PROBLEM SET

1. A manufacturer claims that his measuring instrument has a variability measured by S.D. $\sigma = 2$. During a test the measurements recorded are 4.1, 5.2 and 10.2. Do these data confirm or disprove his claim? Construct a 90% confidence interval to estimate the true population variance.

2. A precision instrument is guaranteed to read accurately to within 2 units. A sample of four instrument readings on the same object yield the measurements 353, 351, 351 and 355. Test the null hypothesis that $\sigma = 0.7$ against the alternative $\sigma > 0.7$ at $\alpha = .05$.

3. Playing 10 rounds of golf on his home course, a golf professional averaged 71.3 with a S.D. 1.32. Test the null hypothesis at $\alpha = .05$ that consistency of his game on his home course is actually measured by $\sigma = 1.20$ against the alternative that he is less consistent.

4. 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 the hypothesis that digits occur with equal frequency in the directory.

5. A survey of 800 families with four children each recorded 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

Test the hypothesis that male and female births are equally likely.

6. The following data give the life of 40 similar car batteries recorded to the nearest length of years

Class	:	1.5-1.9	2.0-2.4	2.5-2.9	3.0-3.4	3.5-3.9	4.0-4.4	4.5-4.9
Frequency	:	2	1	4	15	10	5	3

Test the hypothesis that the frequency distribution of battery lives may be approximated by a normal distribution with mean $\mu = 3.5$ and S.D. $\sigma = 0.7$.

7. The distribution of printing mistakes in the proof of first 392 pages of a book under publishing were found to be as follow:

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 this data and test the goodness-of-fit.

8. To determine whether there is a relationship between an employee's performance in the company's training programme and his or her ultimate success in the job, the company takes a sample of 400 cases and obtains the results shown in the following table:

Success in job	*Performance during training*			
	Below average	*Average*	*Above average*	*Total*
Poor	23	60	29	112
Average	28	79	60	167
Very good	9	49	63	121
Total	60	188	152	400

At $\alpha = 0.01$, test the null hypothesis that performace in the training program and success in the job are independent.

9. To determine the response about student's uniform in the professional colleges, a survey was conducted in four colleges in a metro. The following table gives the response of 200 students from college A, 150 students from college B and 100 students each from colleges C and D:

Response	*College*				
	A	*B*	*C*	*D*	*Total*
For	65	66	40	34	205
Against	42	30	33	42	147
Undecided	93	54	27	24	198
Total	200	150	100	100	550

Test for homogeneity of responses among the four colleges concerning student's uniform in the professional colleges.

10. A random chosen group of 20,000 non-smokers and one of 10,000 smokers were observed over a 10-year period. The following data relate the numbers of them that developed lung cancer during that period.

	Smokers	*Non-smokers*	*Total*
Lung cancer	62	14	76
No lung cancer	9,938	19,986	29,924
Total	10,000	20,000	30,000

At $\alpha = 0.01$, test the hypothesis that smoking and lung cancer are independent.

11. To study whether or not the level of earning is affected by educational attainment, a social scientist randomly selected 100 people from each of three income categories 'low', 'middle', 'high' and then recorded their educational attainment as in the following table:

Educational attainment	Income categories			Total
	Low	Middle	High	
No college	32	20	23	75
UG	13	16	1	30
PG	43	51	60	154
Doctoral	12	13	16	41
Total	100	100	100	300

Do these data indicate that the level of earning is affected by educational attainment? Test at $\alpha = 0.01$.

12. In an experiment on immunization of cattle from tuberculosis the following results were obtained:

	Affected	*Unaffected*
Inoculated	12	26
Not inoculated	16	06

Examine the effect of vaccine in controlling susceptibility to tuberculosis.

13. In an experiment on the immunization of goats from anthrax the following results were obtained:

	Died	*Survived*
Inoculated	2	10
Not inoculated	6	6

Give your conclusion on the efficiency of the vaccine.

14. A new process for producing synthetic diamonds is viable only if the average weight of the diamond is greater than 0.5 karat. The weights of the six diamonds generated are 0.46, 0.61, 0.52, 0.48, 0.57 and 0.54 karat. Test the viability of the process.

15. For a random sample of 16 observations with mean 41.5 inches and the sum of the squares of the deviation from the mean 135 (inches)^2, drawn from a normal population, find the 95% confidence limits for the population mean μ.

16. The following ten observations are from a normal population:

7.4, 7.1, 6.5, 7.5, 7.6, 6.3, 6.9, 7.7, 6.5, 7.0

(a) Find 99% one-sided confidence bound for the population mean μ.

(b) Test $H_0 : \mu = 7.5$ against $H_1 : \mu < 7.5$ at $\alpha = .01$.

(c) Do the results of part (a) support your conclusion in part (b)?

17. A certain tablet administered to each of the 12 patients resulted in the following increase in blood pressure

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

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

18. Below are given the gain in weight (in lbs.) of pigs fed on two diets *A* and *B*

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

Diet B: 44, 34, 22, 10, 47, 31, 40, 36, 32, 35, 18, 21, 35, 29, 22.

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

19. The following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety programme was put into operation

Before:	45	73	46	124	33	57	83	34	26	17
After :	36	60	44	119	35	51	77	29	24	11

At 5% level of significance test whether the safety programme was effective. Also find the 90% confidence interval for the mean improvement in lost worker-hours.

20. Measuring specimens of nylon yarn taken from two spinning machines, it was found that 7 specimens from the first machine had a mean denier of 8.62 with a S.D. of 2.8, while 9 specimens from the second machine had a mean denier of 6.38 with a S.D. of 2.4. Assuming that population sampled are normal with the same variance, test the null hypothesis $\mu_1 - \mu_2 = 1.0$ against the alternative $\mu_1 - \mu_2 > 1.0$ at $\alpha = 0.5$.

21. In a certain experiment to compare two types of animal feed *A* and *B*, the following results of increase in weights were observed in two independent samples of animals each of size 8. Test the hypothesis that food *B* is better than food *A*.

	Increase in weight in lbs							
Food A	49	53	51	52	47	50	52	53
Food B	52	55	52	53	50	54	54	53

22. A coefficient of correlation of 0.2 is obtained from a random sample of 625 pairs of observations.

(i) Is this value of r significant?

(ii) Obtain the 95% confidence limits to the correlation coefficient in population; use that when n is large t-variate is distributed like a standard normal variate.

23. Find the least value of r in a sample of 18 pairs of observations from a bivariate population significant at 5% level.

24. Following data gives the amounts of sulphur monoxide recorded by two instruments *A* and *B* in the atmosphere. Assuming the populations of measurements to be normal, test the hypothesis H_0: $\sigma_A = \sigma_B$ against H_1: $\sigma_A \neq \sigma_B$ at $\alpha = .02$,

Instruments	*Amounts of sulphur monoxide*								
A	0.86	0.82	0.75	0.61	0.89	0.64	0.81	0.68	0.65
B	0.87	0.74	0.63	0.55	0.76	0.70	0.69	0.57	0.53.

25. The following are the values in thousands of an inch obtained by two technicians in 10 successive measurements with the same micrometer. Is one technician significantly more consistent than the other at $\alpha = 0.05$?

Technician A :	503	505	497	505	495	502	499	493	510	501
Technician B :	502	497	492	498	499	495	497	496	498.	

26. The nicotine contents in milligrams of two samples of tobacco were found to be as follows:

Sample A :	24	27	26	21	25	
Sample B :	27	30	28	31	22	36

Can it be claimed that two samples come from the same normal population?

27. For the two samples
105 108 86 103 103 107 124 105 and 89 92 84 97 103 107 111 97
giving the relative output of tin plate workers under two different working conditions, test the hypothesis at $\alpha = .05$, H_0: $\sigma_1^2 = \sigma_2^2$ against the alternative H_1: $\sigma_1^2 > \sigma_2^2$ assuming the two populations to be normal.

ANSWERS

1. $\chi^2 = 5.29$, claim accepted, (3.53, 206.07)
2. $\chi^2 = 22.45$, reject H_0
3. $\chi^2 = 10.89$, H_0 accepted
4. $\chi^2 = 58.542$, H_0 rejected
5. $\chi^2 = 19.63$, H_0 rejected
6. $\chi^2 = 3.05$, H_0 accepted
7. $\chi^2 = 40.937$, H_0 rejected
8. $\chi^2 = 20.179$, H_0 rejected
9. $\chi^2 = 31.17$, Non-homogeneous response
10. $\chi^2 = 79.83$, H_0 rejected
11. $\chi^2 = 19.172$, yes!
12. Vaccine is effective
13. Vaccine is efficient.
14. $t = 1.32$, not viable
15. 41.5 ± 1.6
16. (a) 7.496 (b) $t = -2.849$, reject H_0 (c) yes
17. $t = 2.89$, yes
18. $t = -0.609$; don't differ significantly
19. $t = 4.03$; program is effective, [4.0, 6.4]
20. $t = 0.95$; accepted
21. (i) $t = 2.17$; B is superior (ii) $t = 5.09$; significant (iii) 0.2 ± 0.075
23. 0.47.
24. $F = 1.15$, H_0 accepted
25. $F = 2.4$, No
26. H_0 : $\mu_1 = \mu_2$; $t = 1.9$, not significant
H_0 : $\sigma_1^2 = \sigma_1^2$; $F = 4.08$, not significant, yes
27. $F = 1.26$ accepted.

9 CHAPTER Analysis of Variance

9.1 INTRODUCTION

In the preceding two chapters, we have seen that how large and small sample tests are applied to test the significance of the difference between two sample means of two independent populations. Sometimes we encounter situations in which we wish to study the differences among three or more independent sample means rather than just two, and so, we need an alternative procedure to test the hypothesis that all the samples are drawn from the same population, that is, they have the same mean. For example, say three different types of gaseous fuels are being used to run five vehicles each and the mileage per unit of the fuel covered by each vehicle is given. We may be interested in finding out whether the average mileages given by the three fuels are significantly different, or in other words, whether the samples have come from the same normal population. Thus, here aim is to test the homogeneity of more than two means. In such a situation, a technique known as *Analysis of Variance* (*ANOVA*) developed by Prof. R.A. Fisher is applied. It consists of splitting the total variation into component variations attributed to independent factors of interest to the experimenter and variation due to the experimental error. If the experiment has been properly designed, then these component variations can be used to study the effects of the various factors on the variable of interset.

In Section 9.2, we assume the samples, (not necessarily of equal sizes), are from k distinct populations and we want to test on the basis of the data given that the k population means are equal. Since the mean depends only on a single factor so this forms a *one-way analysis of variance*. In Section 9.3, we consider a model that assumes that there are two factors which effect the mean value of a variable and the model is called *two-way analysis of variance*. In Section 9.4, we consider a model where the response is affected by three factors and consider a specific design, called the *Latin Square Design (LSD)* to analyse it. The chapter ends with review exercises and a problem set.

In all of the models considered in this chapter, we assume that the observations obtained are independent and are normally distributed with the same (though unknown) variance σ^2. Further, the various treatments and environment effects are additive in nature and do not interact with each other.

9.2 ONE-WAY ANALYSIS OF VARIANCE

In this experimental design, independent random samples of sizes $n_1, n_2, ..., n_k$ are selected from k normal populations with means $\mu_1, \mu_2 ..., \mu_k$ and common variance σ^2. The k different populations are classified on the basis of a *single criterion* only such as different treatments or groups, e.g. different regions of a country, different seeds or different machines, etc. The null hypothesis to be tested is

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

against the alternative

$$H_1: \text{At least two of the means are not equal.}$$

That is, we will be testing the null hypothesis that all the population means are equal against the alternative that at least two of them are different.

The complete data of $N = \sum_{i=1}^{k} n_i$ observations can be arranged as in Table 9.1 given below.

Table 9.1

Treatment					Total	Mean
1	x_{11}	x_{12}	–	x_{1n_1}	T_1	$\bar{x}_{1.}$
2	x_{21}	x_{22}	–	x_{2n_2}	T_2	$\bar{x}_{2.}$
⋮	⋮	⋮		⋮	⋮	⋮
i	x_{i1}	x_{i2}	–	x_{in}	T_i	$\bar{x}_{i.}$
⋮	⋮	⋮		⋮	⋮	⋮
k	x_{k1}	x_{k2}	–	x_{kn_k}	T_k	$\bar{x}_{n.}$

9.2.1 Mathematical Model for One-Way ANOVA

Under linear mathematical model, each observation x_{ij} may be written in the form

$$x_{ij} = \mu_i + \varepsilon_{ij} \qquad ...(9.1)$$

where ε_{ij} measures the deviation of the jth observation of the ith sample from the corresponding treatment mean μ_i and represents the error effect due to random causes.

The expression thus can be rewritten as

$$x_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$$
$$\mu + \alpha_i + \varepsilon_{ij}, \qquad ...(9.2)$$

where, $\mu = \frac{1}{N}\sum_{i=i}^{k}(n_i\mu_i)$ represents the general mean effect and μ_i is the fixed effect due to the ith treatment and $\alpha_i = \mu_i - \mu$ is the deviation of the mean of the i th sample from the general mean effect.

Obviously, $$\sum_{i=i}^{k} n_i\alpha_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = \sum_{i=1}^{k} n_i\mu_i - \mu\sum_{i=1}^{k} n_i$$

$$= N\mu - \mu N = 0$$

The null hypothesis that the k population means are equal against the alternative that at least two of the means are unequal may now be replaced by the equivalent hypothesis,

$$H_0: \alpha_1 = \alpha_2 = ... = \alpha_k = 0$$

H_1: At least one of α_i's is not equal to zero.

9.2.2 Statistical Analysis of One-Way ANOVA

Let x_{ij} be the j th observation ($j = 1, 2, ..., n_i$) of the ith sample ($i = 1, 2, ..., k$). The total variation in the experiment, measured by the quantity *total sum of squares* (*TSS*), is given by

$$TSS = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2, \qquad ...(9.3)$$

where $N = n_1 + n_2 + ... + n_k$ is the total number of observations and $\bar{x}_{..} = \frac{1}{N}\sum_i \sum_j x_{ij}$.

This total variation among x_{ij}'s may be attributed to the following two components.

The first component of variation is the variation *between k sample means* (*SSB*) and is given by

$$SSB = \sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2 \qquad ...(9.4)$$

where $\bar{x}_{i.} = \frac{1}{n_i}\sum_j x_{ij}$ is the mean of the sample from the ith population.

The second component of variation is the variation *within the k samples* (*SSW*) and is given by

$$SSW = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 \qquad ...(9.5)$$

However, the three sums given by (9.3), (9.4) and (9.5) are not independent and it can be shown that

$$TSS = SSB + SSW;$$

hence we need to calculate only two of the these three variations.

The degrees of freedom (*df*) for *TSS* are (N – 1), since it involves N squared observations and one degree of freedom is lost for the mean estimated. Similarly, the degrees of freedom for *SSB* are i -1 and the degrees of freedom for *SSW* are

$$\sum_{i=1}^{k}(n_i - 1) = N - k.$$

We observe that

$$df(TSS) = df(SSB) + df(SSW),$$

and hence, the degrees of freedom are additive. Further each mean square (*MS*), obtained by dividing each sum of square by its respective *df*, that is,

$$MSS = \frac{TSS}{N-1}, \quad MSB = \frac{SSB}{K-1}, \quad \text{and} \quad MSW = \frac{SSW}{N-K}$$

provide an unbiased estimate of the *population variance*, say σ^2.

Now since by the assumption all x_{ij}'s are drawn from a normal population with mean μ and variance σ^2 thus, $\frac{1}{\sigma^2}\sum_i\sum_j (x_{ij} - \bar{x}_{..})^2$ is distributed like a chi-square variate with $N-1$ degrees of freedom, that is, $\frac{1}{\sigma^2}\sum_i\sum_j (x_{ij} - \bar{x}_{..})^2 \sim \chi^2_{N-1}$, refer to *Theorem 8.1*

Similarly,

$$\frac{1}{\sigma^2}\sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2 \sim \chi^2_{k-1}, \quad \text{and} \quad \frac{1}{\sigma^2}\sum_i\sum_j (x_{ij} - \bar{x}_{i.})^2 \sim \chi^2_{N-k}$$

The ratio of the mean square between the class means (MSB) to the mean square within the classes (MSW) can be tested using the F-test, where the variance ratio given by

$$F = \frac{\sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_i\sum_i (x_{ij} - \bar{x}_{i.})^2} \cdot \frac{N-k}{k-1} \qquad ...(9.6)$$

is distributed like a F-variate with $v_1 = k-1$ and $v_2 = N-k$ degrees of freedom.

The ANOVA table for the one-way classification is

Source of variation	*df*	*SS*	*MS*	*F*
Between classes	$k-1$	SSB	$MSTR = MSB = SSB/(k-1)$	$\frac{MSTR}{MSE} = \frac{MSB}{MSW} \sim F_{(k-1,\,N-k)}$
Within classes	$N-k$	SSW	$MSE = MSW = SSW/(N-k)$	
Total	$N-1$	TSS		

In case the tabulated value of F for $v_1 = (k-1)$ and $v_2 = (N-k)$ degrees of freedom at the specified level α is less than the calculated value of F, then the hypothesis is accpeted at level α, otherwise, it is rejected.

Remarks:

1. Since all the three sum of squares are independent of change of origin and hence origin may be shifted to any convenient point.
2. The TSS in (9.3) can be expressed as

$$TSS = \sum_i\sum_j x_{ij}^2 - \frac{G^2}{N}, \qquad ...(9.7)$$

where $G = \sum_i\sum_j x_{ij}$ is the grand total of all the N observations.

3. The SSB in (9.4) may be expressed as

$$SSB = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}, \qquad ...(9.8)$$

where $T_i = \sum_j x_{ij}$ are the row totals.

4. The underlying experimental design in case of one-way classification is known as *completely randomized design* (CRD).

Example 9.1: The three drying techniques for curing a glue were studied and the following times were observed

Formula A:	13	10	8	11	8	
Formula B:	13	11	14	14		
Formula C:	4	1	3	4	2	4

At $\alpha = 0.01$, test the hypothesis that the average times for the three formulae are same.

Solution: We form the following table:

Techniques	*Drying Times*						$T_i = \sum_j x_{ij}$	$\sum_j x_{ij}^2$
I	13	10	8	11	8		50	518
II	13	11	14	14			52	682
III	4	1	3	4	2	4	18	62
Total							120	1262

Here, $k = 3$, $n_1 = 5$, $n_2 = 4$, $n_3 = 6$, and thus $N = 15$.

Further, $G = \sum_i\sum_j x_{ij} = 120$ and $\sum_i\sum_j x_{ij}^2 = 1262$. Thus,

$$TSS = \sum_i\sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 1262 - \frac{(120)^2}{15}$$

$$= 1262 - 960 = 302$$

$$SSB = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{(50)^2}{5} + \frac{(52)^2}{4} + \frac{(18)^2}{6} - \frac{(120)^2}{15}$$

$$= 500 + 676 + 54 - 960 = 270$$

and, $SSW = TSS - SSB$

$$= 302 - 270 = 32$$

The ANOVA table is

Source of variation	df	SS	MS	F
Between classes	2	270	135	50.56
Within classes	12	32	2.67	
Total	14	302		

From *Table IV B*, $F_{(2,\,12)[.01]} = 6.93$. Since the value of the test statistic calculated is greater than the value tabulated, the hypothesis that average drying timings are equal in case of the three techniques is rejected.

Example 9.2: Following data represents the total mileages obtained by the vehicles of the same type run on three different gas fuels

Gas 1:	220	251	226	246	260
Gas 2:	244	235	232	242	225
Gas 3:	252	272	250	238	256

At $\alpha = 0.05$, test the hypothesis that the average mileage obtained is not affected by the type of the gas used.

Solution: Shifting the origin to the point 220, we form the following table:

Gas	Mileage					$T_i = \sum_j x_{ij}$	$\sum_j x_{ij}^2$
1	0	31	6	26	40	103	3273
2	24	15	12	22	05	78	1454
3	32	52	30	18	36	168	6248
Total						349	10975

Here, $k = 3$, $n_1 = n_2 = n_3 = 5$, and thus $N = 15$

Further, $G = \sum_i \sum_j x_{ij} = 349$, and $\sum_i \sum_j x_{ij}^2 = 10975$.

Thus, $$TSS = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 10975 - \frac{(349)^2}{15}$$

$$= 10975 - 8120.07$$

$$= 2854.93$$

$$SSB = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{(103)^2 + (78)^2 + (168)^2}{5} - \frac{(349)^2}{15}$$

$$= 8983.40 - 8120.07$$

$$= 863.33$$

and, $SSW = TSS - SSB$

$$= 2854.93 - 863.33$$

$$= 1991.60$$

The ANOVA table is

Source of variation	df	SS	MS	F
Between classes	2	863.33	431.67	2.74
Within classes	12	1991.60	157.72	
Total	14	2854.93		

From *Table IV A*, $F_{(2,\,12)[.05]} = 3.88$. Since the value of the test statistic calculated does not exceed the tabulated value, thus the hypothesis that the average mileage is not affected by the type of gas used may be accepted.

9.3 TWO-WAY ANALYSIS OF VARIANCE

One-way analysis of variance is employed when the experimental units are homogeneous in respect to their configuration and there is only one factor which might influence the response and any other variation in the response is due to random chances or experimental errors. But sometimes the units under study are not homogeneous and are likely to add their own variability to the response. To isolate this source of variation units are divided into relatively homogeneous *classes* (blocks) and one unit within each class is randomly subjected to one specific *factor* (*treatment*). Thus the response of each experimental unit is affected by two considerations, one, because of block, and second, because of treatment, and the analysis is called *two-way analysis of variance*. The complete data can be arranged in m rows and n columns as in Table 9.2, given below and ANOVA is employed to test the independence of row or/and column factor levels.

Table 9.2

Row ↓ \ Column →	1	2	...	j	...	n	Mean	Total (R_i)
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}	$\bar{x}_{1\cdot}$	R_1
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}	$\bar{x}_{2\cdot}$	R_2
⋮	⋮	⋮		⋮		⋮	⋮	⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{in}	$\bar{x}_{i\cdot}$	R_i
⋮	⋮	⋮		⋮		⋮	⋮	⋮
m	x_{m1}	x_{m2}	...	x_{mj}	...	x_{mn}	$\bar{x}_{m\cdot}$	R_m
Mean	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	...	$\bar{x}_{\cdot j}$	...	$\bar{x}_{\cdot n}$	$\bar{x}$	
Total (C_j)	C_1	C_2	...	C_j	...	C_n		G

9.3.1 Mathematical Model for Two-Way ANOVA

Under linear mathematical model each observation x_{ij} may be given in the form

$$x_{ij} = \mu_{ij} + \varepsilon_{ij}, \qquad ...(9.9)$$

where μ_{ij} represents the effect due to the factors assigned in the ith row and jth column and ε_{ij} represents the error effect due to random causes.

Writing, $\mu = \frac{1}{N}\sum_i\sum_j \mu_{ij}$; as the general mean effect

$\mu_i = \frac{1}{n}\sum_{j=1}^{n} \mu_{ij}$; mean effect due to the factor in the ith row.

$\mu_j = \frac{1}{m}\sum_{i=1}^{m} \mu_{ij}$; mean effect due to the factor in the jth column.

and, taking

$$\alpha_i = \mu_i - \mu, \quad \beta_j = \mu_j - \mu, \quad \text{and } \gamma_{ij} = (\mu_{ij} - \mu_i - \mu_j + \mu)$$

as the interaction effects. We can rewrite (9.9) as

$$x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \qquad ...(9.10)$$

with $\sum_{i=1}^{m}\alpha_i = \sum_{j=1}^{n}\beta_j = 0$, and $\sum_{i=1}^{m}\gamma_{ij} = \sum_{j=1}^{n}\gamma_{ij} = 0$

In case of one observation per cell, taking the interaction effect $\gamma_{ij} = 0$, the model (9.10) becomes

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad ...(9.11)$$

Thus, the null hypothesis to be tested in the case of two-way ANOVA are

$$H_0: \alpha_1 = \alpha_2 = ... = \alpha_m$$

and/or

$$H_0: \beta_1 = \beta_2 = ... = \beta_n$$

9.3.2 Statistical Analysis of Two-Way ANOVA

The total variation in the experiment measured by the total sum of square (TSS) is

$$TSS = \sum_i\sum_j (x_{ij} - \bar{x}_{..})^2, \qquad ...(9.12)$$

where $\bar{x}_{..} = \frac{1}{N}\sum_i\sum_j x_{ij}$ is the mean response.

The total variation among x_{ij}'s in this case is attributed to the following three components:

1. The variation *between the rows (SSR)*, given by

$$SSR = n\sum_{i=1}^{m}(\bar{x}_{i\cdot} - \bar{x}_{..})^2, \text{ where } \bar{x}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad ...(9.13)$$

2. The variation *between the columns (SSC)*, given by

$$SSC = m\sum_{j=1}^{n}(\bar{x}_{j} - \bar{x}_{..})^2, \text{ where } \bar{x}_{j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \qquad ...(9.14)$$

3. The variation due to *random chances* or *experimental errors (SSE)*, given by

$$SSE = TSS - SSR - SSC \qquad ...(9.15)$$

The degrees of freedom for TSS are $mn - 1$, for SSR are $m - 1$, for SSC are $n - 1$, and finally since the degrees of freedom are additive, therefore, the degrees of freedom for SSE are

$$(mn - 1) - (m - 1) - (n - 1) = (m - 1)(n - 1).$$

Further, each mean square (MS) obtained by dividing a sum of square with its df provides an unbiased estimate of the population variance as in case of one-way ANOVA. Thus, *the ANOVA table for the two-way classification* is

Source of variation	df	SS	MS	F
Row	$m - 1$	SSR	$MSR = SSR/(m-1)$	$F_R = MSR/MSE \sim F_{((m-1),(m-1)(n-1))}$
Column	$n - 1$	SSC	$MSC = SSC/(n-1)$	$F_C = MSC/MSE \sim F_{((n-1),(m-1)(n-1))}$
Errors	$(m-1)(n-1)$	SSE	$MSE = SSE/(m-1)(n-1)$	
Total	$mn - 1$	TSS		

In case the tabulated value of F for $((m - 1), (m - 1)(n - 1))$ degrees of freedeom at level α is greater than the calculated $F_R = MSR/MSF$, then there is no significant difference between the row factor levels. Similarly by comparing $F_C = MSC/MSE$ with the corresponding tabulated value, we can draw conclusion about column's factor levels.

Remarks:

1. To simplify the calculation various sum of squares can be expressed as follows.

(a) $TSS = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$, ...(9.16)

(b) $SSR = \sum_{i=1}^{m} \frac{R_i^2}{n} - \frac{G^2}{N}$, ...(9.17)

(c) $SSC = \sum_{j=1}^{n} \frac{C_j^2}{m} - \frac{G^2}{N}$, ...(9.18)

where, $G = \sum_i \sum_j x_{ij}$ is the grand total of $N = mn$ observations, $R_i = \sum_{j=1}^{n} x_{ij}$, $i = 1, 2, ..., m$ are the row totals, and $C_j = \sum_{i=1}^{m} x_{ij}$, $j = 1, 2, ... n$ are the column totals.

2. The underlining experimental design in case of two-way classification is known as *randomized block design (RBD)*.

Example 9.3: In an experiment to study the performance of 4 different detergents for cleaning fuel injectors of 3 different models of engines following data was obtained

Detergent	Engine		
	1	2	3
A	45	43	51
B	47	46	52
C	48	50	55
D	42	37	49

Obtain the ANOVA table and test at $\alpha = .05$ whether there are difference in the detergent or in the engines.

Solution: Shifting the origin at the point 45, we obtain the table as

Detergent	Engine			Row total (R_i)
	1	2	3	
A	0	–2	6	4
B	2	1	7	10
C	3	5	10	18
D	–3	–8	4	–7
Col. total (C_j)	2	–4	27	25

Here $m = 4$, $n = 3$, $N = 12$, and $G = 25$. Further

$\sum\sum x_{ij}^2 = 0 + 4 + 36 + 4 + 1 + 49 + 9 + 25 + 100 + 9 + 64 + 16 = 317$

Thus, $TSS = 317 - \frac{(25)^2}{12} = 317 - 52.08 = 264.92$,

$SSR = \frac{16 + 100 + 324 + 49}{3} - \frac{(25)^2}{12} = 163 - 52.08 = 110.92$,

$SSC = \frac{4 + 16 + 729}{4} - \frac{(25)^2}{12} = 187.25 - 52.08 = 135.17$,

and, $SSE = 264.92 - 110.92 - 135.17 = 18.83$.

The ANOVA table is

Source of variation	df	SS	MS	F
Row (Detergent)	3	110.92	MSR = 36.91	$F_R = 11.75$
Col. (Engine)	2	135.17	MSC = 67.58	$F_C = 22.52$
Error	6	18.83	MSE = 3.14	
Total	11	264.92		

From *Table IV A*, $F_{(3, 6)[.05]} = 4.76$. Since F_R calculated exceeds the F tabulated so we conclude that there are differences in the effectiveness of the 4 detergents.

Similarly, from *Table IV A*, $F_{(2, 6)[.05]} = 5.14$, and thus F_C calculated exceeds the F tabulated, therefore, there are differences due to engines also.

Example 9.4: Following data gives the monthly phone costs of four different companies at three different usage levels. Test the hypothesis that there is no difference among the companies by taking $\alpha = .05$

Usage Level	Company			
	A	B	C	D
Low	27	24	31	23
Middle	68	76	65	67
High	308	326	312	300

Solution: The table is

Usage Level	Company				Row total (R_i)
	A	B	C	D	
Low	27	24	31	23	105
Middle	68	76	65	67	276
High	308	326	312	300	1246
Col. total (C_j)	403	426	408	390	1627

Here $m = 3$, $n = 4$, $N = 12$, and $G = 1627$. Further

$$\sum\sum x_{ij}^2 = (27)^2 + (24)^2 + \ldots \ldots + (300)^2$$

$$= 410393$$

Thus, $$TSS = 410393 - \frac{(1627)^2}{12} = 410393 - 220594.08 = 189798.92$$

$$SSR = \frac{(105)^2 + (276)^2 + (1246)^2}{4} - \frac{(1627)^2}{12}$$

$$= 409929.25 - 220594.08 = 189335.17$$

$$SSC = \frac{(403)^2 + (426)^2 + (408)^2 + (390)^2}{3} - \frac{(1627)^2}{12}$$

$$= 220816.33 - 220594.08 = 222.25$$

and, $SSE = TSS - SSR - SSC = 189798.92 - 189335.17 - 222.25 = 241.5$.

Therefore, $$MSC = \frac{222.25}{3} = 74.08$$

$$MSE = \frac{241.5}{6} = 40.25$$

Thus, $$F_C = \frac{MSC}{MSE} = \frac{74.08}{40.25} = 1.84$$

From *Table IVA*, $F_{(3,\ 6)[.05]} = 4.76$. Since tabulated value of F is greater than the value F_c calculated, thus there is no significance difference among the companies, hence the hypothesis may be accepted.

9.4 THREE-WAY ANALYSIS OF VARIANCE. LATIN SQUARE DESIGN

In two-way analysis of variance experimental units are divided into relatively homogeneous blocks and one unit within each block is randomly subjected to one specific treatment. Next, consider a three-way criteria classification problem, say we are interested in the yield of 4 varieties of wheat using 4 different fertilizers over a period of 4 years. The total number of experimental units for a complete three-way ANOVA would by 64. However, by selecting the same number of categories for all three criteria of classification we may consider a *Latin Square Design (LSD)* and can perform the ANOVA using the outputs of only 16 experimental units. A typical Latin square, selected at random from all possible 4×4 squares is the following

D	B	C	A
A	C	B	D
C	D	A	B
B	A	D	C

The four letters A, B, C, and D represent the 4 varieties of wheat. The rows represent the 4 fertilizers and the columns represent the 4 years. These are the two sources of variations which we intend to control. We observe that each letter appears once in each row and once in each column. With such a design the ANOVA enables us to separate the variation due to the different rows (fertilizers) and different columns (years) from the error some of squares and to obtain a test for differences in the yielding capabilities of the 4 varieties of wheat.

9.4.1 Statistical Analysis of an $m \times m$ Latin Square Design

Consider an $m \times m$ Latin square and let x_{ijk} be the response from the experimental unit in the ith row, j th column and k th treatment (letter). The triplet (i, j, k) assumes only m^2 of the possible m^3 values of Latin square selected by the experimenter. If α_i and β_j are the specific effects of the i th row and j th column, γ_k the effect of the kth treatment (letter), μ the grandmean, and ε_{ijk} the random error, then the linear mathematical model can be expressed as

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

Under the constraints that ...(9.19)

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_j \gamma_k = 0,$$

the hypotheses to be tested are

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0,$$
$$H_0: \beta_1 = \beta_2 = \ldots = \beta_m = 0,$$
and, $$H_0: \gamma_1 = \gamma_2 = \ldots = \gamma_m = 0$$

Next, if we write

$\bar{x}_{...}$ = mean of all the m^2 observations

$\bar{x}_{i..}$ = mean of all the m observations in the i th row

$\bar{x}_{.j.}$ = mean of all the m observations in the j th column

$\bar{x}_{..k}$ = mean of all the m observations from the k th treatment (letter)

then we can see very easily that the total sum of square (TSS)

$$\sum_i\sum_j\sum_k (x_{ijk} - \bar{x}_{...})^2 = m\sum_i (\bar{x}_{i..} - \bar{x}_{...})^2 + m\sum_j (\bar{x}_{.j.} - \bar{x}_{...})^2 + m\sum_k (\bar{x}_{..k} - \bar{x}_{...})^2$$

$$+ \sum_i\sum_j\sum_k (\bar{x}_{ijk} - \bar{x}_{i..} - \bar{x}_{.j.} - \bar{x}_{..k} + 2\bar{x}_{...})^2 \quad ...(9.20)$$

the product terms vanish, since the algebraic sum of deviations from respective means are zeros.

Symbolically, it can be expressed as ...(9.21)

$$TSS = SSR + SSC + SST + SSE,$$

where SSR and SSC are the *row* sum of squares and *column* sum of squares respectively, SST is the sum of squares due to *treatments (letters)* and SSE is the sum of squares due to *errors*.

The degrees of freedom are partitioned according to the identity

$$m^2 - 1 = (m - 1) + (m - 1) + (m - 1) + (m - 1)(m - 2)$$

Dividing each of the sum of square on the right hand side of (9.20), with its respective degrees of freedom, we obtain the four independent estimates of the population variance. Thus *the ANOVA table for an $m \times m$ LS D* is

Source of variation	df	SS	MS	F
Row	$m-1$	SSR	MSR = SSR/$(m-1)$	$F_R = \frac{MSR}{MSE} \sim F_{((m-1),(m-1)(m-2))}$
Column	$m-1$	SSC	MSC = SSC/$(m-1)$	$F_C = \frac{MSC}{MSE} \sim F_{((m-1),(m-1)(m-2))}$
Treatment	$m-1$	SST	MST = SST/$(m-1)$	$F_T = \frac{MST}{MSE} \sim F_{((m-1),(m-1)(m-2))}$
Error	$(m-1)(m-2)$	SSE	MSE = SSE/$(m-1)(m-2)$	

In case the tabulated value of F for $[(m-1), (m-1)(m-2)]$ degrees of freedom at level α is greater than the calculated F_T, then there is no significant difference between the treatments (letters) and the hypothesis H_0: $\gamma_1 = \gamma_2 = \ldots = \gamma_m = 0$ may be accepted, otherwise, it is rejected. Similarly, we can test for the hypothesis concerning to the factors along rows and columns.

Remarks

1. The Latin square is designed under the assumption that there is no interaction between different factors, along rows, columns, and treatments. In case it fails then F_T-value calculated in the ANOVA would be inappropriate.
2. To simplify the calculations various sum of squares can be expressed as follows

(a) $$TSS = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{...})^2 = \sum_i \sum_j \sum_k x_{ijk}^2 - \frac{G^2}{N} \quad \ldots(9.20)$$

(b) $$SSR = m \sum_i (\bar{x}_{i..} - \bar{x}_{...})^2 = \frac{1}{m} \sum_i R_i^2 - \frac{G^2}{N} \quad \ldots(9.21)$$

(c) $$SSC = m \sum_j (\bar{x}_{.j.} - \bar{x}_{...})^2 = \frac{1}{m} \sum_j C_j^2 - \frac{G^2}{N} \quad \ldots(9.22)$$

(d) $$SST = m \sum_k (\bar{x}_{..k} - \bar{x}_{...})^2 = \frac{1}{m} \sum_k T_k^2 - \frac{G^2}{N} \quad \ldots(9.23)$$

where $G = \sum_i \sum_j \sum_k x_{ijk}$ is the grand total of $N = m^2$ observations; R_i, $i = 1, 2, \ldots, m$ are the row totals, and C_j, $j = 1, 2, \ldots, m$ are the column totals.

Example 9.5: In the following 3×3 Latin square plan, the letters A, B, and C represent the three methods for soldering copper electrical leads. The rows represent 3 different operators doing the soldering and the columns represent the 3 different solder fluxes used. The data is the number of pounds of tensile force required to separate the soldered leads. Assuming that various sources of variation do not interact, test at 0.05 level of significance whether there are differences in the methods, the operators, or the fluxes.

	Flux 1	Flux 2	Flux 3
Operator 1	A (14.0)	B (16.5)	C (11.0)
Operator 2	C (9.5)	A (17.0)	B (15.0)
Operator 3	B (11.0)	C (12.0)	A (13.5)

Solution: The three treatment totals are:

$$T_A = 44.5, \quad T_B = 42.5 \quad \text{and} \quad T_C = 32.5$$

The three rows totals are:

$$R_1 = 41.5, \quad R_2 = 41.5 \quad \text{and} \quad R_3 = 36.5$$

The three columns totals are:

$$C_1 = 34.5, \quad C_2 = 45.5 \quad \text{and} \quad C_3 = 39.5$$

Grand total, $G = 119.5$, and the total number of observations, $N = 9$.

Also, $$\sum\sum\sum x_{ijk}^2 = (14.0)^2 + (16.5)^2 + (11.0)^2 + (9.5)^2 + (17.0)^2 + (15.0)^2 + (11.0)^2 + (12.0)^2 + (13.5)^2 = 1640.75$$

Thus, $$TSS = 1640.75 - \frac{(119.5)^2}{9} = 54.06$$

$$SSR = \frac{1}{3}[(41.5)^2 + (41.5)^2 + (36.5)^2] - \frac{(119.5)^2}{9}$$
$$= 1592.25 - 1586.69 = 5.56$$

$$SSC = \frac{1}{3}[(34.5)^2 + (45.5)^2 + (39.5)^2] - \frac{(119.5)^2}{9}$$
$$= 1606.92 - 1586.69 = 20.27$$

$$SST = \frac{1}{3}[(44.5)^2 + (42.5)^2 + (32.5)^2] - \frac{(119.5)^2}{9}$$
$$= 1614.25 - 1586.69 = 27.56$$

and, $$SSE = TSS - SSR - SSC - SST$$
$$= 54.06 - 5.56 - 20.27 - 27.56 = 0.67$$

The *ANOVA* table for 3×3 LSD is

Source of Variation	df	SS	MS	F
Rows (Operators)	2	5.56	2.78	$F_R = 8.18$
Columns (Fluxes)	2	20.27	10.14	$F_C = 29.82$
Treatments (Soldering methods)	2	27.56	13.28	$F_T = 39.06$
Error	2	0.67	0.34	

From *Table IVA*, $F_{(2,\,2)[0.5]} = 19$. Thus, $F_R < F_{(2,\,2)[.05]}$, and F_C and $F_T > F_{(2,\,2)[.05]}$, hence at 5% level, variation due to rows (operators) is not significance but variations both due to column (fluxes) and treatments (soldering methods) are significant.

Example 9.6: In the following 4×4 Latin square plan, the letters A, B, C, and D represent 4 varieties of wheat; the rows represent 4 different fertilizers; and the columns represent 4 different years. The data are the yields for the 4 varieties of wheat measured in kilogram per plot. Under the assumption that various sources of variation don't interact, test at $\alpha = 0.05$ the hypothesis that there is no difference in the average yields of the 4 varieties of wheat.

Year→ / Fertilizers ↓	2001	2002	2003	2004
1.	A 70	B 75	C 68	D 81
2.	D 66	A 59	B 55	C 63
3.	C 59	D 66	A 39	B 42
4.	B 41	C 57	D 39	A 55

Solution: Since shifting the origin does not effect the sum of squares, thus shifting it at 60, the Latin square becomes

A 10	B 15	C 8	D 21
D 6	A −1	B −5	C 3
C −1	D 6	A −21	B −18
B −19	C −3	D −21	A −5

The four treatment totals are:

$$T_A = -17, \quad T_B = -27, \quad T_C = 7, \quad \text{and} \quad T_D = 12$$

The four row totals are:

$$R_1 = 54, \quad R_2 = 3, \quad R_3 = -34 \quad \text{and} \quad R_4 = -48$$

The five column totals are:

$$C_1 = -4, \quad C_2 = 17, \quad C_3 = -39 \quad \text{and} \quad C_4 = 1$$

Grand total, $G = -25$, and total number of observations, $N = 16$.

Also, $$\sum_i \sum_j \sum_k x_{ijk}^2 = (10)^2 + (15)^2 + (8)^2 + (21)^2 + (6)^2 + (-1)^2 + (-5)^2 + (3)^2$$
$$+ (-1)^2 + (6)^2 + (-21)^2 + (-18)^2 + (-19)^2 + (-3)^2 + (-21)^2 + (-5)^2$$
$$= 2539$$

Thus, $$\text{TSS} = 2539 - \frac{(-25)^2}{16} = 2539 - 39.06 = 2499.94$$

$$\text{SSR} = \frac{1}{4}[(54)^2 + (3)^2 + (-34)^2 + (-48)^2] - \frac{(-25)^2}{16}$$

$$= 1596.25 - 39.06 = 1557.19$$

$$\text{SSC} = \frac{1}{4}[(-4)^2 + (17)^2 + (-39)^2 + (1)^2] - \frac{(-25)^2}{16}$$
$$= 456.75 - 39.06 = 417.69$$

$$\text{SST} = \frac{1}{4}[(-17)^2 + (-27)^2 + (7)^2 + (12)^2] - \frac{(-25)^2}{16}$$
$$= 302.75 - 36.06 = 263.69.$$

$$\text{SSE} = 2499.94 - 1557.19 - 417.69 - 263.69 = 261.37.$$

The ANOVA table for 4 ×4 LSD is

Source of variation	df	SS	MS	F
Rows (fertilizers)	3	1557.19	519.06	
Columns (years)	3	417.69	139.23	
Treatments (wheat varieties)	3	263.69	87.9	$F_T = 2.02$
Error	6	261.37	43.56	

From *Table IVA* $F_{(3,\,6)[.05]} = 4.76$. Since $F_T < F_{(3,\,6)[.05]}$, thus variation due to treatments is not significant and thus the hypothesis that there is no difference in the average yields of the 4 varieties of wheat may be accepted.

REVIEW EXERCISES

1. What is ANOVA and when is it specifically used? Give the assumptions underlying ANOVA.
2. Work out the ANOVA for a one-way classification, stating the assumptions made.
3. State the mathematical model used in ANOVA for a two-way classification. Mention the hypotheses employed.
4. In case you are testing the equality of several means, using t-test pairwise then what problems are you likely to encounter?
5. What is Latin Square Design (LSD)? Under what conditions can this design be used?
6. Give the statistical analysis of an $m \times m$ LSD.

PROBLEM SET

1. Suppose we wish to compare the means of five populations based on independent random samples each of which contains 8 observations. Insert in an ANOVA table the sources of variation and their respective degrees of freedom. Also mention the assumptions made.
2. A college administrator claims that there is no difference in first-year grade point averages for students entering the college from any of the three different city high schools. The following data give the first-year grade point averages of 12 randomly chosen students, 4 from each of the three high schools. Test the administrator's claim at 5% level of significance.

School 1	School 2	School 3
3.2	3.4	2.8
3.4	3.0	2.6
3.3	3.7	3.0
3.5	3.3	2.7

3. The calcium content of a powdered mineral substance was analyzed five times by each of the three methods with similar deviations:

Method	Per cent calcium				
1	.0279	.0276	.0270	.0275	.0281
2	.0268	.0274	.0267	.0263	.0267
3	.0280	.0279	.0282	.0278	.0283

Use an appropriate test to compare the three methods of measurements at $\alpha = 0.05$.

4. Given the following observations collected according to the one-way analysis variance design

Treatment 1:	6	4	5	5	
Treatment 2:	11	10	13	12	14
Treatment 3:	7	9	11		
Treatment 4:	3	5	1	4	2

Construct the ANOVA table and test the equality of treatments at $\alpha = 0.05$.

5. To compare the prices of nuts in four different states, five suppliers have been randomly selected in each of the four states. The prices per kilogram in rupees are given in the table

States			
A	B	C	D
241	216	230	245
235	220	225	250
238	205	235	238
247	213	228	255
250	220	240	255

Test using $\alpha = 0.05$ whether the data provide sufficient evidence to indicate that the average price per kilogram of nuts differ among the four states.

6. It is suspected that the environmental temperature in which batteries are activated affects their activated life. Thirty homogeneous batteries were tested, six at each of five temperatures and the data shown below were obtained. Carry out the analysis of variance.

Temp.(°C)	Activated life in sec.				
0	55	55	57	54	54	56
25	60	61	60	60	60	60
50	70	72	73	68	77	77
75	72	72	72	70	68	69
100	65	66	60	64	65	65

7. Following data gives the times in seconds to exhaustion of 6 participants when put on a treadmill, who where put on three different diets for a period of six days.

Diet	Participant					
	1	2	3	4	5	6
A	84	35	91	57	56	45
B	91	48	71	45	61	61
C	122	53	110	71	91	122

Perform the ANOVA. Use 0.01 level of significance to determine if there are significant differences among the diets.

8. Following table gives the observations on temperature of a computer chip when four different types of cooling fans were tried on each of the five different computers. Construct the analysis of variance table and test for difference among the cooling fans using $\alpha = 0.05$.

Cooling Fan	Computer				
	A	B	C	D	E
I	26	18	23	12	21
II	26	22	28	21	28
III	24	19	22	21	24
IV	24	21	23	18	19

9. The following data represents the number of different macroinvertebrate species collected at 6 stations, located in the vicinity of a thermal discharge from 1970 to 1977. Test the hypotheses using $\alpha = 0.01$ that the data are unchanging (a) from year to year, and (b) from station to station.

Year	Station					
	1	2	3	4	5	6
1970	53	35	31	37	40	43
1971	36	34	17	21	30	18
1972	47	37	17	31	45	26
1973	55	31	17	23	43	37
1974	40	32	19	26	45	37
1975	52	42	20	27	26	32
1976	39	28	21	21	36	28
1977	40	32	21	21	36	35

10. Set up the ANOVA for the following data of an LSD and state your conclusions at $\alpha = .05$.

A 12	B 19	C 10	D 8.
C 18	B 12	D 6	A 7
B 22	D 10	A 5	C 21
D 12	A 7	C 27	B 17

11. Fill in the blanks in the following ANOVA table of an LSD and state your conclusion at $\alpha = .05$

Source of variation	*df*	*SS*	*MS*	*F*
Rows	–	72	–	2
Columns	–	–	36	–
Treatments	–	180	–	–
Error	6	–	12	
Total	–	–		

12. Set up ANOVA for the following data of an LSD and state your conclusions at $\alpha = .05$.

A 105	B 95	C 125	D 115
C 115	D 125	A 105	B 105
D 115	C 95	B 105	A 115
B 95	A 135	D 95	C 115

ANSWERS

2. Rejected

3. $F = 16.38$, methods are different.

4. $F = 33.55$, treatments are different.

5. $F = 26.44$, prices are different.

6. $F = 70.27$, temperature effects the activated life.

7. $F = 11.86$, significant difference between the diets.

8. $F = 4.24$, not significant.

9. (a) $F = 3.73$, slightly significant (b) $F = 22.48$, significant.

10 CHAPTER Statistical Quality Control

10.1 INTRODUCTION

In the term *statistical quality control* by *quality* we may mean an *attribute of the product that determines its fitness for use*. This may depend on several factors like raw material and machines used, manpower applied, control of the management, etc. Goods of exactly the same quality are not possible to be produced in the continuous flow of any manufacturing process. Indeed every manufacturing process, how good, is characterized by a certain amount of variability, which is of a random nature and which cannot be eliminated completely.

In general, variation in the manufactured product may be attributed to two main causes. One because of *random causes* which are natural or inherent in any manufacturing process. These are influenced by numerous minor factors which can't be controlled. Collectively they create a stable pattern of variation and the range of such variation is known as *natural tolerance of the process*. Second type of variation that sometimes appears, which is far from being inherent to the process, is due to *assignable causes*. The assignable causes may creep in at any stage of the process, right from the arrival of the raw material to the final delivery of goods. Since these causes can be identified and eliminated in a manufacturing process, so the variation due to these is preventable. When the only variation present in the process is due to random causes and not due to assignable causes, we say that the process is in *statistical control*, A key problem in a manufacturing process is to analyze whether a process is in control or is out of control and it is determined by the use of *control charts* proposed by Walter A. Shewhart.

Statistical quality control (S.Q.C.) means planned collection and effective use of data for studying causes of variations in quality and then taking the remedial action whenever assignable causes are present.

The major advantages of bringing a process in good statistical quality control include improvement in the product quality, reduction in cost, better quality assurance at lower inspection cost and preventing frequent and unwarranted adjustments. In addition to these, if a process in control is not good enough then it signals that some radical changes are needed, just making minor changes would not help. Thus statistical quality control reduces wastage of time and material to the absolute minimum by giving an early warning about the occurrence of defects, resulting in less cost of production and ultimately leading to increase in profit.

In Section 10.2 we discuss the control charts and control limits in general. Section 10.3 considers control charts for measurements, and control charts for attributes have been studied in Section 10.4. In the end, a set of review exercises and a problem set has been given.