$$P(x = r) = {}^{n}C_r\, p^r q^{n-r} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!}\, p^r q^{n-r}$$

$$= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)\frac{\lambda^r}{r!}\left(1-\frac{\lambda}{n}\right)^{n-r}\ ;\ np = \lambda \text{ or } p = \lambda/n$$

tends to $\frac{\lambda^r}{r!}e^{-\lambda}$ for a fixed r. Thus the Poisson probability distribution which approximates the binomial distribution is defined by the following probability function:

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}\ ,\ r = 0, 1, 2, \ldots \tag{7-3}$$

where $e = 2.7183$.

Characteristics of Poisson Distribution Since Poisson probability distribution is specified by a process rate λ and the time period t, its mean and variance are identical and are expressed in terms of the parameters: n and p as shown below:

- The *arithmetic mean*, $\mu = E(x)$ of Poisson distribution is given by

$$\mu = \sum x P(x) = \sum x \frac{e^{-\lambda}\lambda^x}{x!},\quad x = 1, 2, 3, \ldots \text{ and } x P(x) = 0 \text{ for } x = 0$$

$$= \lambda e^{-\lambda} + \lambda^2 e^{-\lambda} + \frac{\lambda^3 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^r e^{-\lambda}}{(x-1)!} + \ldots$$

$$= \lambda e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^{r-1}}{(x-1)!} + \ldots\right] = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

  Thus the mean of the distribution is $\mu = \lambda = np$.

- *The variance* σ^2 of Poisson distribution is given by

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

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

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

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Thus the variance of the distribution is $\sigma^2 = \lambda = np$.

The central moments of Poisson distribution can also be determined by the following recursion relation:

$$\mu_r = E(x - \lambda)^r = \sum (x - \lambda)^r e^{-\lambda}\frac{\lambda^x}{x!}$$

Differentially μ_r with respect to λ, we have

$$\frac{d\mu_r}{d\lambda} = -r\mu_{r-1} + \frac{\mu_{r+1}}{\lambda} \text{ or } \mu_{r+1} = \lambda\left[r\mu_{r-1} + \frac{d\mu_r}{d\lambda}\right] \tag{7-4}$$

Substituting $\mu_0 = 1$ and $\mu_1 = 0$ and putting $r = 1$, 2 and 3 in Eqn. (7-4), we have

$$\mu_2 = \mu_3 = \lambda$$

$$\mu_4 = \lambda + 3\lambda^2$$

so that $\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$ and $\gamma_2 = \beta_2 - 3 = \frac{1}{\lambda}$

Hence Poisson distribution is defined by the parameter λ and is positively skewed and leptokurtic. This implies that there is a possibility of infinitely large number of occurrences in a particular time interval, even though the average rate of occurrences is very small. However, as $\lambda \to \infty$, the distribution tends to be symmetrical and mesokurtic.

It is very *rare for more than one event to occur* during a short interval of time. The shorter the duration of interval, the occurrence of two or more events becomes also rare. The probability that exactly one event will occur in such an interval is approximately λ times its duration.

If λ is not an integer and $m = [\lambda]$, the largest integer contained in it, then m is the unique mode of the distribution. But if λ is an integer, the distribution would be bimodal.

The typical application of Poisson distribution is for analysing queuing (or waiting line) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the time interval. While applying Poisson distribution if we consider a time period of different length, the distribution of number of events remains Poisson with the mean proportional to the length of the time period.

Fitting a Poisson Distribution A Poisson distribution can be fitted to the observed values in the data set by simply obtaining values of λ and calculating the probability of zero occurrence. Other probabilities can be calculated by the recurrence relation as follows:

$$f(r) = \frac{e^{-\lambda}\lambda^r}{r!}$$

$$f(r+1) = \frac{e^{-\lambda}\lambda^{r+1}}{(r+1)!}$$

or $$\frac{f(x+1)}{f(r)} = \frac{\lambda}{r+1} \quad \text{or} \quad f(r+1) = \frac{\lambda}{(r+1)} \cdot f(r)\ ;\ r = 0, 1, 2, \ldots$$

Thus, for $r = 0$, $f(1) = \lambda f(0)$,

$$\text{for } r = 1,\ f(2) = \frac{\lambda}{2} f(1) = \frac{\lambda^2}{2} f(0)$$

and so on, where $f(0) = e^{-\lambda}$.

After obtaining the probability for each of the random variable values, multiply each of them by N (total frequency) to get the expected frequency for the respective values.

Example 7.14: What probability model is appropriate to describe a situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this model, what is the probability that a page observed at random will contain at least three misprints?

Solution: Since 100 misprints are distributed randomly throughout the 100 pages of a book, therefore on an average there is only one mistake on a page. This means, the probability of there being a misprint, $p = 1/100$, is very small and the number of words, n, in 100 pages are vary large. Hence, Poisson distribution is best suited in this case.

Average number of misprints in one page, $\lambda = np = 100 \times (1/100) = 1$. Therefore $e^{-\lambda} = e^{-1} = 0.3679$.

Probability of at least three misprints in a page is

$$P(x \geq 3) = 1 - P(x < 3) = 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\}$$

$$= 1 - [e^{-\lambda} + \lambda e^{-\lambda} + \frac{1}{2!}\lambda^2 e^{-\lambda}]$$

$$= 1 - \left\{e^{-1} + e^{-1} + \frac{e^{-1}}{2!}\right\} = 1 - 2.5\, e^{-1} = 1 - 2.5\ (0.3679)$$

$$= 0.0802$$

Example 7.15: A new automated production process has had an average of 1.5 breakdowns per day. Because of the cost associated with a breakdown, management is concerned about the possibility of having three or more breakdowns during a day. Assume that breakdowns occur randomly, that the probability of a breakdown is the same for any two time intervals of equal length, and that breakdowns in one period are inclependent of breakdowns in other periods. What is the probability of having three or more breakdowns during a day? *[HP Univ., MBA, 1995; Kumaon Univ., 1998]*

Solution: Given that, $\lambda = np = 1.5$ breakdowns per day. Thus probability of having three or more breakdowns during a day is given by

$$P(x \geq 3) = 1 - P(x < 3) = 1 - [P(x=0) + P(x=1) + P(x=2)]$$

$$= 1 - \left[\frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!}\right]$$

$$= 1 - e^{-\lambda}\left[1 + \lambda + \frac{1}{2}\lambda^2\right] = 1 - 0.2231\left[1 + 1.5 + \frac{1}{2}(1.5)^2\right]$$

$$= 1 - 0.2231\ (3.625) = 1 - 0.8088 = 0.1912$$

Example 7.16: Suppose a life insurance company insures the lives of 5000 persons aged 42. If studies show the probability that any 42-years old person will die in a given year to be 0.001, find the probability that the company will have to pay at least two claims during a given year.

Solution: Given that, $n = 5000$, $p = 0.001$, so $\lambda = np = 5000 \times 0.001 = 5$. Thus the probability that the company will have to pay at least 2 claims during a given year is given by

$$P(x \geq 2) = 1 - P(x < 2) = 1 - [P(x = 0) + P(x = 1)]$$
$$= 1 - [e^{-\lambda} + \lambda e^{-\lambda}] = 1 - [e^{-5} + 5e^{-5}] = 1 - 6e^{-5}$$
$$= 1 - 6 \times 0.0067 = 0.9598$$

Example 7.17: A manufacturer who produces medicine bottles, finds that 0.1 per cent of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using Poisson distribution, find how many boxes will contain:

(i) no defectives

(ii) at least two defectives *[Delhi Univ., MBA, 1996, 2001]*

Solution: Given that, $p = 1$ per cent $= 0.001$, $n = 500$, $\lambda = np = 500 \times 0.001 = 0.5$

(i) $P[x = 0] = e^{-\lambda} = e^{-0.5} = 0.6065$

Therefore, the required number of boxes are : $0.6065 \times 100 = 61$ (approx.)

(ii) $$P(x > 2) = 1 - P(x \leq 1) = 1 - [P(x = 0) + P(x = 1)]$$
$$= 1 - [e^{\lambda} + \lambda e^{-\lambda}] = 1 - [0.6065 + 0.5(0.6065)]$$
$$= 1 - 0.6065\ (1.5) = 1 - 0.90975 = 0.09025.$$

Therefore, the required number of boxes are $100 \times 0.09025 = 10$ (approx.)

Example 7.18: The following table gives the number of days in a 50-day period during which automobile accidents occurred in a city :

No. of accidents :	0	1	2	3	4
No. of days :	21	18	7	3	1

Fit a Poisson distribution to the data.

[Sukhadia Univ.,MBA,1992; Kumaon Univ., MBA,2000]

Solution: Calculations for fitting of Poisson distribution are shown in the Table 7.3.

Table 7.5: Calculations for Poisson Distribution

Number of Accidents (x)	Number of Days (f)	fx
0	21	0
1	18	18
2	7	14
3	3	09
4	1	04
	$n = 50$	$\Sigma fx = 45$

Thus $$\bar{x}(\text{or } \lambda) = \frac{\Sigma f x}{n} = \frac{45}{50} = 0.9$$

and
$$P(x=0) = e^{-\lambda} = e^{-0.9} = 0.4066$$
$$P(x=1) = \lambda P(x=0) = 0.9(0.4066) = 0.3659$$
$$P(x=2) = \frac{\lambda}{2} P(x=1) = \frac{0.9}{2}(0.3659) = 0.1647$$
$$P(x=3) = \frac{\lambda}{3} P(x=2) = \frac{0.9}{3}(0.1647) = 0.0494$$
$$P(x=4) = \frac{\lambda}{4} P(x=3) = \frac{0.9}{4}(0.0494) = 0.0111$$

In order to fit a Poisson distribution, we shall multiply each of these values by N = 50 (total frequencies). Hence the expected frequencies are:

x :	0	1	2	3	4
f :	0.4066×50	0.3659×50	0.1647×50	0.0494×50	0.0111×50
	$= 20.33$	$= 18.30$	$= 8.23$	$= 2.47$	$= 0.56$

7.5.3 Negative Binomial Probability Distribution

All conditions of binomial distribution are also applicable to the negative binomial distribution except that it describes the number of trials likely to be required to obtain a fixed number of successes. For example, suppose a percentage p of individuals in the population are sampled until exactly r individuals with the certain characteristic are found. The number of individuals in excess of r that are observed or sampled has a negative binomial distribution.

The probability distribution function of the negative binomial distribution is obtained by considering an infinite series of Bernoulli trials with probability of success p of an event on an individual trial. If trials are repeated r times until an event of interest occurs, then the probability that at least m trials will be required to get the event r times (successes) is given by

$P(m, r, p)$ = Probability that an event occurs $(r-1)$ times in the first $m-1$ trials × Probability that the event of interest occurs in the mth trial

$$= {}^{m-1}C_{r-1}\, p^{r-1} q^{m-r} \times p = {}^{m-1}C_{r-1}\, p^r q^{m-r}, \quad m = r, r+1, \ldots \qquad (7\text{-}5)$$

where $r \geq 1$ is a fixed integer.

A random variable having a negative binomial distribution is also referred to as a discrete waiting time random variable. In terms of number of failures, it represents how long one waits for the rth success.

The mean and variance of this distribution are given by

$$\text{Mean, } \mu = \frac{r}{p}, \quad \text{Variance, } \sigma^2 = \frac{rq}{p^2}$$

Example 7.19: A market research agency that conducts interviews by telephone has found from past experience that there is a 0.40 probability that a call made between 2.30 PM and 5.30 PM will be answered. Assuming a Bernoullian process,

(a) calculate the probability that an interviewer's 10th answer comes on his 20th call and that he will receive the first answer on his 3rd call,

(b) what is the expected number of calls necessary to obtain seven answers.

Solution: Let answer to a call be considered 'success'. Then $p = 0.40$

(a) P(10th answer comes on 20th call)

$$= {}^{m-1}C_{r-1}\, p^r q^{m-r}, \quad m = r, r+1, \ldots$$
$$= {}^{19}C_9\, (0.4)^{10} (0.6)^{10}; \quad m = 20 \text{ and } r = 10$$
$$= 0.058$$

P(First answer on 3rd call) $= {}^{2}C_0\, (0.4)^1 (0.6)^2 = 0.144$

(b) Expected number of calls for 7 answers, $\mu = r/p = 7/0.4 = 17.5 \cong 18$ calls

7.5.4 Multinomial Probability Distribution

The binomial distribution discussed earlier is associated with a sequence of n independent repeated Bernoulli trials, each resulting in only two outcomes, and one of the possible outcomes is called success, while multinomial distribution is associated with independent repeated trials that generalize from Bernoulli trials each resulting in two to k outcomes.

Suppose a single trial of an experiment results in only one of the k possible outcomes $O_1, O_2, \ldots, O_k$ with respective probabilities $p_1, p_2, \ldots, p_k$, and the experiment is repeated n times independently. The probability that out of these n trials outcome O_1 occurs x_1 times, O_2 occurs x_2 times and so on, is given by the following discrete density function:

$$P(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\, x_2! \ldots x_k!}\left[p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}\right] \tag{7-6}$$

where $x_1 + x_2 + \ldots + x_k = n$.

Example 7.20: In a factory producing certain items, 30 per cent of the items produced have no defect, 40 per cent have one defect, and 30 per cent have two defects. A random sample of 8 items is taken from a day's output. Find the probability that it will contain 2 items with no defect, 3 items with one defect, and 3 items with two defects.

Solution: We know from the data that $n = 8, p_1 = 0.30, p_2 = 0.40$ and $p_3 = 0.30; x_1 = 2$, $x_2 = 3, x_3 = 3$. Thus the required probability is given by

$$P(x_1=2, x_2=3, x_3=3) = \frac{8!}{2!\,3!\,3!}\left[(0.30)^2 (0.40)^2 (0.30)^3\right]$$
$$= 0.0871$$

7.5.5 Hypergeometric Probability Distribution

For a binomial distribution to be applied, the probability of a success or failure must remain the same for each trial. This is possible only when the number of elements in the population are large relative to the number in the sample, where probability of getting a success on a single trial is equal to the proportion p of successes in the population. However, if the number of element in the population are small in relation to the sample size, i.e. $n/N \geq 0.5$, the probability of a success in a given trial is dependent upon the outcomes of preceding trials. Then the number r of successes follows hypergeometric probability distribution. Thus hypergeometric probability distribution is similar to binomial distribution where probability of success may be different from trial to trial. When sampling is done *without replacement* from a finite population, the Bernoulli process does not apply because there is a systematic change in the probability of success in the reduced size of population.

Let N be the size of population and out of N, m be the total number of elements having a certain characteristic (called success) and the remaining $N - m$ do not have it, such that $p + q = 1$. Suppose a sample of size n is drawn at random without replacement. Then in a random sample of size n, the probability of exactly r sucesses when values of r depend on N, p and n is given by

$$P(x = r) = \frac{{}^{m}C_r \, {}^{N-m}C_{n-r}}{{}^{N}C_n}; \quad r = 0, 1, 2, \ldots, n; \quad \text{and } 0 \leq r \leq m$$

This probability mass function is called *hypergeometric probability distribution.*

The mean and variance of a hypergoemetric distribution are

$$\text{Mean} = n\left(\frac{m}{N}\right) \text{ and Variance} = n\left(\frac{m}{N}\right)\left(\frac{N-m}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Example 7.21: Suppose the HRD manager randomly selects 3 individuals from a group of 10 employees for a special assignment. Assuming that 4 of the employees were assigned to a similar assignment previously, determine the probability that exactly two of the three employees have had previous experience.

Solution: We know from the data that $N = 10, n = 3, r = 2, m = 4$, and $N - m = 6$. Thus the required probability is given by

$$P(x = r \mid N, m, N\text{–}m) = \frac{{}^{m}C_{r}\,{}^{N-m}C_{n-r}}{{}^{N}C_{n}} = \frac{{}^{4}C_{2}\,{}^{10-4}C_{3-2}}{{}^{10}C_{3}}$$

$$= \frac{{}^{4}C_{2}\,{}^{6}C_{1}}{{}^{10}C_{3}} = \frac{\left(\frac{4!}{2!2!}\right)\left(\frac{6!}{1!5!}\right)}{\left(\frac{10!}{3!7!}\right)} = \frac{36}{120} = 0.30$$

Example 7.22: Suppose a particular industrial product is shipped in lots of 20. To determine whether an item is defective a sample of 5 items from each lot is drawn. A lot is rejected if more than one defective item is observed. (If the lot is rejected, each item in the lot is then tested). If a lot contains four defectives, what is the probability that it will be accepted?

Solution: Let r be the number of defectives in the sample size $n = 5$. Given that, $N = 20$, $m = 4$, and $N - m = 16$. Then

$$P(\text{accept the lot}) = P(x \le 1) = P(x = 0) + P(x = 1)$$

$$= \frac{{}^{4}C_{0} \times {}^{16}C_{5}}{{}^{20}C_{5}} + \frac{{}^{4}C_{1} \times {}^{16}C_{4}}{{}^{20}C_{5}} = \frac{\frac{4!}{0!4!} \times \frac{16!}{5!11!}}{\frac{20!}{5!15!}} + \frac{\frac{4!}{1!3!} \times \frac{16!}{4!12!}}{\frac{20!}{5!15!}}$$

$$= \frac{91}{323} + \frac{455}{969} = 0.2817 + 0.4696 = 0.7513$$

Conceptual Questions 7C

13. If x has a Poisson distribution with parameter λ, then show that $E(x)$ and $V(x) = \lambda$. Further, show that the Poisson distribution is a limiting form of the binomial distribution.

14. What is Poisson distribution? Point out its role in business decision-making. Under what conditions will it tend to become a binomial distribution?

[Kumaon Univ., MBA, 1998]

15. When can Poisson distribution be a reasonable approximation of the binomial?

[Delhi Univ., MCom, 1999]

16. Discuss the distinctive features of Poisson distribution. When does a binomial distribution tend to become a Poisson distribution?

17. Under what conditions is the Poisson probability distribution appropriate? How are its mean and variance calculated?

18. What is negative binomial distribution? Distinguish the relationship between the binomial and negative binomial distributions.

19. What is hypergeometric distribution? Explain its properties.

Self-Practice Problems 7C

7.24 The following table shows the number of customers returning the products in a marketing territory. The data is for 100 stores:

No. of returns :	0	1	2	3	4	5	6
No. of stores :	4	14	23	23	18	9	9

Fit a Poisson distribution. *[Lucknow Univ., MBA, 1997]*

7.25 In a certain factory manufacturing razor blades, there is a small chance of 1/150 for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using the Poisson distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets.

7.26 In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the Poisson distribution, find the probability that there will be three or more accidents in a day.

[Coimbatore Univ., MBA, 1997]

7.27 The distribution of typing mistakes committed by a typist is given below. Assuming a Poisson distribution, find out the expected frequencies:

No. of mistakes per page :	0	1	2	3	4	5
No. of pages :	142	156	69	27	5	1

[Rohilkhand Univ., MBA, 1998]

7.28 Find the probability that at most 5 defective bolts will be found in a box of 200 bolts if it is known that 2 per cent of such bolts are expected to be defective [you may take the distribution to be Poisson; $e^{-4} = 0.0183$].

7.29 On an average, one in 400 items is defective. If the items are packed in boxes of 100, what is the probability that any given box of items will contain: (i) no defectives; (ii) less than two defectives; (iii) one or more defectives; and (iv) more than three defectives *[Delhi Univ., MBA, 2000]*

7.30 It is given that 30 per cent of electric bulbs manufactured by a company are defective. Find the probability that a sample of 100 bulbs will contain (i) no defective, and (ii) exactly one defective.

7.31 One-fifth per cent of the blades produced by a blade manufacturing factory turn out to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, and two defective blades respectively in a consignment of 1,00,000 packets. *[Delhi Univ., MBA, 1999, 2003]*

7.32 A factory produces blades in packets of 10. The probability of a blade to be defective is 0.2 per cent. Find the number of packets having two defective blades in a consignment of 10,000 packets.

7.33 When a first proof of 200 pages of an encyclopaedia of 5,000 pages was read, the distribution of printing mistakes was found to be as shown in the first and second columns of the table below. Fit a Poisson distribution to the frequency distribution of printing mistakes. Estimate the total cost of correcting the whole encyclopaedia by using the information given in the first and third columns of the table below:

Misprints on a Page	*Frequency*	*Cost of Detection and Correction Per Page (Rs)*
0	113	1.00
1	62	2.50
2	20	1.50
3	3	3.00
4	1	3.50
5	1	4.00

[MD Univ., MBA, 1998]

7.34 In a certain factory manufacturing razor blades, there is small chance 1/50 for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using an appropriate probability distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets.

7.35 Suppose that a manufactured product has 2 defects per unit of product inspected. Using Poisson distribution, calculate the probabilities of finding a product without any defect, 3 defects, and 4 defects. (Given $e^{-2} = 0.135$) *[Madurai Univ., MCom, 1999]*

7.36 A distributor received a shipment of 12 TV sets. Shortly after this shipment was received, the manufacturer informed that he had inadvertently shipped 3 defective sets. The distributor decided to test 4 sets randomly selected out of 12 sets received.

(a) What is the probability that neither of the 4 sets tested was defective?

(b) What is the mean and variance of defective sets.

7.37 Suppose a population contains 10 elements, 6 of which are defective. A sample of 3 elements is selected. What is the probability that exactly 2 are defective?

7.38 A transport company has a fleet of 15 trucks, used mainly to deliver fruits to wholesale market. Suppose 6 of the 15 trucks have brake problems. Five trucks were selected at random to be tested. What is the probability that 2 of those tested trucks have defective brakes?

7.39 A company has five applicants for two positions: two women and three men. Suppose that the five applicants are equally qualified and that no preference is given for choosing either gender. If r equal the number of women chosen to fill the two positions, then what is the probability distribution of r. Also, determine the mean and variance of this distribution.

Hints and Answers

7.24 Fitting of Poisson distribution

x :	0	1	2	3	4	5	6	
f :	4	14	23	23	18	9	9	= 100
fx :	0	14	46	69	72	45	54	= 300

$\therefore$ $\lambda = 300/100 = 3$. Then

$NP(x = 0) = 100\, e^{-\lambda} = 100(2.7183)^{-3} = 5$;

$P(x = 1) = \lambda\, NP(0) = 15$

$P(x = 2) = NP(x = 1)\dfrac{\lambda}{2} = 22.5$;

$P(x = 3) = NP(x = 2)\dfrac{\lambda}{3} = 22.5$

$P(x = 4) = NP(x = 3)\dfrac{\lambda}{4} = 16.9$;

$P(x = 5) = NP(x = 4)\dfrac{\lambda}{5} = 10.1$

$P(x = 6) = NP(x = 5)\dfrac{\lambda}{6} = 5.1$

7.25 Given that N = 10,000, $p = 1/50$, $n = 10$, $\lambda = np = 10 \times (1/50) = 0.2$

$P(x = 0) = e^{-\lambda} = e^{-0.2} = 0.8187$ (from the table)

$NP(x = 0) = 0.8187 \times 10{,}000 = 8187$

$NP(x = 1) = NP(x = 0) \times \lambda = 8187 \times 0.2 = 1637.4$

$NP(x = 2) = NP(x = 1) \times \lambda/2 = 1637.4 \times (0.2/2)$
$= 163.74$

The approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets is: 10,000 – (8187+163.74+163.74) = (10,000 – 9988.14) = 11.86 or 12.

7.26 The average number of accidents per day = 10/50 = 0.2

$P(x \geq 3$ accidents$) = 1 - P(2$ or less accidents$)$
$= 1 - [P(0) + P(1) + P(2)]$

$$= -\left[e^{-2} + 2e^{-2} + \frac{e^{-2} \times 0.2 \times 0.2}{2}\right]$$

$= 1 - e^{-2}[1 + 0.2 + 0.02] = 1 - 0.8187 \times 1.22$ (From table of $e^{-\lambda}$)

$= 1 - 0.998 = 0.002$

7.27 $\lambda = \Sigma fx / N = 400/400 = 1$,

$P(x = 0) = e^{-\lambda} = 0.3679$

$NP(x = 0) = 147.16$

$NP(x = 1) = NP(x = 0)\lambda = 147.16$

$$NP(x = 2) = NP(x = 1)\frac{\lambda}{2} = 73.58$$

$$NP(x = 3) = NP(x = 2)\frac{\lambda}{3} = 24.53$$

$$NP(x = 4) = NP(x = 3)\frac{\lambda}{4} = 6.13$$

$$NP(x = 5) = NP(x = 4)\frac{\lambda}{5} = 1.23$$

Expected frequencies as per the distribution are:

No. of mistakes per page	:	0	1	2	3	4	5
No. of pages	:	147	147	74	25	6	1

7.28 p(defective bolt) = 2% = 0.02. Given $n = 200$, so $\lambda = np = 200 \times 0.02 = 4$

$$P(0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-4} = 0.0183$$

$$P(x \le 5) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5)$$

$$= e^{-4}\left(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} + \frac{5^5}{5!}\right)$$

$= 0.0183 \times (643/15) = 0.7844.$

7.30 $\lambda = np = 100 \times 0.30 = 3$

$P(x = 0) = e^{-\lambda} = e^{-3} = 0.05$;

$P(x = 1) = \lambda P(x = 0) = 3 \times 0.03 = 0.15$

7.31 Given $n = 10, p = 1/500, \lambda = np = 10/500 = 0.02$

(i) $P(x = 0) = e^{-\lambda} = e^{-0.02} = 0.9802$

$NP(x = 0) = 1,00,000 \times 0.9802$

$= 98020$ packets

(ii) $P(x = 1) = \lambda P(x = 0)$

$= \lambda e^{-\lambda} = 0.02 \times 0.9802$

$= 0.019604$

$NP(x = 1) = 1,00,000 \times 0.019604$

$= 1960$ packets

(iii) $$P(x = 2) = \frac{\lambda^2}{2} P(x = 0) = \frac{(0.02)^2}{2} \times 0.9802$$

$= 0.00019604$

$NP(x = 2) = 1,00,000 \times 0.00019604$

$= 19.60 \cong 20$ packets

7.32 Given $n = 10, p = 0.002, \lambda = np = 10 \times 0.002 = 0.02$.

$$P(x = 2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-0.02}(0.02)^2}{2!}$$

$= 0.000196$

The required number of packets having two defective blades each in a consignment of 10,000 packets $= 10,000 \times 0.000196 \cong 2$.

7.33 No. of mis-

Prints (x)	:	0	1	2	3	4	5	
Frequency (f)	:	113	62	20	3	1	1	= 200 (=N)
fx	:	0	62	40	09	04	05	= 120

$\therefore \quad \bar{x} = \Sigma fx/N = 120/200 = 0.6 (= \lambda)$

For fitting of Poisson distribution, calculating

$$NP(x = 0) = 200\left[\frac{e^{-\lambda}\lambda^0}{0!}\right] = 200\, e^{-0.6}$$

$= 200\,(0.5488) = 109.76$

$NP(x = 1) = 200\, P_0 \times \lambda = 200 \times 109.76 \times 0.6$

$= 65.856$

$$NP(x = 2) = 200P_1 \times \frac{\lambda}{2} = 65.856 \times \frac{0.6}{2}$$

$= 19.756$

$$NP(x = 3) = 200\, P_2 \times \frac{\lambda}{3} = 19.756 \times \frac{0.6}{3}$$

$= 3.951$

$$NP(x = 4) = 200\, P_3 \times \frac{\lambda}{4} = 3.951 \times \frac{0.6}{4}$$

$= 0.5927$

$$NP(x = 5) = 200\, P_4 \times \frac{\lambda}{5} = 0.5927 \times \frac{0.6}{5}$$

$= 0.0711$

The total cost of correcting the first proof of the whole encyclopaedia will be

No. of Misprints/Page	*Rate/Page* (x)	*No. of Pages* (f)	*Total Cost of Correcting Proof* (fx)
0	1.00	109.760	109.760
1	1.50	65.856	98.784
2	2.50	19.756	49.390
3	3.00	3.951	11.853
4	3.50	0.592	2.074
5	4.00	0.071	0.284
			Rs 272.145

7.34 Given N = 10,000, $p = 1/50$ and $n = 10$.

Thus $\lambda = np = 0.20$ and

$NP\,(x = 0) = 10,000\, e^{-\lambda} = 10,000\, e^{-0.20} = 8187$;

$P(x = 1) = NP(x = 0) \times \lambda = 8187 \times 0.2 = 1637.4.$

$$NP(x = 2) = NP(x = 1) \times \frac{\lambda}{2} = 1637.4 \times \frac{0.2}{2}$$

$= 163.74$

The approximate number of packets containing not more than 2 defective blades in a consignment of 10,000 packets will be: 10,000 – (8187 + 1637.40 + 163.74) = 12 approx.

7.35 Given average number of defects, $\lambda = 2$.

$P(x = 0) = e^{-\lambda} = e^{-2} = 0.135$;

$P(x = 1) = P(x = 0) \times \lambda = 0.135 \times 2 = 0.27$

$$P(x = 2) = P(x = 1) \times \frac{\lambda}{2} = 0.27 \times \frac{2}{2} = 0.27$$

$$P(x = 3) = P(x = 2) \times \frac{\lambda}{3} = 0.27 \times \frac{2}{3} = 0.18$$

$$P(x = 4) = P(x = 3) \times \frac{\lambda}{4} = 0.18 \times \frac{2}{4} = 0.09$$

7.36 Given $N = 12, n = 4, m = 3$ and $N - m = 9$.

$$\text{(a) } P(x = 0) = \frac{{}^3C_0 \times {}^9C_4}{{}^{12}C_4} = \frac{\frac{3!}{0!3!} \times \frac{9!}{4!5!}}{\frac{12!}{4!8!}}$$

$$= \frac{1 \times 126}{495} = \frac{14}{55}$$

(b) Mean, $\mu = n\left(\frac{m}{N}\right) = 4\left(\frac{3}{12}\right) = 1$

$$\text{Variance, } \sigma^2 = n\left(\frac{m}{N}\right)\left(\frac{N-m}{N}\right)\left(\frac{N-n}{N-1}\right)$$

$$= 4\left(\frac{3}{12}\right)\left(\frac{9}{12}\right)\left(\frac{8}{11}\right) = 0.5455$$

7.37 $P(x = 2) = \frac{{}^6C_2 \times {}^4C_1}{{}^{10}C_3} = \frac{15 \times 4}{120} = 0.50$

7.38 $P(x = 2) = \frac{{}^9C_3 \times {}^6C_2}{{}^{15}C_5} = \frac{84 \times 15}{3003} = 0.4196$

7.39 Given $N = 5, n = 2, m = 2, N - m = 3$

$$P(x = r) = \frac{{}^mC_r \times {}^{N-m}C_{n-r}}{{}^NC_n} = \frac{{}^2C_r \times {}^3C_{2-r}}{{}^5C_2}; r = 0, 1, 2$$

Mean, $\mu = 2\left(\frac{2}{5}\right) = 0.8$;

Variance $= 2\left(\frac{2}{5}\right)\left(\frac{3}{5}\right)\left(\frac{3}{4}\right) = 0.6$

7.6 CONTINUOUS PROBABILITY DISTRIBUTIONS

If a random variable is discrete, then it is possible to assign a specific probability to each of its value and get the probability distribution for it. The sum of all the probabilities associated with the different values of the random variable is 1. However, not all experiments result in random variables that are discrete. Continuous random variables such as height, time, weight, monetary values, length of life of a particular product, etc. can take large number of obserable values corresponding to points on a line interval much like the infinite number of gains of sand on a beach. The sum of probability to each of these infinitely large values is no longer sum to 1.

Unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of their specified values. Instead, probability distribution is created by distributing one unit of probability along the real line, much like distributing a handful of sand along a line. The probability of measurements (e.g. gains of sand) piles up in certain places resulting into a probability distribution called *probability density function*. Such distribution is used to find probabilities that the random variable falls into a specified interval of values. The depth or density of the probability that varies with the random variable (x) may be described by a mathematical formula.

The probability density function for a continuous random variable x is a curve such that the area under the curve over an interval equals the probability that x falls into that interval, i.e. the probability that x is in that interval can be found by summing the probabilities in that interval. Certain characteristics of probability density function for the continuous random variable, x are follows:

(i) The area under a continuous probability distribution is equal to 1.

(ii) The probability $P(a \leq x \geq b)$ that random variable x value will fall into a particular interval from a to b is equal to the area under the density curve between the points (values) a and b.

Nature seems to follows a predicatable pattern for many kinds of measurments. Most numerical values of a random variable are spread arround the center, and greater the distance a numerical value has from the center, the fewer numerical values have that

Normal distribution: A continuous probability distribution in which the curve is bell-shaped having a single peak. The mean of the distribution lies at the center of the curve and the curve is symmetrical around a vertical line erected at the mean. The tails of the curve extend indefinitely parallel to the horizontal axis.

specific value. A frequency distribution of values of random variable observed in nature which follows this pattern is approximately bell shaped. A special case of distribution of measurements is called a **normal curve (or distribution).**

If a population of numerical values follows a normal curve and x is the randomly selected numerical value from the population, then x is said to be *normal random variable*, which has a normal probability distribution.

W.J. Youden, a well-known statistician expressed his views about normal distribution as follows:

THE
NORMAL
LAW OF ERROR
STANDS OUT IN
THE EXPERIENCE OF
MANKIND AS ONE OF THE
BROADEST GENERALISATION OF
NATURAL PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT RESEARCHES IN THE
PHYSICAL AND SOCIAL SCIENCES AND IN MEDICINE
AGRICULTURE AND ENGINEERING. IT IS AN INDISPENSABLE
TOOL FOR THE ANALYSIS AND THE INTERPRETATION OF
THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Artistically, it also look like a normal curve

The normal distribution also known as *Gaussian distribution* is due to the work of German mathematician Karl Friedrich Gauss during the early part of the 19th century. Normal distribution provides an adequate representation of a continuous phenomenon or process such as daily changes in the stock market index, frequency of arrivals of customers at a bank, frequency of telephone calls into a switch board, customer servicing times, and so on.

7.6.1 Normal Probability Distribution Function

The formula that generates normal probability distribution is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-1/2)[(x-\mu)/\sigma]^2}, \quad -\infty < x < \infty \tag{7-6}$$

where π = constant 3.1416
e = constant 2.7183
μ = mean of the normal distribution
σ = standard of normal distribution

The $f(x)$ values represent the relative frequencies (height of the curve) within which values of random variable x occur. The graph of a normal probability distribution with mean μ and standard deviation σ is shown in Fig. 7.7. The distribution is symmetric about its mean μ that locates at the centre.

Since the total area under the normal probability distribution is equal to 1, the symmetry implies that the area on either side of μ is 50 per cent or 0.5. The *shape* of the distribution is determined by μ and σ values.

In symbols, if a random variable x follows normal probability distribution with mean μ and standard deviation σ, then it is also expressed as: $x \sim N(\mu, \sigma)$.

Characteristics of the Normal Probability Distribution There is a family of normal distributions. Each normal distribution may have a different mean μ or standard deviation σ. A unique normal distribution may be defined by assigning specific values to the mean

μ and standard deviation σ in the normal probability density function (7-6). Large value of σ reduce the height of the curve and increase the spread; small values of σ increase the height of the curve and reduce the spread. Figure 7.6(a) shows three normal distributions with different values of the mean μ and a fixed standard deviation σ, while in Fig. 7.6(b) normal distributions are shown with different values of the standard deviation σ and a fixed mean μ.

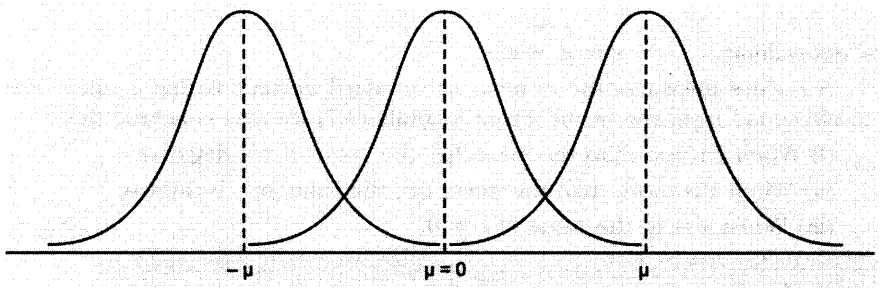


Figure 7.6 (a)
Normal Distributions with Different Mean Values But Fixed Standard Deviation

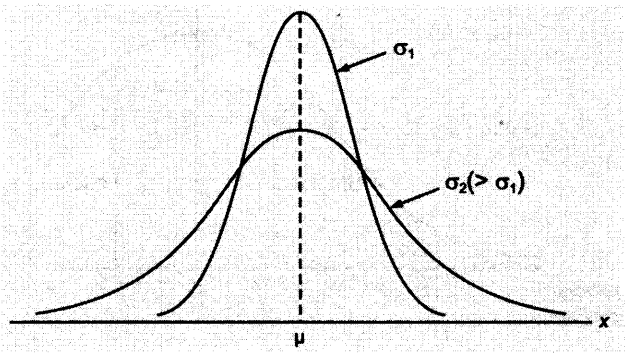


Figure 7.6 (b)
Normal Distributions with Fixed Mean and Variable Standard Deviation

From Fig. 7.6(a) and 7.6(b) the following characteristics of a normal distribution and its density function may be derived:

(i) For every pair of values of μ and σ, the curve of normal probability density function is bell shaped and symmetric.

(ii) The normal curve is symmetrical around a vertical line erected at the mean μ with respect to the area under it, that is, fifty per cent of the area of the curve lies on both sides of the mean and reflect the mirror image of the shape of the curve on both sides of the mean μ. This implies that the probability of any individual outcome above or below the mean will be same. Thus, for any normal random variable x,

$$P(x \leq \mu) = P(x \geq \mu) = 0.50$$

(iii) Since the normal curve is symmetric, the mean, median, and mode for the normal distribution are equal because the highest value of the probability density function occurs when value of a random variable, $x = \mu$.

(iv) The two tails of the normal curve extend to infinity in both directions and theoretically never touch the horizontal axis.

(v) The mean of the normal distribution may be negative, zero, or positive as shown in Fig. 7.6(a).

(vi) The mean μ determines the *central location* of the normal distribution, while standard deviation σ determines its *spread*. The larger the value of the standard deviation σ, the wider and flatter is the normal curve, thus showing more variability in the data, as shown in Fig. 7.6(b). Thus standard deviation σ determines the range of values that any random variable is likely to assume.

(vii) The area under the normal curve represents probabilities for the normal random variable, and therefore, the total area under the curve for the normal probability distribution is 1.

Standard Normal Probability Distribution: To deal with problems where the normal probability distribution is applicable more simply, it is necessary that a random variable x is standardized by expressing its value as the number of standard deviations (σ) it lies to the left or right of its mean (μ). The *standardized normal random variable*, z (also called *z–statistic*, *z–score* or *normal variate*) is defined as:

$$z = \frac{x - \mu}{\sigma} \qquad (7\text{-}7)$$

or equivalently $x = \mu + z\sigma$

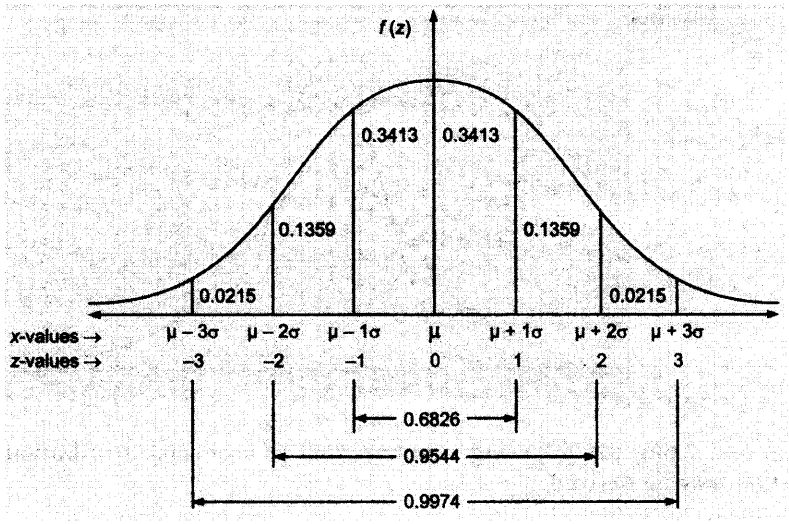
A z–score measures the number of standard deviations that a value of the random variable x fall from the mean. From formula (7.7) we may conclude that

(i) When x is less than the mean (μ), the value of z is negative

(ii) When x is more than the mean (μ), the value of z is positive

(iii) When $x = \mu$, the value of $z = 0$.

Figure 7.7
Standard Normal Distribution



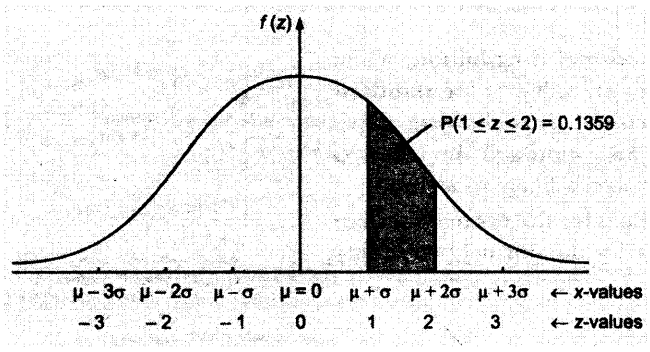
Any normal probability distribution with a set of μ and σ value with random variable can be converted into a distribution called **standard normal probability distribution** z, as shown in Fig. 7.7, with mean $\mu_z = 0$ and standard deviation $\sigma_z = 1$ with the help of the formula (7–7).

Standard normal probability distribution: A normal probability distribution with mean mean equal to zero and standard deviation equal to one.

A z-value measures the distance between a particular value of random variable x and the mean (μ) in units of the standard deviation (σ). With the value of z obtained by using the formula (7.7), we can find the area or probability of a random varible under the normal curve by referring to the standard distribution in Appendix. For example $z = \pm 2$ implies that the value of x is 2 standard deviations above or below the mean (μ).

Area Under the Normal Curve Since the range of normal distribution is infinite in both the directions away from μ, the *pdf* function $f(x)$ is never equal to zero. As x moves away from μ, $f(x)$ approaches x-axis but never actually touches it.

Figure 7.8
Diagram for Finding P(1 < z < 2)



The area under the standard normal distribution between the mean $z = 0$ and a specified positive value of z, say z_0 is the probability $P(0 \le z \le z_0)$ and can be read off directly from standard normal (z) tables. For example, area between $1 \le z \le 2$ is the proportion of the area under the curve which lies between the vertical lines erected at two points along the x–axis. A portion of the table is shown in Table 7.6. For example, as shown in Fig. 7.8, if x is σ away from μ, that is, the distance between x and μ is one standard deviation or $(x - \mu)/\sigma = 1$, then 34.134 per cent of the distribution lies between x and μ. Similarly, if x is at

2σ away from μ, that is, $(x - \mu)/\sigma = 2$, then the area will include 47.725 per cent of the distribution, and so on, as shown in Table 7.6.

Table 7.6: Area Under the Normal Curve

$z = \frac{x-\mu}{\sigma}$	*Area Under Normal Curve Between x and μ*
1.0	0.34134
2.0	0.47725
3.0	0.49875
4.0	0.49997

Since the normal distribution is symmetrical, Table 7.6 indicates that about 68.26 per cent of the normal distribution lies within the range $\mu - \sigma$ to $\mu + \sigma$. The other relationships derived from Table 7.6 are shown in Table 7.7 and in Fig. 7.7.

Table 7.7: Percentage of the Area of the Normal Distribution Lying within the Given Range

Number of Standard Deviations from Mean	*Approximate Percentage of Area under Normal Curve*
$x \pm \sigma$	68.26
$x \pm 2\sigma$	95.45
$x \pm 3\sigma$	99.75

The standard normal distribution is a symmetrical distribution and therefore

$$P(0 \le z \le a) = P(-a \le z \le 0) \text{ for any value } a.$$

For example,

$$P(1 \le z \le 2) = P(z \le 2) - P(z \le 1)$$
$$= 0.9772 - 0.8413 = 0.1359$$

The value of $P(1 \le z \le 2)$ is shown is Fig. 7.8.

7.6.2 Approximation of Binomial and Poisson Distributions to Normal Distribution

The binomial distribution approaches a normal distribution with standardized variable, that is,

where $$z = \frac{x - np}{\sqrt{npq}} \sim N(0, 1)$$

However this approximation works well when both $np \ge 10$ and $npq \ge 10$

Similarly, Poisson distribution also approaches a normal distribution with standardized variable, that is,

$$z = \frac{x - \lambda}{\sqrt{\lambda}} \sim N(0, 1)$$

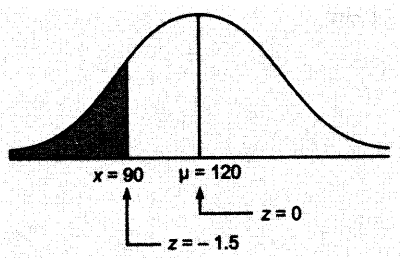
Example 7.23: 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

(a) How many bulbs will expire in less than 90 days?

(b) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10% should expire before replacement?

Solution: (a) Given, $\mu = 120$, $\sigma = 20$, and $x = 90$. Then

$$z = \frac{x - \mu}{\sigma} = \frac{90 - 120}{20} = -1.5$$



The area under the normal curve between $z = 0$ and $z = -1.5$ is 0.4332. Therefore area to the left of – 1.5 is 0.5 – 0.4332 = 0.0668. Thus the expected number of bulbs to expire in less than 90 days will be 0.0668 × 1000 = 67 (approx.).

(b) The value of z corresponding to an area 0.4 (0.5 – 0.10). Under the normal curve is 1.28. Therefore

$$z = \frac{x-\mu}{\sigma} \text{ or } -1.28 = \frac{x-120}{20} \text{ or } x = 120 - 20(-1.28) = 94$$

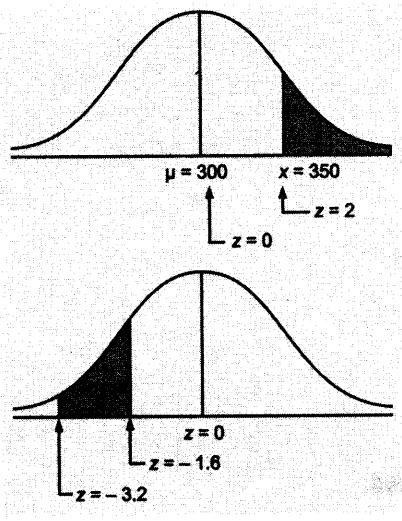
Hence, the bulbs will have to be replaced after 94 days.

Example 7.24: The lifetimes of certain kinds of electronic devices have a mean of 300 hours and standard deviation of 25 hours. Assuming that the distribution of these lifetimes, which are measured to the nearest hour, can be approximated closely with a normal curve

(a) Find the probability that any one of these electronic devices will have a lifetime of more than 350 hours.

(b) What percentage will have lifetimes of 300 hours or less?

(c) What percentage will have lifetimes from 220 or 260 hours?



Solution: (a) Given, $\mu = 300$, $\sigma = 25$, and $x = 350$. Then

$$z = \frac{x-\mu}{\sigma} = \frac{350-300}{25} = 2$$

The area under the normal curve between $z = 0$ and $z = 2$ is 0.9772. Thus the required probability is, 1 – 0.9772 = 0.0228.

$$\text{(b) } z = \frac{x-\mu}{\sigma} = \frac{300-300}{25} = 0$$

Therefore, the required percentage is, 0.5000 × 100 = 50%.

(c) Given, $x_1 = 220$, $x_2 = 260$, $\mu = 300$ and $\sigma = 25$. Thus

$$z_1 = \frac{220-300}{25} = -3.2 \text{ and } z_2 = \frac{260-300}{25} = -1.6$$

From the normal table, we have

$$P(z = -1.6) = 0.4452 \text{ and } P(z = -3.2) = 0.4903$$

Thus the required probability is

$$P(z = -3.2) - P(z = -1.6) = 0.4903 - 0.4452 = 0.0541$$

Hence the required percentage = 0.0541 × 100 = 5.41 per cent.

Example 7.25: In a certain examination, the percentage of passes and distinctions were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively (assume the distribution of marks to be normal).

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25 per cent of them should be given another opportunity of being examined.

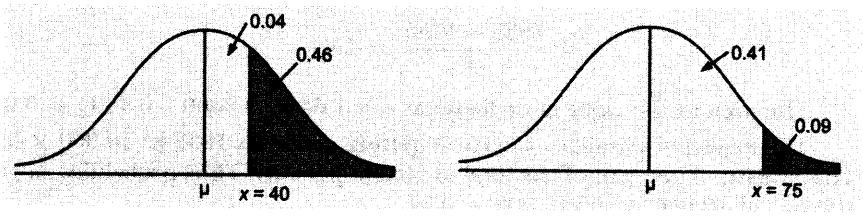
Solution: (a) Let μ be the mean and σ be the standard deviation of the normal distribution. The area to the right of the ordinate at $x = 40$ is 0.46 and hence the area between the mean and the ordinate at $x = 40$ is 0.04.

Now from the normal table, corresponding to 0.04, the standard normal variate, $z = 0.1$. Therefore, we have

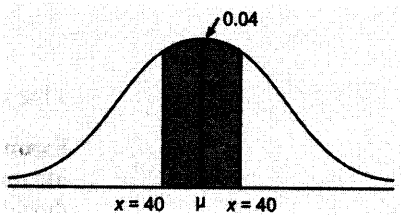
$$\frac{40-\mu}{\sigma} = 0.1 \text{ or } 40 - \mu = 0.1\sigma$$

Similarly, $\dfrac{75-\mu}{\sigma} = 1.34$ or $75 - \mu = 1.34\,\sigma$

Solving these equations, we get $\sigma = 28.23$ and $\mu = 37.18$ or 37.



(b) Let us assume that x_1 is the minimum qualifying marks for re-examination of the failed candidates.



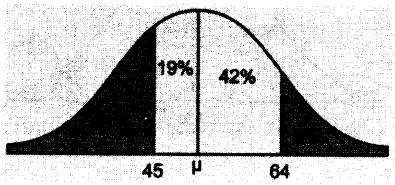
The area to the right of $x = 40$ is 46 per cent. Thus the percentage of students failing = 54 and this is the area to the left of 40. We want that the best 25 per cent of these failed candidates should be given a chance to reappear. Suppose this area is equal to the shaded area in the diagram. This area is, 25 per cent of 54 = 13.5 per cent = 0.1350.

The area between mean and ordinate at $x_1 = -(0.1350 - 0.04) = -0.0950$ (negative sign is included because the area lies to the left of the mean ordinates).

Corresponding to this area, the standard normal variate $z = -0.0378$. Thus, we write

$$\frac{x_1 - \mu}{\sigma} = -0.0378$$

$$x_1 = \mu - 0.0378\,\sigma$$
$$= 37.2 - (0.0378 \times 28.23)$$
$$= 37.2 - 1.067 = 36.133 \text{ or } 36 \text{ (approx.)}$$



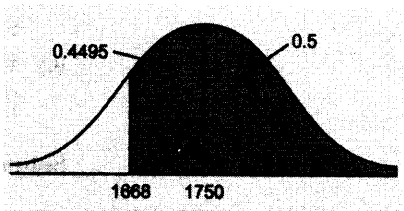
Example 7.26: In a normal distribution 31 per cent of the items are under 45 and 8 per cent are over 64. Find the mean and standard deviation of the distribution. *[Delhi Univ., MBA, 1999]*

Solution: Since 31 per cent of the items are under 45, therefore the left of the ordinate at $x = 45$ is 0.31, and obviously the area to the right of the ordinate up to the mean is (0.5 – 0.31) = 0.19. The value of z corresponding to this area is 0.5. Hence

$$z = \frac{45 - \mu}{\sigma} = -0.5 \text{ or } -\mu + 0.5\sigma = -45$$

As 8 per cent of the items are above 64, therefore area to the right of the ordinate at 64 is 0.08. Area to the left of the ordinate at $x = 64$ up to mean ordinate is (0.5 – 0.08) = 0.42 and the value of z corresponding to this area is 1.4. Hence

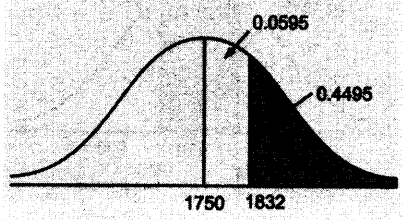
$$z = \frac{64 - \mu}{\sigma} = 1.4 \text{ or } -\mu - 1.4\sigma = -64$$



From these two equations, we get $1.9\,\sigma = 19$ or $\sigma = 10$. Putting $\sigma = 10$ in the first equation, we get $\mu - 0.5 \times 10 = 45$ or $\mu = 50$.

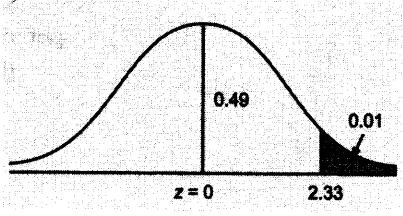
Thus, mean of the distribution is 50 and standard deviation 10.

Example 7.27: The income of a group of 10,000 persons was found to be normally distributed with mean Rs 1750 p.m. and standard deviation Rs 50. Show that of this group 95% had income exceeding Rs 1668 and only 5 per cent had income exceeding Rs 1832. What was the lowest income among the richest 100? *[Delhi Univ., MBA, 1997]*



Solution: (a) Given that, $x = 1668$, $\mu = 1750$ and $\sigma = 50$. Therefore, the standard normal variate corresponding to $x = 1668$, is

$$z = \frac{x - \mu}{\sigma} = \frac{1668 - 1750}{50} = -1.64$$



The area to the right of the ordinate at $z = -1.64$ (or $x = 1668$) is (0.4495 + 0.5000) = 0.9495 (because $z = -1.64$ to its right covers 95 per cent area).

The expected number of persons getting above Rs 1668 are $10{,}000 \times 0.9495$ = 9495. This is about 95 per cent of the total of 10,000 persons.

(ii) The standard normal variate corresponding to $x = 1832$ is

$$z = \frac{1832 - 1750}{50} = 1.64$$

The area to the right of ordinate at $z = 1.64$ is: 0.5000 – 0.4495 = 0.0505

The expected number of persons getting above Rs 1832 is: 10,000 × 0.0505 = 505. This is about 5 per cent of the total of 10,000 persons. Thus probability of getting richest 100 out of 10,000 is 100/10,000 = 0.01.

The standard normal variate having 0.01 area to its right is, $z = 2.33$. Hence

$$2.33 = \frac{x - 1750}{50}$$

$$x = 2.33 \times 50 + 1750 = \text{Rs } 1866 \text{ approx.}$$

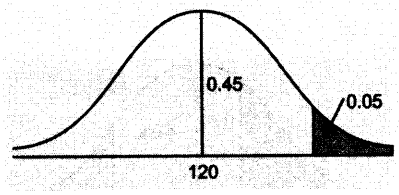
This implies that the lowest among the richest 100 is getting Rs 1866 per month.

Example 7.28: A wholesale distributor of fertilizer products finds that the annual demand for one type of fertilizer is normally distributed with a mean of 120 tonnes and standard deviation of 16 tonnes. If he orders only once a year, what quantity should be ordered to ensure that there is only a 5 per cent chance of running short?

[Delhi Univ., MBA, 1998, 2000]

Solution: Let x be the annual demand (in tonnes) for one type of fertilizer. Therefore

$$z = \frac{x - 120}{16}$$



The desired area of 5 per cent is shown in the figure. Since the area between the mean and the given value of x is 0.45, therefore from the normal table this area of 0.45 corresponds to $z = 1.64$.

Substituting this value of $z = 1.64$ in standard normal variate, we get

$$1.64 = \frac{x - 120}{16}$$

$$\text{or} \quad x = 120 + (1.64)\,(16) = 146.24 \text{ tonnes.}$$

If it is necessary to order in whole units, then the wholesale distributor should order 147 tonnes.

Example 7.29: Assume that the test scores from a college admissions test are normally distributed with a mean of 450 and a standard deviation of 100.

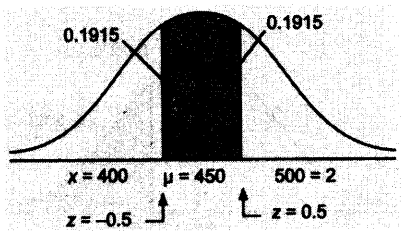
(a) What percentage of people taking the test score are between 400 and 500?

(b) Suppose someone received a score of 630. What percentage of the people taking the test score better? What percentage score worse?

(c) If a particular university will not admit any one scoring below 480, what percentage of the persons taking the test would be acceptable to the university?

[Delhi Univ. MBA, 2003]

Solution: (a) Given $\mu = 450$ and $\sigma = 100$. Let x be the test score. Then



$$z_1 = \frac{x - \mu}{\sigma} = \frac{500 - 450}{100} = 0.5$$

and

$$z_2 = \frac{x - \mu}{\sigma} = \frac{400 - 450}{100} = -0.5$$

The area under the normal curve between $z = 0$ and $z = 0.5$ is 0.1915

The required probability that the score falls between 400 and 500 is

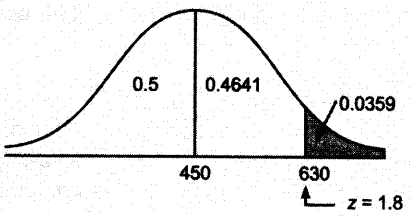
$$P(400 \le x \le 500) = P(-0.5 \le z \le 0.5) = 0.1915 + 0.1915 = 0.3830$$

So the percentage of the people taking the test score between 400 and 500 is 38.30 per cent.

(b) Given $x = 630$, $\mu = 450$ and $\sigma = 100$. Thus

$$z = \frac{x - \mu}{\sigma} = \frac{630 - 450}{100} = 1.8$$

The area under the normal curve between $z = 0$ and $z = 1.8$ is 0.4641.



The probability that people taking the test score better is given by

$$P(x \geq 630) = P(z \geq 1.8) = 0.5000 + 0.4641 = 0.9640$$

That is, 96.40 percent people score better

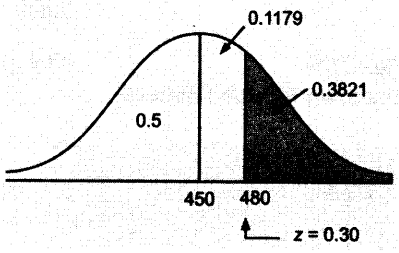
The probability that people taking the test score worse is given by

$$P(x \leq 630) = P(z \leq 1.8) = 0.5000 - 0.4641 = 0.0.0359$$

That is, 3.59 per cent people score worse

(c) Given $x = 480$, $\mu = 450$ and $\sigma = 100$. Thus

$$z = \frac{x-\mu}{\sigma} = \frac{480-450}{100} = 0.30$$



The area under the normal curve between $z = 0$ and $z = 0.30$ is 0.1179. So

$$P(x \geq 480) = P(z \geq 0.30) = 0.5000 + 0.1179 = 0.6179$$

The percentage of people who score more than 480 and are acceptable to the university is 61.79 per cent.

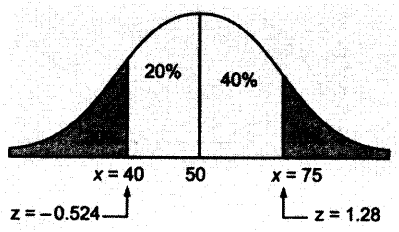
Example 7.30: The results of particular examination are given below in a summary form:

Result	*Per cent of Candidates*
• Passed with distinction	10
• Passed with out distinction	60
• Failed	30

It is known that a candidate fails in the examination if he obtains less than 40 markes (out of 100) while he must obtain at least 75 markes in order to pass with distinction. Determine the mean and standard deviation of the distribution of marks, assuming this to be normal.

Solution: The given data are illustrated in the figure

Since 30 per cent candidates who obtained less than 40 marks (out of 100) failed in the examination, from the figure we have



$$z = \frac{x-\mu}{\sigma} \text{ or } -0.524 = \frac{40-\mu}{\sigma} \text{ or } \mu - 0.524\sigma = 40$$

(Table value of z corresponding to 20 per cent area under the normal curve is 0.524)

Also 10 per cent condidates who obtained more than 75 marks passed with distinction, from the figure we have

$$z = \frac{x-\mu}{\sigma} \text{ or } 1.28 = \frac{75-\mu}{\sigma} \text{ or } \mu + 1.28\sigma = 75$$

(Table value of z corresponding to 40 per cent area under normal curve is 1.28)

Solving these equations, we get mean $\mu = 50.17$ and standard deviation $\sigma = 19.4$

7.6.3 Uniform (Rectangular) Distribution

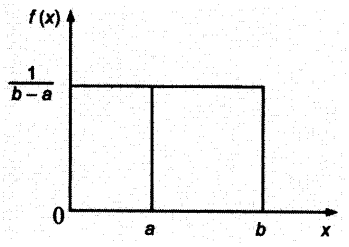
The simplest case of a continuous distribution is the uniform distribution. The general expression for the *pdf* (range of values) for a continuous random variable which is uniformly distributed over the interval between a to b is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

This distribution is also known as *constant distribution* because the probability is constant $[= 1/(b-a)]$ at every point of the interval (a, b) and is independent of whatever value the variable may take within the interval.

The general form of the rectangular probability distribution is shown in Fig. 7.9.

Figure 7.9
Rectangular Probability Distribution



The mean and variance of this distribution are given by

$$\text{Mean} = \frac{(b+a)}{2} \text{ and Variance} = \frac{(b+a)^2}{12}$$

This distribution is useful when the probability of occurrences of an event is constant whatever be the value of the variable, that is, all possible values of the continuous variable are assumed equally likely.

7.6.4 Exponential Probability Distribution

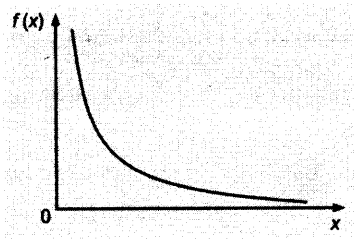
The probability density function (*pdf*) for exponential probability distribution is

$$f(x) = \begin{cases} \mu e^{-\mu x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where μ (>0) is a given parameter. This distribution is also referred to as *negative exponential distribution.* It is particularly useful in the queuing (waiting line) theory.

The graph of its *pdf* slopes downward to the right from its maximum at $x = 0$, where $f(x) = \mu$, as shown in Fig. 7.10.

Figure 7.10
Exponential Probability Distribution



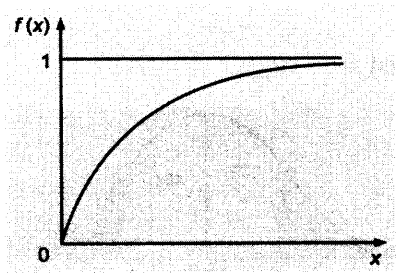
The exponential distribution has the mean, $1/\mu$ and variance, $1/\mu^2$. The *cumulative density function* (*cdf*) of the exponential distribution is

$$F(x) = \int_0^x \mu e^{-\mu x}\, dx$$

$$= \left[-e^{-\mu x}\right]_0^x = 1 - e^{-\mu x}$$

The graph of *cdf* is shown in Fig. 7.11.

Figure 7.11
Exponential Probability Distribution



The typical applications of *cdf* of exponential functions are found in representing a saturation phenomenon. That is, the situations where the effect of successive increments of the input x (e.g., size of advertising effort) show diminishing returns (e.g., resulting sales) as the total amount of x increases, and eventually, additional input increments have no effect.

Exponential distribution is closely related with the Poisson distribution. For example, if the Poisson random variable represents the *number of arrivals* per unit time at a service window, the exponential random variable will represent the *time between two successive arrivals.*

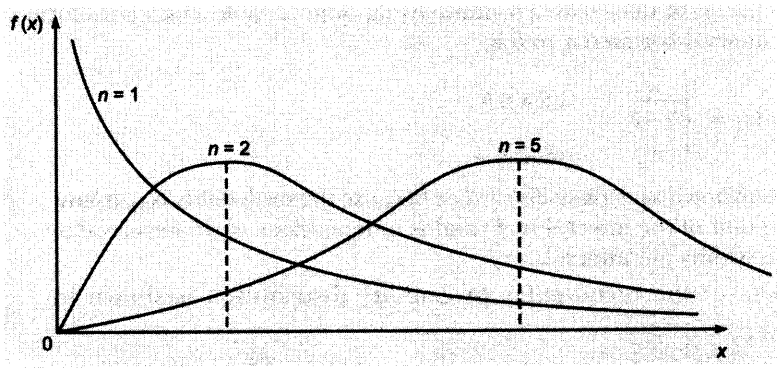
7.6.5 Gamma (or Erlang) Distribution

The probability density function (*pdf*) for the gamma (or Erlang) distribution is

$$f(x) = \frac{\mu(\mu x)^{n-1} e^{-\mu x}}{(n-1)!}, \; x > 0 \text{ and } \mu \geq 0$$

Gamma distribution is derived by the sum of n identically distributed and independent exponential random variables. Here it may be noted that the *pdf* of gamma distribution reduces to the exponential density function for $n = 1$. This means, the exponential distribution is the special case of the gamma distribution, where $n = 1$.

Figure 7.12
Gama Distribution pdf's for $\mu = 1$



The graphs of the *pdf*'s for the gamma distribution for $\mu = 1$ and selected values of n are shown in Fig. 7.12.

In the gamma distribution *pdf*'s, the parameter μ changes the relative scales of the two axes, and the parameter n determines the location of the peak of the curve. However, for all values of these two parameters, the area under the curve is equal to 1.

The expected value and variance of this distribution are: $E(x) = n/\mu$; $Var(x) = n/\mu^2$

7.6.6 Beta Distribution

The probability density function (*pdf*) for beta distribution is

$$f(x) = \frac{x^{m-1}(1-x)^{n-1}}{\beta(m,n)}; \; 0 \le x \le 1 \; ; \; m > 0; \; n > 0$$

where $$\beta(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx$$

is the beta function, whose value may be obtained directly from the table of the beta function.

The expected value and variance of random variable x in this case are given by

$$E(x) = \frac{m}{m+n} \quad \text{and} \quad Var(x) = \frac{mn}{(m+n)^2(m+n+1)}$$

This distribution is commonly used to describe the random variable whose possible values lie in a restricted interval of numbers. A typical use of this distribution is found in the use of PERT where activity times are estimated within a specific range.

Conceptual Questions 7D

20. State the conditions under which a binomial distribution tends to (i) Poisson distribution, (ii) normal distribution. Write down the probability functions of binomial and Poisson distributions.

21. Normal distribution is symmetric with a single peak. Does this mean that all symmetric distributions are normal? Explain.

22. When finding probabilities with a normal curve we always deal with intervals; the probability of a single value of x is defined equal to zero. Why is this so?

23. When finding a normal probability, is there a difference between the values of $P(a < x < b)$ and $P(a \le x \le b)$, where a and b represent two numbers? Why or why not?

24. What are the parameters of normal distribution? What information is provided by these parameters?

25. What are the chief properties of normal distribution? Describe briefly the importance of normal distribution in statistical analysis. *[Delhi Univ., MBA, 1990]*

26. Discuss the distinctive features of the binomial, Poisson, and normal distributions. When does a binomial distribution tend to become a normal distribution? *[Shukhadia Univ., MBA 1995; Kumaon Univ., MBA, 2000]*

27. Briefly describe the characteristics of the normal probability distribution. Why does it occupy such a prominent place in statistics?

Self-Practice Problems 7D

7.40 A cigarette company wants to promote the sales of X's cigarettes (brand) with a special advertising campaign. Fifty out of every thousand cigarettes are rolled up in gold foil and randomly mixed with the regular (special king-sized, mentholated) cigarettes. The company offers to trade a new pack of cigarettes for each gold cigarette a smoker finds in a pack of brand X. What is the probability that buyers of brand X will find X = 0, 1, 2, 3, . . . gold cigarettes in a single pack of 10?

7.41 You are in charge of rationing in a State affected by food shortage. The following reports were received from investigators:

Daily calories of food available per adult during current period

Area	*Mean*	*S.D.*
A	2000	350
B	1750	100

The estimated daily requirement of an adult is taken as

2500 calories and the absolute minimum is 1000. Comment on the reported figures and determine which area in your opinion needs more urgent attention.

7.42 Assume that on an average one telephone number out of fifteen is busy. What is the probability that if six randomly selected telephone numbers are called
(a) not more than three will be busy?
(b) at least three of them will be busy?

7.43 Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches. How many soldiers in a regiment of 1,000 would you expect to be over six feet tall?

7.44 The income of a group of 10,000 persons was found to be normally distributed with mean = Rs 750 p.m. and standard deviation = Rs 50. Show that in this group about 95 per cent had income exceeding Rs. 668 and only 5 per cent had income exceeding Rs 832. What was the lowest income among the richest 100?

[Delhi Univ., MBA, 1995]

7.45 In an intelligence test administered to 1000 students, the average score was 42 and standard deviation 24.
Find (a) the number of students exceeding a score of 50, (b) the number of students lying between 30 and 54, (c) the value of the score exceeded by the top 100 students.

7.46 A aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find:
(a) the number of candidates whose scores exceeds 58.
(b) the number of candidates whose scores lie between 30 and 66.

7.47 There are 600 business students in the postgraduate department of a university, and the probability for any student to need a copy of a particular textbook from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be greater than 0.90 that none of the students needing a copy from the library has to come back disappointed. (Use normal approximation to the binomial probability law.)

[Kurukshetra Univ., MCom, 1998]

7.48 A workshop produces 2000 units of an item per day. The average weight of units is 130 kg with a standard deviation of 10 kg. Assuming normal distribution, how many units are expected to weigh less than 142 kg?

[Delhi Univ., MBA, 1996]

7.49 Suppose a tire manufacturer wants to set a minimum kilometer guarantee on its new AT 100 tire. Tests reveal the mean kilometer is 47,900 with a standard deviation of 2050 kms and the distribution is a normal distribution. The namufacturer wants to set the minimum guaranteed kilometer so that no more than 4 percent of the tires will have to be replaced. What minimum guaranteed kilometer should the manufacturer announce?

7.50 The annual commissions per salesperson employed by a pharmaceutical company, which is a manufacturer of cough syrup, averaged Rs 40,000, with a standard deviation of Rs 5000. What per cent of the salespersons earn between Rs 32,000 and Rs 42,000?

7.51 Management of a company is considering adopting a bonus system to increases production. One suggestion is to pay a bonus on the highest 5 per cent of production based on past experience. Past records indicate that, on the average, 4000 units of a small assembly are produced during a week. The distribution of the weekly production is approximately normal with a standard deviation of 60 units. If the bonus is paid on the upper 5 per cent of production, the bonus will be paid on how many units or more?

Hints and Answers

7.40 An experiment, with $n = 10$ trials, probability of finding a golden cigarette (a success) is $p = 50/100 = 0.05$. The expected number of golden cigarettes per pack is, $\lambda = np = 10(0.05)$.

Number of Golden Cigarettes per Pack	*Probability*
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016

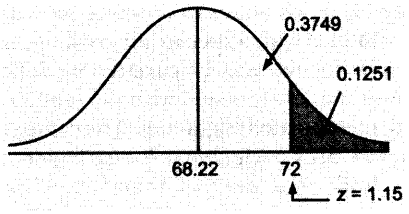
7.41

	Area A	*Area B*
	Mean ± 3σ	Mean ± 3σ
	= 2,000 ± (3 × 350)	= 1,750 ± (3 × 100)
or	950 and 3,050 calories	1,450 and 2,050 calories

Since the estimated requirement is minimum of 1,000 calories, area A needs more urgent attention.

7.43 Assuming that the distribution of height is normal. Given that, $x = 72$ inches, $\mu = 68.22$, $\sigma = \sqrt{10.8} = 3.286$. Therefore

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 68.22}{3.286} = 1.15$$



Area to the right of $z = 1.15$ from the normal table is $(0.5000 - 0.3749) = 0.1251$. Probability of getting

soldiers above six feet is 0.1251 and their expected number is 0.1251 × 1000 = 125.

7.44 Given, $x = 668$, $\mu = 750$, $\sigma = 50$. Therefore

$$z = \frac{x-\mu}{\sigma} = \frac{668 - 750}{50} = -1.64$$

Area to the right of $z = -1.64$ is (0.4495 + 0.5000) = 0.9495.

Expected number of persons getting above Rs 668 = 10,000 × 0.9495 = 9495, which is about 95 per cent of the total, that is, 10,000. Also,

$$z = \frac{832 - 750}{50} = 1.64$$

Area to the right of $z = 1.64$ is 0.5000–0.4495 = 0.0505

Expected number of persons getting above Rs 832 = 10,000 × 0.0505 = 505, which is approximately 5%.

Probability of getting richest 100 = 10/1000 = 0.01.

Value of standard normal variate for $z = 0.01$, to its right = 2.33

$$2.33 = \frac{x - 750}{50}$$

$$x = (2.33 \times 50) + 750 = \text{Rs } 866.5$$

Hence the lowest income of the richest 100 persons is Rs 866.50.

7.45 (a) Given $\mu = 42$, $x = 50$, $\sigma = 24$. Thus

$$z = \frac{x - \mu}{\sigma} = \frac{50 - 42}{24} = 0.333$$

Area to the right of $z = 0.333$ under the normal curve is 0.5 – 0.1304 = 0.3696

Expected number of children exceeding a score of 50 are 0.3696 × 1,000 = 370.

(b) Standard normal variate for score 30

$$z = \frac{x - \mu}{\sigma} = \frac{30 - 42}{24} = -0.5$$

Standard normal variate for score 54

$$z = \frac{x - \mu}{\sigma} = \frac{54 - 42}{24} = 0.5$$

Area between $z = 0$ to $z = 0.5 = 0.1915$

Area between $z = -0.5$ to $z = 0$ is 0.1915

Area between $z = -0.5$ to $z = 0.5$ is

$$0.1915 + 0.1915 = 0.3830$$

7.46 (a) Number of candidates whose score exceeds 58.

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 42}{24} = 0.667$$

Area to the right of $z = 0.667$ under the normal curve is (0.5 – 0.2476) = 0.2524

Number of candidates whose score exceeds 60 is: 100 × 0.2524 = 252.4 or 252

(b) Number of candidates whose score lies between 30 and 66.

Standard normal variate corresponding to 30,

$$z = \frac{30 - 42}{24} = -0.5$$

Standard normal variate corresponding to 66,

$$z = \frac{66 - 42}{24} = 1$$

Area between $z = -0.5$ and $z = 1$, is

$$0.1915 + 0.3413 = 0.5328$$

Number of candidates whose score lies between 30 and 66 : 1000 × 0.5328 = 532.8 or 533

7.47 Let n be the number of students and p the probability for an student to need a copy of a particular textbook from the university library. Given that $\mu = np = 600 \times 0.05 = 30$, $\sigma = \sqrt{npq} = \sqrt{600 \times 0.05 \times 0.95} = 5.34$.

Let x = number of copies of a textbook required on any day. Thus,

$$z = \frac{x - 30}{5.34} > 1.28 \text{ (95 per cent probability for } x)$$

$x - 30 > 6.835$, i.e. $x > 36.835 \cong 37$ (approx.)

Hence the library should keep at least 37 copies of the book to ensure that the probability is more than 90 per cent that none of the students needing a copy from the library has to come back disappointed.

7.48 Given N = 2000, $\mu = 130$, $\sigma = 10$ and $x = 142$,

$$z = \frac{x-\mu}{\sigma} = \frac{142-130}{10} = 1.2 \cong 0.3849$$

$P(x \le 142) = 0.5 + 0.3849 = 0.8849$

Expected number of units weighing less than 142 kg is 2000 × 0.8849 = 1,770 approx.

7.49 Given $\mu = 47{,}900$, $\sigma = 2050$

$$P(x \le 0.04) = P\left[z \le \frac{x - 47{,}900}{2050}\right] \text{ or } -1.75 = \frac{x - 47{,}900}{2050}$$

or $x = 44{,}312$

The area under the normal curve to the left of μ is 0.5. So the area between x and μ is 0.5 – 0.04 = 0.46 ≅ 0.4599. This area corresponds to $z = -1.75$.

7.50 Given $\mu = 40$, $\sigma = 5$. Thus $P(3200 \le x \le 42{,}000) =$

$$P\left[\frac{42-40}{5} \le z \le \frac{32-40}{5}\right] = P[0.40 \le z \le -1.60] =$$

0.1554 + 0.4452 = 0.6006, i.e. 60% approx.

7.51 $z = \frac{x-\mu}{\sigma}$ or $1.65 = \frac{x - 4000}{60}$ or $x = 4{,}099$ units.

Formulae Used

1. **Expected value of a random variable x**

$$E(x) = \Sigma\, x.P(x)$$

where x = value of the random variable
$P(x)$ = probability that the random variable will take on the value x.

2. Binomial probability distribution
   - Probability of r success in n Bernoulli trials

$$P(x = r) = {}^{n}C_r p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

where p = probability of success
q = probability of failure, $q = 1 - p$

   - Mean and standard deviation of binomial distribution

Mean $\mu = np$

Standard deviation $\sigma = \sqrt{npq}$

4. Poisson probability distribution
   - Probability of getting exactly r occurrences of random event

$$P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

where $\lambda = np$, mean number of occurrences per interval of time
e = 2.71828, a constant that represents the base of the natural logarithm system

   - Mean and standard deviation of Poisson distribution

$\lambda = np, \sigma = np$

5. Normal distribution formula:
Number of standard deviations σ a value of random variable x is away from the mean μ of normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

Chapter Concepts Quiz

True or False

1. The expected value of a random variable describes the long range weighted average of its values. (T/F)
2. The mean of the binomial distribution is greater than its variance. (T/F)
3. A binomial distribution is positively stewed when $p > 0.5$. (T/F)
4. The mean, median, and mode always coincide in the normal distribution. (T/F)
5. The expected value of a random variable is always a non-negative number. (T/F)
6. The binomial distribution is symmetrical for any value of p (probability of success). (T/F)
7. Poisson distribution generally describes arrivals at a service facility. (T/F)
8. In a Bernoulli process, the probability of success must equal the probability of failure. (T/F)
9. The symmetry of the normal distribution about its mean ensure that its tails extend indefinitely in both the positive and negative directions. (T/F)
10. All normal distributions are defined by two measures – the mean and the standard deviation. (T/F)
11. The expected value of a discrete random variable may be determined by taking an average of the values of the random variable. (T/F)
12. Within 2σ limits from mean, the area under a normal curve is 95.45 per cent. (T/F)
13. Any course of action that maximizes expected gain also minimizes expected loss. (T/F)
14. The value of normal variate for some value of the random variable x lying in a normal distribution is the area between x and the mean μ of the distribution. (T/F)
15. For a given binomial distribution with n fixed, if $p < 0.5$, then distribution will be skewed to the right. (T/F)

Multiple Choice

16. In a binomial distribution if n is fixed and $p > 0.5$, then
    (a) the distribution will be skewed to left
    (b) the distribution will be skewed to right
    (c) the distribution will be symmetric
    (d) cannot say anything
17. The binomial distribution is symmetric when:
    (a) $p < 0.5$ (b) $p > 0.5$
    (c) $p = 0.5$ (d) p has any value
18. Which of the following is the characteristic of the the probability distribution of a random variable?
    (a) $0 \le P(A_i) \le 1$, for all i
    (b) $\Sigma P(A_i) = 1$, for all i
    (c) The outcomes of each trial are independent
    (d) all of these
19. A Bernoulli process has all but which of the following properties?
    (a) each trial has only two possible outcomes
    (b) the probability of a success on any trial remains fixed over time
    (c) the probability of success on any trial is equal to the probability of failure
    (d) trials are statistically independent
20. The symmetry of the normal distribution about its mean indicates that:
    (a) the distribution is bell-shaped
    (b) the area under the curve on both sides of the mean is equal
    (c) the two tails extend indefinitely on either sides of the mean
    (d) all of the above
21. All normal distributions are:
    (a) bell-shaped

(b) symmetrical
(c) defined by its parameters μ and σ
(d) all of the above

22. For a Poisson distribution $P(x) = \frac{e^{-5}(5)^2}{2!}$, the mean value is:
(a) 2 (b) 5
(c) 10 (d) none of the above

23. For a binomial distribution $P(x) = {}^{10}C_r\,(0.5)^r\,(0.5)^{10-r}$, $r = 0, 1, 2, \ldots, 10$, the mean value is:
(a) 4 (b) 5
(c) 10 (d) 15

24. For a binomial distribution $P(x) = {}^{10}C_r\,(0.5)^r\,(0.5)^{10-r}$, $r = 0, 1, 2, \ldots, 10$, the standard deviation is:
(a) 2.5 (b) $\sqrt{2.5}$
(c) 2 (d) $\sqrt{2}$

25. For a normal curve with $\mu = 50$ and $\sigma = 4$, how much area will be to the left of μ?
(a) 50 per cent (b) 68.45 per cent
(c) 95.27 per cent (d) cannot be determined

26. The area under the normal curve covered within $\mu \pm 3\sigma$ limits is:
(a) 0.6827 (b) 0.9545
(c) 0.9973 (d) 1.000

27. For a standard normal probability distribution, the mean μ and standard deviation are:
(a) $\mu = 0, \sigma = 1$
(b) $\mu = 16, \sigma = 4$
(c) $\mu = 25, \sigma = 5$
(d) $\mu = 100, \sigma = 10$

28. The standard deviation of the binomial distribution is:
(a) np (b) $\sqrt{np}$
(c) npq (d) $\sqrt{npq}$

29. For a binomial distribution, the relationship between its mean μ and variance σ^2 is:
(a) $\mu > \sigma^2$ (b) $\mu < \sigma^2$
(c) $\mu = \sigma^2$ (d) none of these

30. For a normal distribution if $\mu = 30$, then its mode value is:
(a) 15 (b) 30
(c) 60 (d) none of these

31. Which of the following is a characteristic of the probability distribution for any random variable?
(a) The sum of all probabilities is 1
(b) A random variable may have more than one probability
(c) The mean always lies between the mode and the median
(d) The tails of the distribution extend infinitely without touching horizontal axis

32. Which of the following a necessary condition for use of a poisson distribution?
(a) Probability of an event in a short interval of time is constant
(b) The number of events in any interval of time is independent of successes in other intervals.
(c) Probability of two or more events in the short interval of time is zero
(d) All of these

33. The standard deviation of binomial distribution depends on
(a) probability of success (b) probability of failure
(c) number of trials (d) all of these

34. A binomial distribution may be approximated by a poisson distribution provided
(a) n is small and p is large
(b) n is large and p is small
(c) n is large and p is large
(d) n is small and p is small

35. Which is the characteristic of the normal distribution?
(a) For every pair of values μ and σ, the curve of the distribution is bell shaped and symmetric
(b) The mean of the normal distribution may be negative, or positive
(c) For any normal random variable x, $P(x \le \mu) = P(x \ge \mu) = 0.50$
(d) All of these

Concepts Quiz Answers

1. T	**2.** T	**3.** T	**4.** T	**5.** F	**6.** F	**7.** T	**8.** F	**9.** T
10. F	**11.** F	**12.** T	**13.** F	**14.** T	**15.** T	**16.** (a)	**17.** (c)	**18.** (d)
19. (c)	**20.** (d)	**21.** (d)	**22.** (b)	**23.** (b)	**24.** (b)	**25.** (a)	**26.** (c)	**27.** (a)
28. (d)	**29.** (a)	**30.** (b)	**31.** (a)	**32.** (d)	**33.** (d)	**34.** (b)	**35.** (d)	

Review Self-Practice Problems

7.52 Five hundred television sets are inspected as they come off the production line and the number of defects per set is recorded below:

Number of defects:	0	1	2	3	4
Number of sets:	368	72	52	7	1

Estimate the average number of defects per set and the

expected frequencies of 0, 1, 2, 3, and 4 defects assuming Poisson distribution.

[Sukhadia Univ., MBA; Delhi Univ., MBA, 1997]

7.53 The useful life of a certain brand of radial tyre has been found to follow a normal distribution with mean μ = 38,000 km and standard deviations = 3000 km. If a dealer orders 500 tyres for sale, then

(a) find the probability that a randomly chosen tyre will have a useful life of at least 35,000 km.

(b) find the approximate number of tyres that will last between 40,000 and 45,000 km.

(c) If an individual buys 2 tyres, then what is the probability that these tyres will last at least 38,000 km each?

7.54 The amount of time consumed by an individual at a bank ATM is found to be normally distributed with mean μ = 130 seconds and standard deviation σ = 45 seconds.

(a) What is the probability that a randomly selected individual will consume less than 100 seconds at the ATM?

(b) What is the probability that a randomly selected individual will spend between 2 to 3 minutes at the ATM?

(c) Within what length of time do 20 per cent of individuals complete their job at the ATM?

(d) What is the least amount of time required for individuals with top 5 per cent of required time?

7.55 An aptitude test for selecting officers in a bank was conducted on 1000 candidates. The average score is 42 and the standard deviation of scores is 24. Assuming normal distribution for the scores, find the

(a) number of candidates whose scores exceed 58.

(b) number of candidates whose scores lies between 30 and 66. *[Karnataka Univ., BCom, 1995]*

7.56 The mean inside diameter of a sample of 500 washers produced by a machine is 5.02 mm and the standard deviation is 0.05 mm. The purpose for which these washers are intended allows a maximum tolerance in the diameter of 4.96 to 5.08 mm, otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine, assuming the diameters are normally distributed.

7.57 In a binomial distribution consisting of 5 independent trials, the probability of 1 and 2 successes are 0.4096 and 0.2048, respectively. Find the parameter p of the distribution

7.58 If the probability of defective bolts be 1/10, find the following for the binomial distribution of defective bolts in a total of 400 bolts: (a) mean, (b) standard deviation, and (c) moment coefficient of skewness.

7.59 The probability of a bomb hitting a target is 0.20. Two bombs are enough to destroy a bridge. If 6 bombs are aimed at the bridge, find the probability that the bridge will be destroyed.

7.60 In an Indian university, it has been found that 25 per cent of the students come from upper income families (U), 35 per cent from middle income families (M), and 40 per cent from lower income families (L). A sample of 10 students is taken at random. What is the probability that the sample will contain 5 students from U, 2 from M and 3 from L?

7.61 The distribution of the total time a light bulb will burn from the time it is installed is known to be exponential with mean time between failure of the bulbs equal to 1000 hours. (a) What is the probability that a bulb will burn more than 1000 hours? and (b) what is the probability that the life will lie between 100 hours and 120 hours?

7.62 Past experience says that the average life of a bulb (assumed to be continuous random variable following exponential distribution) is 110 hours. Calculate the probability that the bulb will work for almost 25 hours.

[IGNOU, MS-51, 2001]

7.63 A firm uses a large fleet of delivery vehicles. Their record over a period of time (during which fleet size utilization may be assumed to have remained suitably constant) shows that the average number of vehicles unserviceable per day is 3. Estimate the probability on a given day when

(a) all vehicles will be serviceable.

(b) more than 2 vehicles will be unserviceable.

7.64 The director, quality control of automobile company, while conducting spot checking of automatic transmission, removed ten transmissions from the pool of components and checked for manufacturing defects. In the past, only 2 per cent of the transmissions had such flaws. (Assume that flaws occur independently in different transmissions.)

(a) What is the probability that sample contains more than two transmissions with manufacturing flaws?

(b) What is the probability that none of the selected transmission has any manufacturing flaw?

7.65 The Vice-President, HRD of an insurance company, has developed a new training programme that is entirely self-paced. New employees work at various stages at their own pace; completion occurs when the material is learned. The programme has been especially effective in speeding up the training process, as an employee's salary during training is only 67 per cent of that earned upon completion of the programme. In the last several years, the average completion time of the programme has been in 44 days, with a standard deviation of 12 days.

(a) What is the probability that an employee will finish the programme between 33 and 42 days?

(b) What is the probability of finishing the programme in fewer than 30 days?

7.66 In the past 2 months, on an average, only 3 per cent of all cheques sent for clearance by a Group Housing Welfare Society (GHWS) have bounced. This month, the GHWS received 200 cheques. What is the probability that exactly ten of these cheques bounced?

7.67 The sales manager of an exclusive shop that sells leather clothing decides at the beginning of the winter season, how many full-length leather coats to order. These coats cost 500 each, and will sell for 1000 each. Any coat left over at the end of the season will have to be sold at a 20 per cent discount in order to make room for summer inventory. From past experience, he knows that the demand for the coats has the following probability distribution:

Number of coats demanded:	8	10	12	14	16
Probability:	0.10	0.20	0.25	0.30	0.15

He also knows that there is never any problem with selling all leftover coats at discount.

(a) If he decides to order 14 coats, what is the expected profit?

(b) How would the answer to part (a) change if the leftover coats were sold at a 40 per cent discount?

7.68 Mr. Tiwari is campaign manager for a candidate for Lok Sabha. General impression is that the candidate has the support of 40 per cent of registered voters. A random sample of 300 registered voters shows that 34 per cent would vote for the candidate. If 40 per cent of voters really are allied with the candidate, what is the probability that a sample of 300 voters would indicate 34 per cent or fewer on his side? Is it likely that the 40 per cent estimate is correct?

Hints and Answers

7.52 No. of defects (x) : 0 1 2 3 4

No. of sets (f) : 368 72 52 7 1

$= 500\ (= N)$

fx : 0 72 104 21 4

$= 201\ (\Sigma fx)$

Average number of defects per set,

$$\lambda = \frac{\Sigma fx}{N} = \frac{201}{500} = 0.402.$$

Expected frequencies for 0, 1, 2, 3 and 4 defects are

$NP(x = 0) = 500\, e^{-\lambda} = 500\, e^{-0.402}$

$= 500 \times 0.6689 = 334.45$

$NP(x = 1) = NP(x=0) \times \lambda$

$= 334.45 \times 0.402 = 134.45$

$$NP(x = 2) = NP(x=1) \times \frac{\lambda}{2}$$

$$= 134.45 \times \frac{0.402}{2} = 27.02$$

$$NP(x = 3) = NP(x = 2) \times \frac{\lambda}{3}$$

$$= 27.02 \times \frac{0.402}{3} = 3.62$$

$$NP(x = 4) = NP(x = 3) \times \frac{\lambda}{4}$$

$$= 3.62 \times \frac{0.402}{4} = 3.36$$

7.53 (a) $z = \dfrac{x-\mu}{\sigma} = \dfrac{35{,}000 - 38{,}000}{3{,}000} = -1.00$

$P(x \geq 35{,}000) = P(z \geq -1.00)$

$= 0.500 + 0.3413 = 0.8413$

(b) $z_1 = \dfrac{x_1 - \mu}{\sigma} = \dfrac{40{,}000 - 38{,}000}{3{,}000} = 0.67$

$z_2 = \dfrac{x_1 - \mu}{\sigma} = \dfrac{45{,}000 - 38{,}000}{3{,}000} = 2.33$

$P(40{,}000 \leq x \leq 45{,}000) = P(0.67 \leq z \leq 2.33)$

$= 0.4901 - 0.2486 = 0.2415$

(c) $P(x \geq 38{,}000) = P(z \geq 0) = 0.500$

P(2 tyres each will last at least 38,000 km) $= (0.5000)^2 = 0.2500$ (based on the multiplication rule for the joint occurrence of independent events)

7.54 (a) $z = \dfrac{x-\mu}{\sigma} = \dfrac{100 - 130}{45} = -0.67$

$P(x < 100) = P(z < -0.67) = 0.5000 - 0.2486$

$= 0.2574$

(b) $z_1 = \dfrac{120 - 130}{45} = -0.22;$

$z_2 = \dfrac{180 - 130}{45} = 1.11$

$P(120 \leq x \leq 180) = P(-0.22 \leq z \leq 1.11)$

$= 0.0871 + 0.3655 = 0.4536$

(c) $x = \mu + z\sigma = 130 + (-0.84)\ 45 = 92$ seconds

(d) $x = \mu + z\sigma = 130 + (1.65)\ 45 = 204$ seconds

7.55 (a) $z = \dfrac{x-\mu}{\sigma} = \dfrac{58 - 42}{24} = 0.67$

$P(x > 58) = P(z > 0.67) = 0.5000 - 0.2476$

$= 0.2524$

Expected number of candidates whose score exceeds 58 is $1000\ (0.2524) = 2524$

(b) $z_1 = \dfrac{30 - 42}{24} = -0.50;\ z_2 = \dfrac{66 - 42}{24} = 1.00$

$P(30 \leq x \leq 66) = P(-0.50 \leq z \leq 1.00)$

$= 0.1915 + 0.3413 = 0.5328$

Expected number of candidates whose scores lie between 30 and 66 is $1000\ (0.5328) = 5328$.

7.56 $z_1 = \dfrac{x-\mu}{\sigma} = \dfrac{4.96 - 5.02}{0.05} = -1.20;$

$z_2 = \dfrac{5.08 - 5.02}{0.05} = 1.20$

$P(4.96 \leq x \leq 5.08) = P(-1.20 \leq z \leq 1.20)$

$= 2P(0 \leq z \leq 1.20) = 2(0.3849)$

$= 7698$ or 76.98%

Percentage of defective washers $= 100 - 76.98$

$= 23.02\%.$

7.57 Given $n = 5$; $f(x = 1) = {}^nC_1 p^n q^{n-1}$

$= {}^5C_1 p^1 q^4 = 5pq^4 = 0.4096$

$f(x = 2) = {}^nC_2 p^2 q^{n-2} = {}^5C_2 p^2 q^3 = 10p^2q^3 = 0.2048$

Thus $\dfrac{f(x=2)}{f(x=2)} = \dfrac{10\, p^2q^3}{5pq^4} = \dfrac{0.2048}{0.4096}$ or $\dfrac{2p}{q} = \dfrac{1}{2}$

or $4p = q\ (= 1 - p)$, i.e. $p = 1/5$

7.58 Given $n = 400, p = 1/10 = 0.10, q = 9/10 = 0.90$

(a) Mean $\mu = np = 400 \times (1/10) = 40$

(b) Standard deviation

$$\sigma = \sqrt{npq} = \sqrt{400\,(1/10)\,(9/10)} = 6$$

(c) Moment coefficient of skewness

$$= \frac{q-p}{\sqrt{npq}} = \frac{0.90-0.10}{6} = 0.133$$

7.59 Given $p = 0.20$, $q = 0.80$ and $n = 6$. The bridge is destroyed if at least 2 of the bombs hit it. The required probability is

$$\begin{aligned} P(x \geq 2) &= P(x=1) + P(x=2) + \ldots + P(x \geq 6) \\ &= 1 - [P(x=0) + P(x=1)] \\ &= 1 - [{}^6C_0\,(0.80)^6 + {}^6C_1\,(0.20)\,(0.80)^5] \\ &= 1 - \frac{2048}{3125} = 0.345 \end{aligned}$$

7.60 Required probability $= \dfrac{10!}{5!\,3!\,2!}(0.25)^2\,(0.35)^2\,(0.40)^3$

$= 0.0193$

(Based on the rule of multinomial rule of probability)

7.61 Given, mean time between failures $1/\lambda = 1000$ or $\lambda = 1/1000$ bulbs per hours ; $t = 1000$ hours

(a) $P(t > 1000) = 1 - P(x \leq 1000) = 1 - (1 - e^{\lambda t})$

$= e^{-\lambda t} = e^{-(1/1000)1000} = e^{-1} = 0.3680$

(b) $P(100 \leq t \leq 120) = (1 - e^{-\lambda t_1}) - (1 - e^{-\lambda t_2})$

$= \{1 - e^{-(1/1000)120}\}$

$\{1 - e^{-(1/1000)100}\}$

$= 0.1132 - 0.0952 = 0.018$

7.62 Given, mean life of a bulb $= 1/\lambda = 100$ or $\lambda = 1/100$; $t = 25$ hours

$F(t \leq T) = 1 - e^{-\lambda t} = 1 - e^{-(1/110)25} = 1 - e^{0.23}$

$= 1 - 0.7945 = 0.2055$

7.63 (a) $P(x=0) = \dfrac{e^{-3}\,(3)^0}{0!} = 0.0497$

(b) $P(x > 2) = 1 - P(x \leq 2)$

$= 1 - [P(x=0) + P(x=1) + P(x=2)]$

$= 1 - \left[e^{-3} + 3e^{-3} + \frac{9}{2}e^{-3}\right]$

$= 1 - e^{-3}\left(1 + 3 + \frac{9}{2}\right) = 1 - \frac{11}{2}(0.0497)$

$= 1 - 0.4224 = 0.5776$

7.64 (a) Given, $p = 0.02, q = 0.98, n = 10$

$P(x > 2 \text{ flaws}) = 1 - [P(x=0) + P(x=1) + P(x=2)]$

$= 1 - [{}^{10}C_0\,(0.98)^{10} + {}^{10}C_1\,(0.02)$

$(0.98)^9 + {}^{10}C_2\,(0.02)^2\,(0.98)^8]$

$= 1 - [0.8171 + 0.1667 + 0.0153]$

$= 0.0009$

(b) $P(x = 0 \text{ flaw}) = {}^nC_0 p^0 q^n = {}^{10}C_0\,(0.02)^0\,(0.98)^{10}$

$= 10\,(0.98)^{10} = 0.8171$

7.65 Given, average completion time of the programme, $\mu = 44$ days and standard deviation, $\sigma = 12$ days

(a) $P(33 \leq x \leq 42) = P\left[\dfrac{x_1 - \mu}{\sigma} \leq z \leq \dfrac{x_2 - \mu}{\sigma}\right]$

$= P\left[\dfrac{33-44}{12} \leq z \leq \dfrac{42-44}{12}\right]$

$= P[0.92 \leq z \leq -0.17]$

$= 0.3212 - 0.0675 = 0.2537$

(b) $P(x < 30) = P\left[z < \dfrac{x-\mu}{\sigma}\right] = P\left[z < \dfrac{30-42}{12}\right]$

$= P(z < -1.17) = 0.5000 - 0.3790$

$= 0.1210$

7.66 Given $n = 200, p = 0.03, \lambda = np = 200(0.03) = 6$.

$$P(x=10) = \frac{e^{-\lambda}\,\lambda^r}{r!} = \frac{e^{-6}\,(6)^{10}}{10!} = 0.0413$$

7.67 Expected profit when 14 coats are ordered

Number of coats ordered	:	8	10	12	14	
Probability	:	0.10	0.20	0.25	0.45	
(a) Profit per cast (Rs)	:	1160	1240	1320	1400	
Total expected profit (Rs)	:	116	248	330	630	= 1324
(b) Profit per Coat (Rs)	:	920	1080	1240	1400	
Total expected profit (Rs):		92	216	310	620	= 1248

7.68 Given $n=300$, $p=0.40$; $\mu=np=300(0.40)=120$;

$\sigma = \sqrt{npq} = \sqrt{120\,(06)} = 8.48$

$P(x \leq 0.34 \times 300 = 102)$

$= P\left[z < \dfrac{x-\mu}{\sigma}\right] = P\left[z < \dfrac{102-120}{8.48}\right]$

$= P\,[z \leq -2.12] = 0.5000 - 0.4830$

$= 0.0170$

Since the probability that the sample would indicate 34 per cent or less is very small, it is unlikely that the 40 per cent estimate is correct.

Chapter 8

By a small sample we may judge of the whole piece.
—Cervantes

Nine times out of ten, in the arts as in life, there is actually no truth to be discovered; there is only error to be exposed.
—H. L. Mencken

Sampling and Sampling Distributions

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- distinguish between population parameter and sample statistics
- apply the Central Limit Theorem
- know various procedures of sampling that provide an attractive means of learning about a population or process.
- develop the concept of a sampling distribution that helps you understand the methods and underlaying thinking of statistical inference.

8.1 INTRODUCTION

So far we introduced certain statistical methods (Chapters 2 to 5) to analyse a data set and the concepts of probability and its distributions (Chapters 6 and 7) to increase our knowledge about unknown features (characteristics) of a population or a process. In statistical inference we use random sample or samples to extract information about the population from which it is drawn. The information we extract is in the form of summary statistics: a sample mean, a sample standard deviation or other measures computed from the sample. Sample statistics are treated as estimator of population parameters – μ, σ, p, etc.

Sampling The process of selecting a sample from a population is called *sampling*. In sampling, a representative *sample* or *portion* of elements of a population or process is selected and then analysed. Based on sample results, called *sample statistics*, *statistical inferences* are made about the population characteristic. For instance, a political analyst selects specific or random set of peoples for interviews to estimate the proportion of the votes that each candidate may get from the population of voters; an auditor selects a sample of vouchers and calculates the sample mean for estimating population average amount; or a doctor examines a few drops of blood to draw conclusions about the nature of disease or blood constitution of the whole body.

8.2 REASONS OF SAMPLE SURVEY

A census is a count of all the elements in a population. Few examples of census are: population of eligible voters; census of consumer preference to a particular product, buying habits of adult Indians. Some of the reasons to prefer sample survey instead of census are given below.

1. ***Movement of Population Element*** The population of fish, birds, snakes, mosquitoes, etc. are large and are constantly moving, being born and dying. So instead of attempting to count all elements of such populations, it is desirable to make estimates using techinques such as counting birds at a place picked at random, setting nets at predetermined places, etc.
2. ***Cost and/or Time Required to Contact the Whole Population*** Time required to contact the whole population. A census involves a complete count of every individual member of the population of interest, such as persons in a state, households in a town, shops in a city, students in a college, and so on. Apart from the cost and the large amount of resources (such as enumerators, clerical assistance, etc.) that are required, the main problem is the time required to process the data. Hence the results are known after a big gap of time.
3. ***Destructive Nature of Certain Tests*** The census becomes extremely difficult, if not impossible, when the population of interest is either infinite in terms of size (number); constantly changing; in a state of movement; or observation results required destruction. For example, sometimes it is required to test the strength of some manufactured item by applying a stress until the unit breaks. The amount of stress that results in breakage is the value of the observation that is recorded. If this procedure is applied to an entire population, there would be nothing left. This type of testing is called destructive testing and requires that a sample be used in such cases.

8.3 TYPES OF BIAS DURING SAMPLE SURVEY

Not all surveys produce trust worthy results. Results based on a survey are *biased* when the method used to obtain those results would consistently produce values that are either too high or too low. Following are the common types of bias that might occur in surveys:

1. ***Undercoverage Bias*** This bias occur when a random sample chosen does not represent the population of interest. For instance, passengers at a railway station are surveyed to determine attitude towards buying station ticket, the results are not likely to represent all passengers at railway stations.
2. ***Non-response Bias*** This bias occurs when only a small number of respondents respond or return their questionnaire. The sample results would be biased because only those respondent who were particularly concerned about the subject chose to respond.
3. ***Wording Bias*** This bias occurs when respondents respond differently from how they truely feel. It may be due to the reason that the questionnaire contains questions that tend to confuse the respondents. For instance, questions for the survey about the use of drug, payment of income tax, abusive behavior, etc. must be worded and conducted carefully to minimise response bias.

8.3.1 Sampling and Non-Sampling Errors

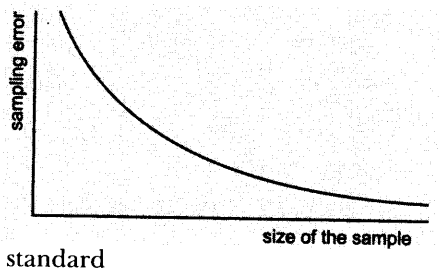
Sampling error The absolute value of the difference between an unbiased estimate and the corresponding population parameter, such as $|\bar{x}-\mu|$, $|\bar{p}-p|$, etc.

Any statistical inference based on sample results (statistics) may not always be correct, because sample results are either based on partial or incomplete analysis of the population features (or characteristics). This error is referred to as the **sampling error** because each sample taken may produce a different estimate of the population characteristic compared to those results that would have been obtained by a complete enumeration of the population. It is, therefore, necessary to measure these errors so as to have an exact idea about the reliability of sample-based estimates

of population features. The likelihood that a sampling error exceeds any specified magnitude must always be specified in terms of a probability value, say 5%. This acceptable margin of error is then used to produce a *confidence* in the decision-maker to arrive at certain conclusions with the limited data at his disposal. In general, in the business context, decision-makers wish to be 95 per cent or more confident that the range of values of sample results reflect the true characteristic of the population or process of interest.

Non-sampling errors arise during census as well as sampling surveys due to biases and mistakes such as (i) incorrect enumeration of population members, (ii) non-random selection of samples, (iii) use of incomplete, vague, or faulty questionnaire for data collection, or (iv) wrong editing, coding, and presenting of the responses received through the questionnaire. The sampling errors can be minimized if (i) the questionnaire contains precise and unambiguous questions, (ii) the questionnaire is administered carefully, (iii) the interviewers are given proper training, and (iv) the responses are correctly processed.

Figure 8.1
Measurement of Sampling Error



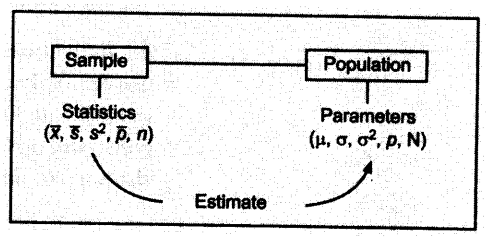
Measurement of Sampling Error A measure of sampling error is provided by the standard error of the estimate. Estimation of sampling error can reduce the element of uncertainty associated with interpretation of data. In most cases, the degree of precision or the level of error, would depend on the size of the sample.

The standard error of estimate is inversely proportional to the square root of the sample size. In other words, as the sample size increases, element of error is reduced. Figure 8.1 illustrates this concept.

8.4 POPULATION PARAMETERS AND SAMPLE STATISTICS

Parameters An exact, but generally unknown measure (or value) which describes the entire population or process characteristics is called a *parameter.* For example, quantities such as mean μ, variance σ^2, standard deviation σ, median, mode, and proportion p computed from a data set (also called population) are called parameters. A parameter is usually denoted with letters of the lower case Greek alphabet, such as mean μ and standard deviation σ.

Figure 8.2
Estimation Relationship between Sample and Population Measures



Sample Statistics A measure (or value) found from analysing sample data in called a *sample statistic* or simply a *statistic*. Inferential statistical methods attempt to estimate population parameters using sample statistics. **Sample statistics** are usually denoted by Roman letters such as mean $\bar{x}$, standard deviation s, variance s^2 and proportion $\bar{p}$.

Sample statistic A sample measure, such as mean $\bar{x}$, standard deviation, s, proportion $\bar{p}$, and so on.

The value of every statistic varies randomly from one sample to another whereas the value of a parameter is considered as constant. The value for statistic calculated from any sample depends on the particular random sample drawn from a population. Thus probabilities are attached to possible outcomes in order to assess the reliability or sample error associated with a statistical inference about a population based on a sample. Figure 8.2 shows the estimation relationships between sample statistics and the population parameters.

8.5 PRINCIPLES OF SAMPLING

The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:

(i) Principle of statistical regularity
(ii) Principle of inertia of large numbers

8.5.1 Principle of Statistical Regularity

This principle is based on the mathematical theory of probability. According to King, *'The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to process the characteristic of the large group.'* This principle, emphasises on two factors:

(i) ***Sample Size Should be Large*** As the size of sample inereases it becomes more and more representative of parent population and shows its characteristics. However, in actual practice, large samples are more expansive. Thus a balance has to be maintained between the sample size, degree of accuracy desired and financial resourses available.

(ii) ***Samples Must be Drawn Randomly*** The random sample is the one in which elements of the population are drawn in a such way that each combination of elements has an equal probability of being selected in the sample. When the term random sample is used without any specification, it usually refers to a *simple random sample.* The selection of samples based on this principle can reduce the amount of efforts required in arriving at a conclusion about the characteristic of a large population. For example, to understand the book buying habit of students in a college, instead of approaching every student, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

8.5.2 Principle of Inertia of Large Numbers

This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory. This principle states that, under similar conditions, *as the sample size (number of observations in a sample) get large enough, the statistical inference is likely to be more accurate and stable.* For example, if a coin is tossed a large number of times, then relative frequency of occurrence of head and tail is expected to be equal.

8.6 SAMPLING METHODS

As mentioned above, sampling methods compared to census provides an attractive means of learning about a population or process in terms of reduced cost, time and greater accuracy. The representation basis and the element selection techniques from the given population, classify several sampling methods into two categories as shown in Tabel 8.1.

Table 8.1: Types of Sampling Methods

Element Selection	*Representation Basis*	
	Probability (Random)	*Non-probability (Non-random)*
• Unrestrieted	Simple random sampling	Convenience sampling
• Restricted	Complex random sampling	Purposive sampling
	• Stratified sampling	• Quota sampling
	• Cluster sampling	• Judgement sampling
	• Systematic sampling	
	• Multi-stage sampling	

8.6.1 Probability Sampling Methods

Several probability sampling methods for selecting samples from a population or process are as follows:

Simple Random (Unrestricted) Sampling In this method, every member (or element) of the population has an equal and independent chance of being selected again and

again when a sample is drawn from the population. To draw a random sample, we need a complete list of all elements in the population of interest so that each element can be identified by a distinct number. Such a list is called *frame for experiment*. The frame for experiment allows us to draw elements from the population by randomly generating the numbers of the elements to be included in the sample.

For instance, in drawing the random sample of 50 students from a population of 3500 students in a college we make a list of all 3500 students and assign each student an identification number. This gives us a list of 3500 numbers, called frame for experiment. Then we generate by computer or by other means a set of 50 random numbers in the range of values from 1 and 3500. The procedure gives every set of 50 students in the population an equal chance of being included in the sample. Selecting a random sample is analogous to using a gambling device to generate numbers from this list.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with this method is that all elements of the population have to be available for selection, which many a times is not possible.

Stratified Sampling This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called *strata* that are relevent, appropriate and meaningful in the context of the study. A simple random sample, called a *sub-sample,* is then drawn from each *strata* or *group*, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum are disproportionate to the respective numbers in the population. The basis for forming the strata such as location, age, industry type, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

This sampling procedure is more efficient than the simple random sampling procedure because, for the same sample size, we get more representativeness from each important segment of the population and obtain more valuable and differentiated information with respect to each strata. For instance, if the president of a company is concerned about low motivational levels or high absentee rate among the employees, it makes sense to stratify the population of organizational members according to their job levels. Assume that the 750 employees were divided into six strata as shown in Table 8.2. Let 100 employees are to be selected for study, then number to be selected from each strata is shown in Table 8.2

Table 8.2: Proportionate and Disproportionate Stratified Random Samples

Strata	*Job Level*	*Number of Employees (Elements)*	*Number of Employees in the Sample*	
			Proportionate Sample	*Disproportionate Sample*
1	Top management	15	(15/750)×100 = 2	3
2	Middle-level management	30	(30/750)×100 = 4	10
3	Lower-level management	55	(55/750)×100 = 7	15
4	Supervisors	105	(105/750)×100 =14	25
5	Clerks	510	(510/750)×100 =68	37
6	Secretaries	35	(35/750)×100 = 5	10
		750	100	100

When the data are collected and the analysis completed, it is likely that the members of a particular group are found to be not motivated. This information will help in taking action at the right level and think of better ways to motivate that group members which otherwise would not have been possible.

Disproportionate sampling decisions are made either when strata are either too small, too large, or when there is more variability suspected within a particular stratum. For example, the educational levels in a particular strata might be expected to influence perceptions, so more people will be sampled at this level. Disproportionate sampling is done when it is easier, and less expensive to collect data from one or more strata than from others.

For this method of sampling to be more effective in terms of reliability, efficiency, and precision, any stratification should be done which ensures

(i) maximum uniformity among members of each strata,

(ii) largest degree of variability among various strata.

Cluster Sampling This method, sometimes known as *area sampling method*, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in the population so that each member can be identified by a distinct number). The entire population to be analysed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as *clusters*. The elements of a cluster are called *elementary units*. These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of a cluster.

If several groups with intragroup heterogeneity and intergroup homogeneity are found, then a random sampling of the clusters or groups can be done with information gathered from each of the elements in the randomly chosen clusters. Cluster samples offer more heterogeneity within groups and more homogeneity among groups—the reverse of what we find in stratified random sampling, where there is homogeneity within each group and heterogeneity across groups.

For instance, committees formed from various departments in an organization to offer inputs to make decisions on product development, budget allocations, marketing strategies, etc are examples of different clusters. Each of these clusters or groups contains a heterogeneous collection of members with different interests, orientations, values, philosophy, and vested interests. Based on individual and combined perceptions, it is possible to make final decision on strategic moves for the organization.

In summary, cluster sampling involves preparing only a list of clusters instead of a list of individual elements. For examples, (i) residential blocks (colonies) are commonly used to cluster in surveys that require door-to-door interviews, (ii) airlines sometimes select randomly a set of flights to distribute questionnaire to every passenger on those flights to measure customor satisfaction. In this situation, each flight is a cluster. It is much easier for the airline to choose a random sample of flights than to identify and locate a random sample of individual passengers to distribute questionnaire.

Multistage Sampling This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the *first stage* would be to choose as a sample a specific state (region). The size of the sample, that is the number of interviews, from each region would be determined by the relative populations in each region. In the *second stage*, a limited number of towns/cities in each of the regions would be selected, and then in the *third stage*, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is being drawn, it is customary to classify all the urban areas in the region in such

a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

Systematic Sampling This procedure is useful when elements of the population are already physically arranged in some order, such as an alphabelized list of people with driving licenses, list of bank customers by account numbers. In these cases one element is chosen at random from first k element and then every kth element (member) is included in the sample. The value k is called the *sampling interval.* For example, suppose a sample size of 50 is desired from a population consisting of 100 accounts receivable. The sampling interval is $k = N/n = 1000/50 = 20$. Thus a sample of 50 accounts is identified by moving systematically through the population and identifying every 20th account after the first randomly selected account number.

8.6.2 Non-Random Sampling Methods

Several non-random sampling methods for selecting samples from a population or process are as follows:

Convenience Sampling In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for his project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

Purposive Sampling Instead of obtaining information from those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets–respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

Judgement Sampling Judgement sampling involves the selection of respondents who are in the best position to provide the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conventiently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information.

However, the validity of the sample results depend on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

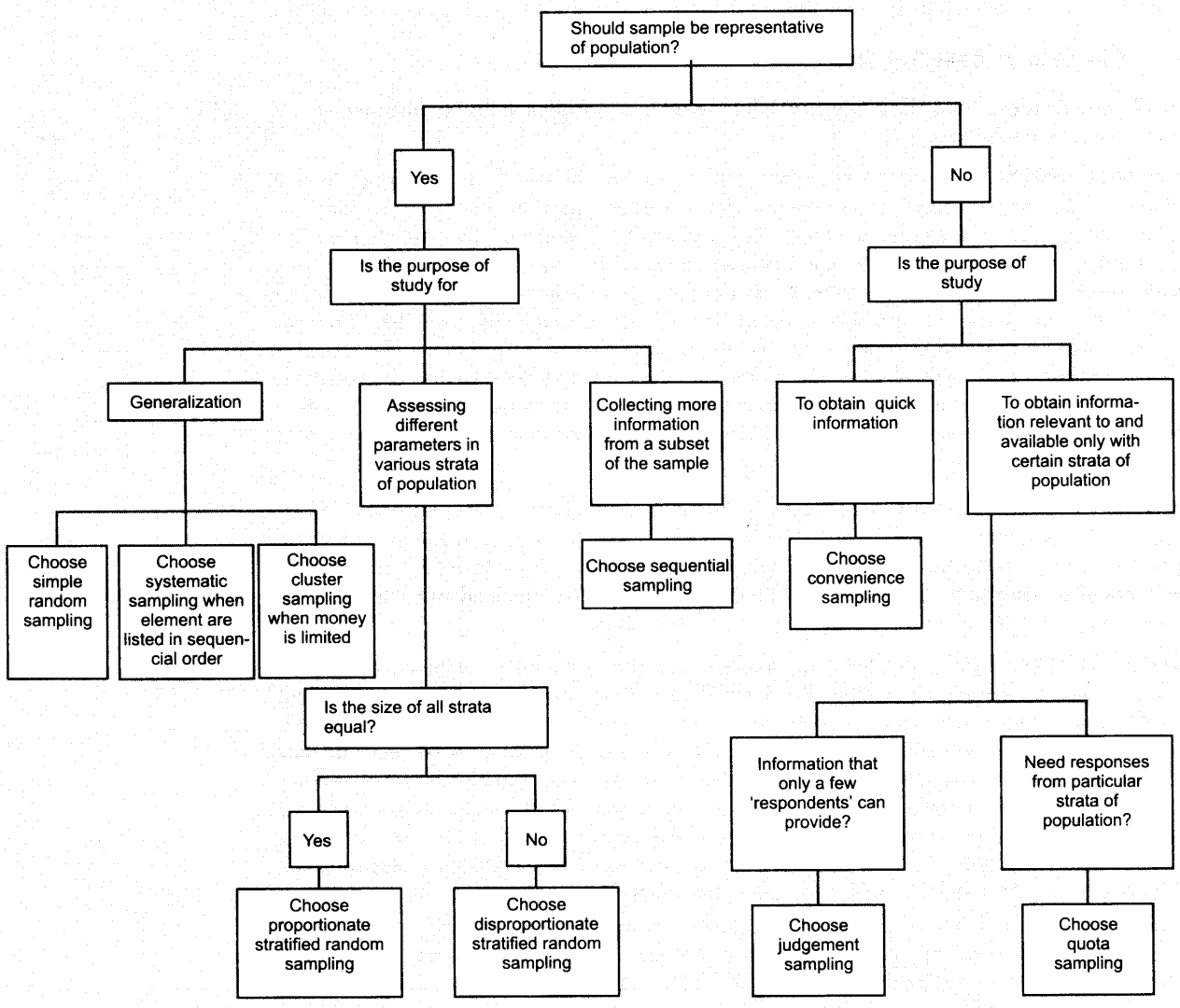
Quota Sampling Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which is essential for the study. For example, the investigator may choose to interview ten men and ten women in such a way that two of them have annual income of more than two lakh rupees five of them have annual income between one and two lakh rupees and thirteen whose annual income is below one lakh rupees. Furthermore, some of them should be between 25 and 35 years of age, others between 36 and 45 years of age, and the balance over 45 years. This means that the investigator's choice of respondent is partly dictated by these 'controls'.

Quota sampling has been criticized because it does not satisfy the fundamental requirement of a sample, that is, it should be random. Consequently, it is not possible to achieve precision of results on any valid basis.

8.6.3 Choice of Sampling Methods

The choice of particular sampling method (procedure) must be decided according to various factors such as: nature of study, size of the population, size of the sample, availability of resources, degree of precision desired, etc. A choice plan is shown in Fig. 8.3.

Figure 8.3
Guidelines to Choose Sample



Judging the Reliability of a Sample The reliability of a sample can be determined in the following ways to ensure dependable results:

(i) A number of samples may be taken from the same population and the results of various samples compared. If there is not much variation in the results of the different samples, it is a measure of its reliability.

(ii) Sub-sample may be taken from the main sample and studied. If the results of the sub-samples are similar to those given by the main sample it gives a measure of its reliability.

(iii) If some mathmatical properties are found in the distribution under study, the sample result can be compared with expected values obtained on the

basis of mathmatical relationship and if the difference between them is not significant, the sample has given dependable results.

In probability distributions where binomial, normal, Poisson or any other theoretical probability distribution is applicable, sample results can be compared with the expected values to get an idea about the reliability of the sample.

8.7 SAMPLING DISTRIBUTIONS

In Chapter 3 we have discussed several statistical methods to calculate parameters such as the mean and standard deviation of the population of interest. These values were used to describe the characteristics of the population. If a population is very large and the description of its characteristics is not possible by the census method, then to arrive at the statistical inference, samples of a given size are drawn repeatedly from the population and a particular '*statistic*' is computed for each sample. The computed value of a particular statistic will differ from sample to sample. In other words, if the same statistic is computed for each of the samples, the value is likely to vary from sample to sample. Thus, theoretically it would be possible to construct a frequency table showing the values assumed by the statistic and the frequency of their occurrence. This *distribution of values of a statistic is called a* **sampling distribution**, because the values are the outcome of a process of sampling. Since the values of statistic are the result of several simple random samples, therefore these are random variables.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

Suppose all possible random samples of size n are drawn from a population of size N, and the 'mean' values computed. This process will generate a set of ${}^{N}C_{n} = N!/n!\,(N - n)!$ sample means, which can be arranged in the form of a distribution. This distribution would have its mean denoted by $\mu_{\bar{x}}$ and standard deviation is denoted by $\sigma_{\bar{x}}$ (also called *standard error*). We may follow this procedure to compute any other statistic from all possible samples of given size drawn from a population.

The concept of sampling distribution can be related to the various probability distributions. Probability distributions are the theoretical distributions of random variables that are used to describe characteristics of populations or processes under certain specified conditions. That is, probability distributions are helpful in determining the probabilities of outcomes of random variables when populations or processes that generate these outcomes satisfy certain conditions. For example, if a population has a normal distribution, then the phenomenon that describes the normal probability distribution provides a useful description of the distribution of population values. Thus when mean values obtained from samples are distributed normally, it implies that this distribution is useful for describing the characteristics (or properties) of sampling distribution. Consequently, these properties, which are also the properties of sampling distribution, help to frame rules for making statistical inferences about a population on the basis of a single sample drawn from it, that is, without even repeating the sampling process. The sampling distribution of a sample statistic also helps in describing the extent of error associated with an estimate of the value of population parameters.

8.7.1 Standard Error of Statistic

Since sampling distribution describes how values of a sample statistic, say mean, is scattered around its own mean $\mu_{\bar{x}}$, therefore its standard deviation $\sigma_{\bar{x}}$ is called the *standard error* to distinguish it from the standard deviation σ of a population. The population standard deviation describes the variation among values of the members of the population, whereas the standard deviation of sampling distribution measures the variability among values of the sample statistic (such as mean values, proportion values) due to sampling errors. Thus knowledge of sampling distribution of a sample statistic enables us to determine the probability of sampling error of the given magnitude. Consequently standard deviation of sampling distribution of a sample statistic measures sampling error and is also known as *standard error of statistic*.

The standard error of statistic measures not only the amount of chance error in the sampling process but also the accuracy desired. One of the most common inferential procedures is *estimation* (discussed in Chapter 9). In estimation, the value of the statistic is used as an estimate of the value of the population parameter.

8.7.2 Distinction between Population, Sample Distributions, and Sampling Distributions

In the previous sections, we introduced sampling methods to drawn samples of the same size repeatedly from a population, and compute for each sample the statistic of interest. Hence from the distribution of population we can derive a *sampling distribution of statistic of interest.* This distribution has its own mean and standard deviation (also called *standard error*). Such distributions describe the relative frequency of occurrence of values of a sample statistic and hence help to estimate universal parameters.

Population distribution is the distribution of values of its elements members and has mean denoted by μ, variance σ^2 and standard deviation σ.

Sample distribution is the distribution of measured values of statistic in random samples drawn from a given population. Each sample distribution is a discrete distribution [as shown in Fig 8.4(b)] because the value of the sample mean would vary from sample to sample. This variability serves as the basis for the random sampling distribution. In Fig. 8.4(b) only five such samples are shown, however, there could be several such cases. In such distributions the arithmetic mean represents the average of all possible sample means or the 'mean of means' denoted by $\bar{x}$; the standard deviation which measures the variability among all possible values of the sample values, is considered as a good approximation of the population's standard deviations σ. To estimate σ of the population to greater accuracy the formula

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \text{ is used instead of } s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

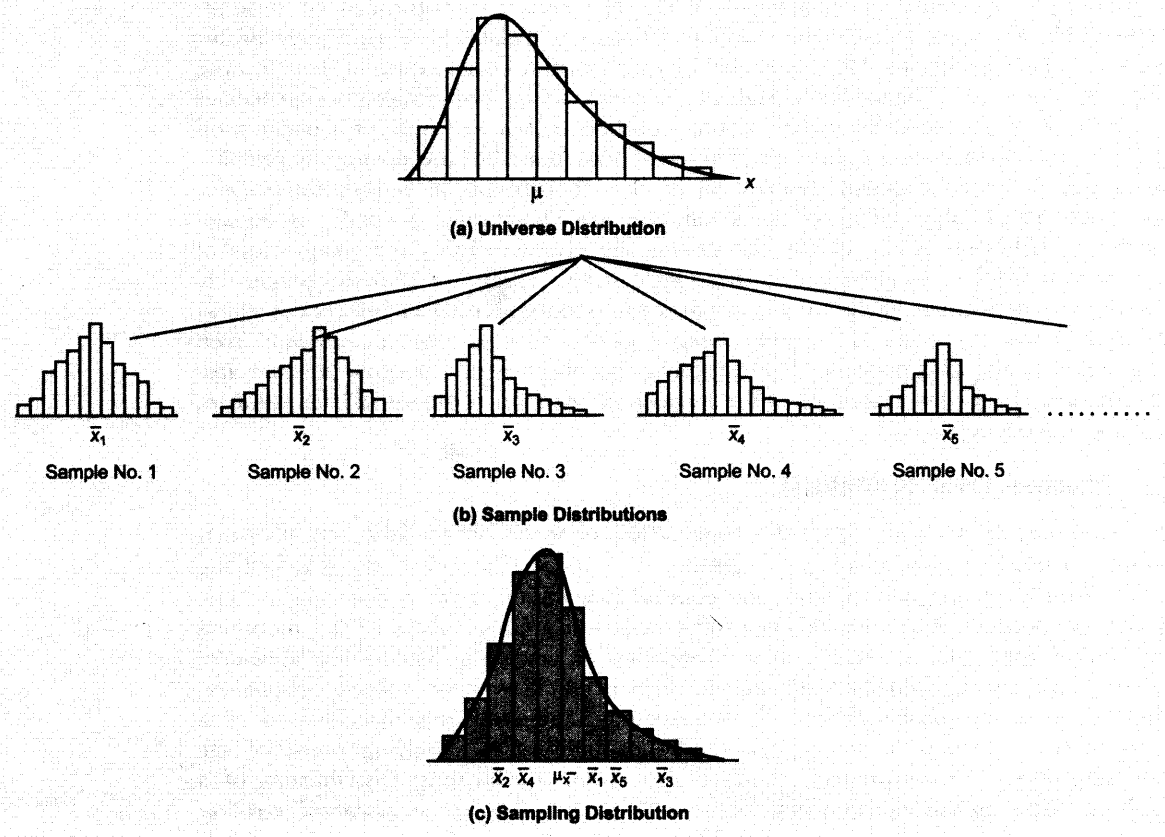


Figure 8.4
Random Sampling Distribution of Sample Mean

where n is the size of sample. The new value $n - 1$ in the denominator results into higher value of s than the observed value s of the sample. Here $n - 1$ is also known as *degree of freedom.* The number of degrees of freedom, $df = n - 1$ indicate the number of values that are free to vary in a random sample.

Sampling distribution is the distribution of all possible values of a statistic from all the distinct possible samples of equal size drawn from a population or a process as shown in Fig. 8.4(c). The sampling distribution of the mean values has its own arithmetic mean denoted by $\mu_{\bar{x}}$ (mu sub x bar) or $\bar{\bar{x}}$ (mean of mean values) and standard deviation $\sigma_{\bar{x}}$ (sigma sub x bar) or s. The standard deviation of the sampling distribution indicates how different samples would be distributed. The calculation of these sampling distribution statistics is based on the following properties:

(i) The arithmetic mean $\mu_{\bar{x}}$ of sampling distribution of mean values is equal to the population mean μ regardless of the form of population distribution, that is, $\mu_{\bar{x}} = \mu$.

(ii) The sampling distribution has a standard deviation (also called standard error or sampling error) equal to the population standard deviation divided by the square root of the sample size, that is, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Remember that a standard deviation is the spread of the values around the average in a *single sample*, where as the standard error is the spread of the averages around the average of averages in a *sampling distribution.*

(iii) The sampling distribution of sample mean values from normally distributed populations is the normal distribution for samples of all sizes.

Sampling error provides some idea of the precision of a statistical estimate. A low sampling error means a relatively less variability or range in the sampling distribution. Since we never actually see the sampling distribution, calculation of sampling error is based on the caluculation of the standard deviation of the sample. Thus greater the sample's standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size: greater the sample size, smaller the standard error

A sample of size $n \geq 30$ is generally considered to be a large sample for statistical analysis whereas a sample of size $n < 30$ is considered to be a small sample. It may be noted from the formula of $\sigma_{\bar{x}}$ that its value tends to be smaller as the size of sample n increases and vice-versa.

When standard deviation σ of population is not known, the standard deviation s of the sample, which closely approximates σ value, is used to compute standard error, that is, $\sigma_{\bar{x}} = s/\sqrt{n}$.

Conceptual Questions 8A

1. Briefly explain
   (a) The fundamental reason for sampling
   (b) Some of the reasons why a sample is chosen instead of testing the entire population
2. What is the relationship between the population mean, the mean of a sample, and the mean of the distribution of the sample mean?
3. Is it possible to develop a sampling distribution for other statistics besides sample mean? Explain.
4. How does the standard error of mean measure sampling error? Is the amount of sampling error in the sample mean affected by the amount of variability in the universe? Explain.
5. If only one sample is selected in a sampling problem, how is it possible to have an entire distribution of the sample mean?
6. What is sampling? Explain the importance in solving business problems. Critically examine the well-known methods of probability sampling and non-probability sampling. *[Delhi Univ., MBA, 1998]*
7. Point out the differences between a sample survey and a census survey. Under what conditions are these undertaken? Explain the law which forms the basis of sampling. *[Delhi Univ., MBA, 1999]*
8. Explain with the help of an example, the concept of sampling distribution of a sample statistic and point out its role in managerial decision-making. *[Delhi Univ., MBA, 2000]*

9. Why does the sampling distribution of mean follow a normal distribution for a large sample size even though the population may not be normally distributed?
10. Explain the concept of standard error. Discuss the role of standard error in large sample theory.
11. What do you mean by sampling distribution of a statistic and its standard error? Give the expressions for the standard error of the sample mean.
12. Bring out the importance of sampling distribution and the concept of standard error in statistical application.
13. Explain the principles of 'Inertia of Large Numbers' and 'Statistical Regularity'.
14. Enumerate the various methods of sampling and describe two of them mentioning the situations where each one is to be used.
15. Distinguish between sampling and non-sampling errors. What are their sources? How can these errors be controlled?
16. (a) What is the distinction between a sampling distribution and a probability distribution?
    (b) What is the distinction between a standard deviation and a standard error?
17. Is the standard deviation of sampling distribution of mean the same as the standard deviation of the population? Explain.
18. Explain the terms 'population' and 'sample'. Explain, why is it sometimes necessary and often desirable to collect information about the population by conducting a sample survey instead of complete enumeration?
19. What are the main steps involved in a sample survey. Discuss different sources of error in such surveys and point out how these errors can be controlled.

8.8 SAMPLING DISTRIBUTION OF SAMPLE MEAN

In general, the sampling distribution of sample means depending on the distribution of the population or process from which samples are drawn. If a population or process is normally distributed, then sampling distribution of sample means is also normally distributed regardless of the sample size. Even if the population or process in not distributed normally, the sampling distribution of sample mean tends to be distributed normally as the sample size is sufficiently large.

8.8.1 Sampling Distribution of Mean When Population has Non-Normal Distribution

Central limit theorem A result that enables the use of normal probability distribution to approximate the sampling distribution of $\bar{x}$ and $\bar{p}$.

If population is not normally distributed, then we make use of the **central limit theorem** to describe the random nature of the sample mean for large samples without knowledge of the population distribution. The Central Limit Theorem states that

- When the random samples of observations are drawn from a non-normal population with finite mean μ and standard deviation σ, and as the sample size n is increased, the sampling distribution of sample mean $\bar{x}$ is approximately normally distributed, with mean and standard deviation as:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{x}}$$

Regardless of its shape, the sampling distribution of sample mean $\bar{x}$ always has a mean identical to the sampled population, i.e. $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{x}$. This implies that *the spread of the distribution of sample means is considerably less than the spread of the sampled population.*

The central limit theorem is useful in statistical inference. When the sample size is sufficiently large, estimations such as 'average' or 'proportion' that are used to make inferences about population parameters are expected to have sampling distribution that is approximately normal. The behaviour of these estimations can be described in repeated sampling and are used to evaluate the probability of observing certain sample results using the normal distribution as follows:

$$\text{Standard normal random variable, } z = \frac{\text{Estimator} - \text{Mean}}{\text{Standard deviation}}$$

As stated in the central limit theorem that the approximation to normal distribution is valid as long as the sample size is sufficiently 'large' – but how large? There is no clear understanding about the size of n. However, following guidelines are helpful in deciding an appropriate value of n:

(i) If the sampled population is *normal*, then the sampling distribution of mean $\bar{x}$ will also be normal, regardless of the size of sample.

(ii) If the sampled population is approximately *symmetric*, then the sampling distribution of mean $\bar{x}$ becomes approximately normal for relatively small values of n.

(iii) If the sampled population is *skewed*, the sample size n must be larger, with at least 30 before the sampling distribution of mean $\bar{x}$ becomes approximately normal

Thus the standard normal variate, $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ approximate the standard normal distribution, where μ and σ are the population mean and standard deviation, respectively.

8.8.2 Sampling Distribution of Mean When Population has Normal Distribution

Population Standard Deviation σ is Known As mentioned earlier that no matter what the population distribution is, for any given sample of size n taken from a population with mean μ and standard deviation σ, the sampling distribution of a sample statistic, such as mean and standard deviation are defined respectively by

- Mean of the distribution of sample means or expected value of the mean $\mu_{\bar{x}}$ or $E(\bar{x}) = \mu$
- Standard deviation (or error) of the distribution of sample means or standard error of the mean $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

If all possible samples of size n are drawn *with replacement* from a population having normal distribution with mean μ and standard deviation σ, then it can be shown that the sampling distribution of mean $\bar{x}$ and standard error $\sigma_{\bar{x}}$ will also be normally distributed irrespective of the size of the sample. This result is true because any linear combination of normal random variables is also a normal random variable. In particular, if the sampling distribution of $\bar{x}$ is normal, the standard error of the mean $\sigma_{\bar{x}}$ can be used in conjunction with normal distribution to determine the probabilities of various values of sample mean. For this purpose, the value of sample mean $\bar{x}$ is first converted into a value z on the standard normal distribution to know how any single mean value deviates from the mean $\bar{x}$ of sample mean values, by using the formula

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

since $\sigma_{\bar{x}}$ measures the dispersion (standard deviation) of values of sample means in the sampling distribution of the means, it can be said that

- $\bar{x} \pm \sigma_{\bar{x}}$ covers about the middle 68 per cent of the total possible sample means
- $\bar{x} \pm 1.96\,\sigma_{\bar{x}}$ covers about the middle 95 per cent of the total possible sample means

The procedure for making statistical inference using sampling distribution about the population mean μ based on mean $\bar{x}$ of sample means is summarized as follows:

- If the population standard deviation σ value is known and either
  (a) population distribution is normal, or
  (b) population distribution is not normal, but the sample size n is large ($n \geq 30$), then the sampling distribution of mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, is very close to the standard normal distribution given by

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- If the population is finite with N elements whose mean is μ and varience. is $\sigma^2$and the samples of fixed size n are drawn *without replacement*, then the standard deviation (also called standard error) of sampling distribution of mean $\bar{x}$ can be modified to adjust the continued change in the size of the population N due to the several draws of samples of size n as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Finite population correction factor The term $\sqrt{(N-n)/(N-1)}$ is multiplied with for $\sigma_{\bar{x}}$ and $\sigma_{\bar{p}}$ a finite population is being sampled. In general, ignore the finite population correction factor whenever $n/N \leq 0.05$.

The term $\sqrt{(N-n)/(N-1)}$ is called the **finite population multiplier or finite correction factor**. In general, this factor has little effect on reducing the amount of sampling error when the size of the sample is less than 5 per cent of the population size. But if N is large relative to the sample size n, $\sqrt{(N-n)/(N-1)}$ is approximately equal to 1.

Population Standard Deviation σ is Not Known While calculating standard error $\sigma_{\bar{x}}$ of normally distributed sampling distribution, so far we have assumed that the population standard deviation σ is known. However, if σ is not known, the value of the normal variate z cannot be calculated for a specific sample. In such a case, the standard deviation of population σ must be estimated using the sample standard deviation σ. Thus the standard error of the sampling distribution of mean $\bar{x}$ becomes

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Since the value of $\sigma_{\bar{x}}$ varies according to each sample standard deviation, therefore instead of using the conversion formula

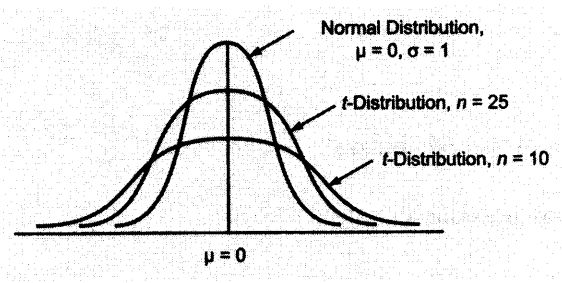
$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

we use following formula, called 'Student's t-distribution'

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

where $s = \sqrt{\Sigma(x-\bar{x})^2/(n-1)}$.

Figure 8.5
Comparison of t-Distributions with Standard Normal Distribution



In contrast to the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, the t-distribution is a family of symmetrical distributions centred around the mean $\mu = 0$. The shape of the distribution depends on two statistics: $\bar{x}$ and s. However, the value of s varies with sample size n. The higher the sample size n, s will be a more accurate estimate of population standard deviation σ and vice-versa. Figure 8.5. illustrates a comparison of t-distribution with that of the standard normal distribution

Degrees of freedom The number of unrestricted chances for variation in the measurement being made.

Degrees of Freedom The devisor $(n-1)$ in the formula for the sample variance s^2 is called number of *degress of freedom* (*df*) associated with s^2. The number of degrees of freedom refers to the *number of unrestricted chances for variation in the measurement being made, i.e.* number of independent squared deviations in s^2 that are available for estimating σ^2. In other words, it refers to the number of values that are free to vary in a random sample. The shape of t-distribution varies with **degrees of freedom**. Obviously more is the sample size n, higher is the degrees of freedom.

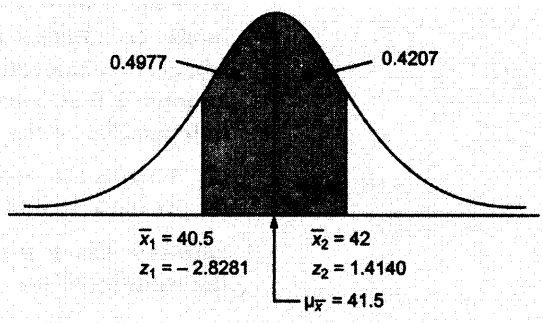
Example 8.1: The mean length of life of a certain cutting tool is 41.5 hours with a standard deviation of 2.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours? *[Delhi Univ., MBA, 2003]*

Solution: We are given the following information

μ = 41.5 hours, σ = 2.5 hours, and n = 50

It is required to find the probability that the mean length of life, $\bar{x}$, of the cutting tool lies between 40.5 hours and 42 hours, that is, $P(40.5 \leq \bar{x} \leq 42)$.

Figure 8.6
Normal curve



Based upon the given information, the statistics of the sampling distribution are computed as:

$$\mu_{\bar{x}} = \mu = 41.5$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{50}} = \frac{2.5}{7.0711} = 0.3536$$

The population distribution is unknown, but sample size $n = 50$ is large enough to apply the central limit theorem. Hence, the normal distribution can be used to find the required probability as shown by the shaded area in Fig. 8.6.

$$P(40.5 \leq \bar{x} \leq 42) = P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} \leq z \leq \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}}\right]$$

$$= P\left[\frac{40.5 - 41.5}{0.3536} \leq z \leq \frac{42 - 41.5}{0.3536}\right]$$

$$= P[-2.8281 \leq z \leq 1.4140]$$

$$= P[z \geq -2.8281] + P[z \leq 1.4140]$$

$$= 0.4977 + 0.4207 = 0.9184$$

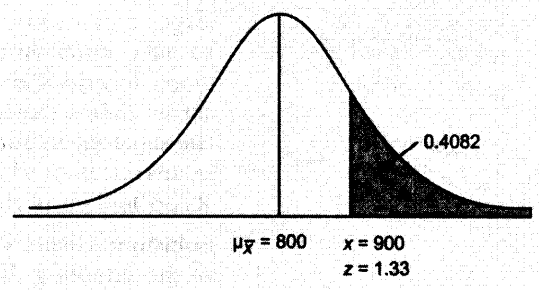
Thus 0.9184 is the probability of the tool of having a mean life between the required hours.

Example 8.2: A continuous manufacturing process produces items whose weights are normally distributed with a mean weight of 800 gms and a standard deviation of 300 gms. A random sample of 16 items is to be drawn from the process.

(a) What is the probability that the arithmetic mean of the sample exceeds 900 gms? Interpret the results.

(b) Find the values of the sample arithmetic mean within which the middle 95 per cent of all sample means will fall.

Figure 8.7
Normal curve



Solution: (a) We are given the following information

$$\mu = 800 \text{ g}, \ \sigma = 300 \text{ g}, \text{ and } n = 16$$

Since population is normally distributed, the distribution of sample mean is normal with mean and standard deviation equal to

$$\mu_{\bar{x}} = \mu = 800$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{16}} = \frac{300}{4} = 75$$

The required probability, $P(\bar{x} > 900)$ is represented by the shaded area in Fig. 8.7 of a normal curve. Hence

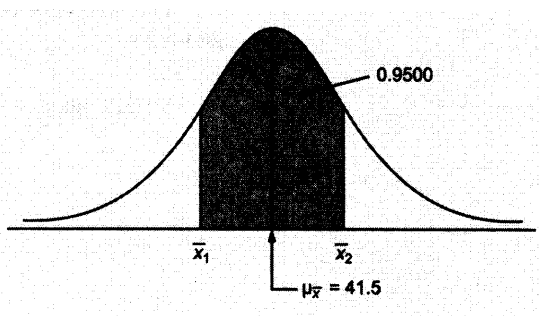
$$P(\bar{x} > 900) = P\left[z > \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{900 - 800}{75}\right]$$

$$= P[z > 1.33]$$

$$= 0.5000 - 0.4082 = 0.0918$$

Hence, 9.18 per cent of all possible samples of size $n = 16$ will have a sample mean value greater than 900 g.

Figure 8.8
Normal curve



(b) Since $z = 1.96$ for the middle 95 per cent area under the normal curve as shown in Fig. 8.8, therefore using the formula for z to solve for the values of $\bar{x}$ in terms of the known values are as follows:

$$\bar{x}_1 = \mu_{\bar{x}} - z\sigma_{\bar{x}}$$

$$= 800 - 1.96(75) = 653 \text{ g}$$

and

$$\bar{x}_2 = \mu_{\bar{x}} + z\sigma_{\bar{x}}$$

$$= 800 + 1.96(75) = 947 \text{ g}$$

Example 8.3: An oil refinary has backup monitors to keep track of the refinary flows continuously and to prevent machine malfunctions from disrupting the process. One particular monitor has an average life of 4300 haurs and a standard deviation of 730 haurs. In addition to the primary monitor, the refinary has set up two standby units, which are duplicates of the primary one. In the case of malfunction of one of the monitors, another will automatically take over in its place. The operating life of each monitor is independent of the other.

(a) What is the probability that a given set of monitors will last at least 13,000 hours?
(b) At most 12,630 hours?

Solution: Given, $\mu = 4300$ hours, $\sigma = 730$ hours, $n = 3$. Based upon the given information the statistics of the sampling distribution are computed as:

$$\text{Mean, } \mu_{\bar{x}} = \mu = 4300$$

and $$\text{Standard deviation, } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{x}} = \frac{730}{\sqrt{3}} = \frac{730}{1.732} = 421.48$$

(a) For a set of monitors to last 13,000 hours, they must each last 13,000/3 = 4333.33 hours on average. The required probability is caluculated as follow:

$$P(\bar{x} \geq 4.333.33) = P\left[\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \geq \frac{4333.33-4300}{421.48}\right]$$

$$= P\,[\,z \geq 0.08] = 0.5 - 0.0319 = 0.4681$$

(b) For the set to last at most 12,630 hours, the average life can not exceed 12,630/3 = 4210 hours. The required probability is calculated as follows:

$$P(\bar{x} \leq 4210) = P\left[\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \leq \frac{4210-4300}{421.48}\right]$$

$$= P\,[\,z \leq -0.213] = 0.5 - 0.0832 = 0.4168$$

Example 8.4: Big Bazar, a chain of 130 shoping malls has been bought out by another larger nationwide supermarket chain. Before the deal is finalized, the larger chain wants to have some assurance that Big Bazar will be a consistent money maker. The larger chain has decided to look at the financial records of 25 of the Big Bazar outlets. Big Bazar claims that each outlet's profits have an apporximately normal distribution with the same mean and a standard deviation of Rs 40 million. If the Big Bazar management is correct, then what is the probability that the sample mean for 25 outlets will fall within Rs 30 million of the actual mean?

Solution: Given N = 130, $n = 25$, $\sigma = 40$. Based upon the given information the statistics of the sampling distribution are compated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{40}{\sqrt{25}}\sqrt{\frac{130-25}{130-1}}$$

$$= \frac{40}{5}\sqrt{\frac{105}{129}} = 8 \times 0.902 = 13.72$$

The probability that the sample mean for 25 stores will fall within Rs 30 million is given by

$$P(\mu - 30 \leq \bar{x} \leq \mu + 30) = P\left[\frac{-30}{13.72} \leq \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \leq \frac{30}{13.72}\right]$$

$$= P\,(-2.18 \leq z \leq 2.18)$$

$$= 0.4854 + 0.4854 = 0.9708$$

Example 8.5: Chief Executive officer (CEO) of a life insurance company wants to undertake a survey of the huge number of insurance policies that the company has underwritten. The company makes a yearly profit on each policy that is distributed with mean Rs 8000 and standard deviation Rs 300. It is desired that the survey must be large enough to reduce the standard error to no more than 1.5 per cent of the population mean. How large should sample be?

Solution: Given μ = Rs 8000, and σ = Rs 300. The aim is to find sample size n be large enough so that

$$\text{Standard error of estimate, } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \leq 1.5 \text{ per cent of Rs } 8000$$

$$\text{or} \qquad \frac{300}{\sqrt{x}} \leq 0.015 \times 8000 = 120$$

$$300 \leq 120\sqrt{n} \text{ or } \sqrt{n} \geq 25, \text{ or } n \geq 625$$

Thus, a sample size of at least 625 insurance policies is needed.

Example 8.6: Safal, a tea manufacturing company is interested in determining the consumption rate of tea per household in Delhi. The management believes that yearly consumption per household is normally distributed with an unknown mean μ and standard deviation of 1.50 kg

(a) If a sample of 25 household is taken to record their consumption of tea for one year, what is the probability that the sample mean is within 500 gms of the population mean?

(b) How large a sample must be in order to be 98 per cent certain that the sample mean is within 500 gms of the population mean?

Solution: Given μ = 500 gms, n = 25 and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.5/\sqrt{25} = 0.25$ kg.

(a) Probability that the sample mean is within 500 gms or 0.5 kg of the population mean is calculated as follows:

$$P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) = P\left[\frac{-0.5}{\sigma/\sqrt{n}} \leq z \leq \frac{0.5}{\sigma/\sqrt{n}}\right]$$

$$= P\left[\frac{-0.5}{0.25} \leq z \leq \frac{0.5}{0.25}\right] = P[-2 \leq z \leq 2]$$

$$= 0.4772 + 0.4772 = 0.9544$$

(b) For 98 per cent confidence, the sample size is calculated as follows:

$$P(\mu - 0.5 \leq \bar{x} \leq \mu + 0.5) = P\left[\frac{-0.5}{1.5/\sqrt{n}} \leq z \leq \frac{0.5}{1.5/\sqrt{n}}\right]$$

Since z = 2.33 for 98 per cent area under normal curve, therefore

$$2.33 = \frac{0.5}{1.5/\sqrt{n}} \quad \text{or} \quad 2.33 = 0.33\sqrt{n}$$

$$n = (2.33/0.33)^2 = 49.84$$

Hence, the management of the company should sample at least 50 households.

Example 8.7: A motorcycle manufacturing company claims that its particular brand of motorcycle gave an average highway km per litre rating of 90. An independent agency tested it to verify the claim. Under controlled conditions, the motorcycle was driven for a distance of 100 km on each of 25 different occasions. The actual kms per litre achieved during the trip were recorded on each occasion. Over the 25 trials, the average and standard deviation turned out to be 87 and 5 kms per litre, respectively. It is believed that the distribution of the actual highway km per litre for this motorcycle is close to a normal distribution.

If the rating of 90 km per litre of the agency is correct, find the probability that the average kms per litre over a random sample of 25 trials would be 87 or less.

Solution: Since the population standard deviation σ is unknown, t-Student's test will be applicable to calculate the desired probability $P(\bar{x} \leq 87)$ as follows:

$$P(\bar{x} \leq 87) = P\left[t \leq \frac{\bar{x} - \mu}{s/\sqrt{n}}\right] = P\left[t \leq \frac{87 - 90}{5/\sqrt{25}}\right]$$

$$= P[t \leq -3]$$

with degrees of freedom $(n - 1) = (25 - 1) = 24$.

The desired probability of $t \leq -3.00$ with $df = 24$ from t-distribution table is 0.0031. Hence, the probability that the average km per litre is less than or equal to 87 is very small.

8.8.3 Sampling Distribution of Difference Between Two Sample Means

The concept of sampling distribution of sample mean introduced earlier in this chapter can also be used to compare a population of size N_1 having mean μ_1 and standard deviation σ_1 with another similar type of population of size N_2 having mean μ_2 and standard deviation σ_2.

Let $\bar{x}_1$ and $\bar{x}_2$ be the mean of sampling distribution of mean of two populations, respectively. Then the difference between their mean values μ_1 and μ_2 can be estimated by generalizing the formula of standard normal variable as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{\bar{x}_1} - \mu_{\bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

where $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$ (mean of sampling distribution of difference of two means)

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$ (Standard error of sampling distribution of two means)

n_1, n_2 = independent random samples drawn from first and second population, respectively.

Since random samples are drawn independently from two populations with replacement, therefore the sampling distribution of the difference of two means $\bar{x}_1 - \bar{x}_2$ will be normal provided sample size is sufficiently large.

The standard error of sampling distributions of some other statistic is given below:

Sampling Distribution	Standard Error and Mean	Remarks
• Median	$\sigma_{Med} = 1.2533 \frac{\sigma}{\sqrt{n}}$ $\mu_{Med} = \mu$	• For a large sample size $n \geq 30$, the sampling distribution of median approaches normal distribution. This result is true only if the population is normal or approximately normal.
• Sample standard deviation	(i) $\sigma_s = \frac{\sigma}{\sqrt{2n}}$ (ii) $\sigma_s = \sqrt{\frac{\mu_4 - \mu_2^2}{4n\mu_2}}$ (iii) $\mu_s = \sigma$	• For a large sample size $n \geq 100$, the sampling distribution is close to normal distribution. • If population is normaly distributed, then σ_s is calculated using (i), otherwise (ii). • μ_2 and μ_4 are second and fourth moments, where $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$.

Example 8.8: Car stereos of manufacturer A have a mean lifetime of 1400 hours with a standard deviation of 200 hours, while those of manufacturer B have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If a random sample of 125 stereos of each manufacturer are tested, what is the probability that manufacturer A's stereos will have a mean lifetime which is at least (a) 160 hours more than manufacturer B's stereos and (b) 250 hours more than the manufacturer B's stereos? *[Delhi Univ., MBA, 1999]*

Solution: We are given the following information

Manufacturer A: $\mu_1 = 1400$ hours, $\sigma_1 = 200$ hours, $n_1 = 125$

Manufacturer B: $\mu_2 = 1200$ hours, $\sigma_2 = 100$ hours, $n_2 = 125$

Thus, $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 1400 - 1200 = 200$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}} = \sqrt{80 + 320} = \sqrt{400} = 20$$

(a) Let $\bar{x}_1 - \bar{x}_2$ be the difference in mean lifetime of stereo manufactured by the two manufacturers. Then we are required to find the probability that this difference is more than or equal to 160 hours, as shown in Fig. 8.9. That is,

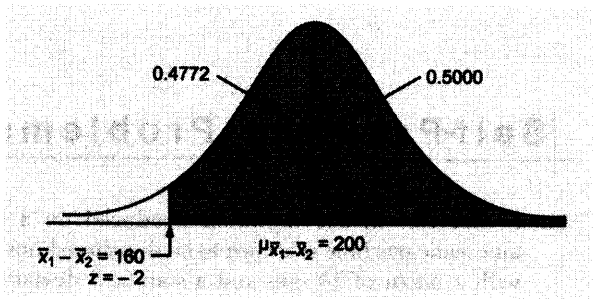


Figure 8.9
Normal curve

$$P[(\bar{x}_1 - \bar{x}_2) \geq 160] = P\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]$$

$$= P\left[z \geq \frac{160 - 200}{20}\right]$$

$$= P[z \geq -2]$$

$$= 0.5000 + 0.4772 = 0.9772 \text{ (Area under normal curve)}$$

Hence, the probability is very high that the mean lifetime of the stereos of A is 160 hours more than that of B.

(b) Proceeding in the same manner as in part (a) as follows:

$$P[(\bar{x}_1 - \bar{x}_2) \geq 250] = P\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]$$

$$= P\left[z \geq \frac{250 - 200}{20}\right] = P[z \geq 2.5]$$

$$= 0.500 - 0.4938$$

$$= 0.0062$$

(Area under normal curve)

Hence, the probability is very less that the mean lifetime of the stereos of A is 250 hours more than that of B as shown in Fig. 8.10.

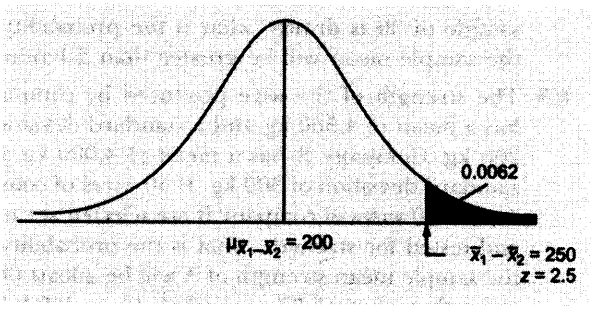


Figure 8.10
Normal curve

Example 8.9: The particular brand of ball bearings weighs 0.5 kg with a standard deviation of 0.02 kg. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2 gms.

Solution: We are given the following information

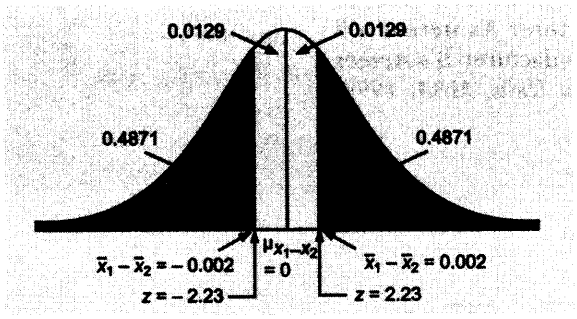
Lot 1: $\mu_{\bar{x}_1} = \mu_1 = 0.50$kg; $\sigma_1 = 0.02$ kg and $n_1 = 100$

Lot 2: $\mu_{\bar{x}_2} = \mu_2 = 0.50$kg; $\sigma_2 = 0.02$ kg and $n_1 = 100$

Thus $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 0$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

Figure 8.11
Normal curve



A difference of 2 gms in two lots is equivalent to a difference of 2/100 = 0.002 kg in mean weights. It is possible if $\bar{x}_1 - \bar{x}_2 \leq 0.002$ or $\bar{x}_1 - \bar{x}_2 \geq -0.002$. Then the required probability that each ball bearing will differ by more than 2 gms is calculated as follows and shown in Fig. 8.11

$$P\left[-0.002 \leq \bar{x}_1 - \bar{x}_2 \leq 0.002\right]$$

$$= P\left[\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \leq z \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]$$

$$= P\left[\frac{-0.002}{0.000895} \leq z \leq \frac{0.002}{0.000895}\right]$$

$$= P\left[-2.33 \leq z \leq 2.33\right]$$

$$= 2\left[0.5000 - 0.4871\right] = 0.0258$$

Self-Practice Problems 8A

8.1 A diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If a random sample of size 5 is picked up, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm?

[Delhi Univ., MBA, 1997]

8.2 The time between two arrivals at a queuing system is normally distributed with a mean of 2 minutes and standard deviation 0.25 minute. If a random sample of 36 is drawn, what is the probability that the sample mean will be greater than 2.1 minutes?

8.3 The strength of the wire produced by company A has a mean of 4,500 kg and a standard deviation of 200 kg. Company B has a mean of 4,000 kg and a standard deviation of 300 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will be atleast 600 kg more than that of B? *[Delhi Univ., MBA, 2000]*

8.4 For a certain aptitude test, it is known from past experience that the average score is 1000 and the standard deviation is 125. If the test is administered to 100 randomly selected individuals, what is the probability that the value of the average score for this sample will lie in the interval 970 and 1030? Assume that the population distribution is normal.

8.5 A manufacturing process produces ball bearings with mean 5 cm and standard deviation 0.005 cm. A random sample of 9 bearings is selected to measure their average diameter and find it to be 5.004 cm. What is the probability that the average diameter of 9 randomly selected bearings would be at least 5.004 cm?

8.6 A population of items has an unknown distribution but a known mean and standard deviation of 50 and 100, respectively. Based upon a randomly drawn sample of 81 items drawn from the population, what is the probability that the sample arithmetic mean does not exceed 40?

8.7 A marketing research team has determined the standard error of sampling distribution of mean for a proposed market research sample size of 100 consumers. However, this standard error is twice the level that the management of the organization considers acceptable. What can be done to achieve an acceptable standard error for mean?

8.8 Assume that the height of 300 soldiers in an army batallion are normally distributed with mean 68 inches and standard deviation 3 inches. If 80 samples consisting of 25 soldiers each are taken, what would be the expected mean and standard deviation of the resulting sampling distribution of means if the sampling is done (a) with replacement and (b) without replacement?

8.9 How well have equity mutual funds performed in the past compared with BSE Stock Index? A random sample of 36 funds averages a 16.9 per cent annual investment return for 2001–2 with a standard deviation of 3.6 per cent annual return. The BSE Stock Index grew at an annual average rate of 16.3 per cent over the same period. Do these data show that, on the average, the mutual funds out-performed the BSE Stock Index during this period?

8.10 The average annual starting salary for an MBA is Rs 3,42,000. Assume that for the population of MBA (Marketing majors), the average annual starting salary is μ = 3,40,000 and the standard deviation is σ = 20,000. What is the probability that a simple random sample of MBA (Marketing majors) will have a sample mean within ± Rs 2,500 of the population mean for each sample sizes: 50,100 and 200? What is your conclusion? *[Delhi Univ., MBA, 2003]*

Hints and Answers

8.1 Given $\mu_{\bar{x}} = \mu = 10, \sigma = 0.1$ and $n = 10$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.1/\sqrt{5} = 0.047$$

$$P[9.95 \le \bar{x} \le 10.05]$$

$$= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}_1}} \le z \le \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}_2}}\right]$$

$$= P\left[\frac{9.95 - 10}{0.047} \le z \le \frac{10.05 - 10}{0.047}\right]$$

$$= P[-1.12 \le z \le 1.12]$$

$$= P[z \ge -1.12] + P[z \le 1.12]$$

$$= 0.3686 + 0.3686 = 0.7372$$

8.2 Given $\mu_{\bar{x}} = \mu = 2, \sigma = 0.25$ and $n = 36$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}\,\frac{3,37,000 - 3,40,000}{20,000/\sqrt{50}} = 0.25/\sqrt{36} = 0.042$$

$$P[\bar{x} \ge 2.1] = P\left[z \ge \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \ge \frac{2.1 - 2}{0.042}\right]$$

$$= P[z \ge 2.38] = 0.5000 - 0.4913$$

$$= 0.0087$$

8.3 Given $\mu_1 = 4500, \sigma_1 = 200$ and $n_1 = 50$; $\mu_2 = 4000, \sigma_2 = 300$ and $n_2 = 100$. Then

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 4500 - 4000 = 500$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40,000}{50} + \frac{90,000}{100}}$$

$$= 41.23$$

$$P[(\bar{x}_1 - \bar{x}_2) \ge 600] = P\left[z \ge \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]$$

$$= P\left[z \ge \frac{600 - 500}{41.23}\right]$$

$$= P(z \ge 2.43)$$

$$= 0.5000 - 0.4925 = 0.0075$$

8.4 Given $\mu_{\bar{x}} = \mu = 1000, \sigma = 125$ and $n = 100$. Thus $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 125/\sqrt{100} = 12.5$

$$P(970 \le \bar{x} \le 1030)$$

$$= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} \le z \le \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}}\right]$$

$$= P\left[\frac{970 - 1000}{12.5} \le z \le \frac{1030 - 1000}{12.5}\right]$$

$$= P(-2.4 \le z \le 2.4)$$

$$= P(z \le 2.4) + P(z \ge -2.4)$$

$$= 0.4918 + 0.4918 = 0.9836$$

8.5 Given $\mu_{\bar{x}} = \mu = 5, \sigma = 0.005$ and $n = 9$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.005/\sqrt{9} = 0.0017$$

$$P(\bar{x} \ge 5.004) = P\left[z \ge \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right]$$

$$= P\left[z \ge \frac{5.004 - 5.000}{0.0017}\right]$$

$$= P(z \ge 2.4) = 1 - P(z \le 2.4)$$

$$= 1 - 0.9918 = 0.0082$$

8.6 Given $\mu_{\bar{x}} = \mu = 50, \sigma = 100$ and $n = 81$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 100/\sqrt{81} = 11.1$$

$$P(\bar{x} \le 40) = P\left[z \ge \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \le \frac{40 - 50}{11.1}\right]$$

$$= P(z \le -0.90) = 0.5000 - 0.3159 = 0.1841$$

8.7 Since standard error is inversely proportional to the square root of the sample size, therefore to reduce the standard error determined by the market research team, the sample size should be increased to $n = 400$ (four times of $n = 100$).

8.8 The number of possible samples of size 25 each from a group of 3000 soldiers with and without replacement are $(3000)^{25}$ and ${}^{300}C_{25}$, respectively. These numbers are much larger than 80—actually drawn samples. Thus we will get only an experimental sampling distribution of means rather than true sampling distribution. Hence mean and standard deviation would be close to those of the theoretical distribution. That is:

(a) $\mu_{\bar{x}} = \mu = 68$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/\sqrt{25} = 0.60$

(b) $\mu_{\bar{x}} = \mu = 68$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

$$= \frac{3}{\sqrt{25}}\sqrt{\frac{3000 - 25}{3000 - 1}} = 1.19$$

8.9 Given $\mu_{\bar{x}} = \mu = 16.9, \sigma = 3.6$ and $n = 36$. Thus

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3.6/\sqrt{36} = 0.60$$

$$P(\bar{x} \ge 16.3) = P\left[z \ge \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \ge \frac{16.3 - 16.9}{0.60}\right]$$

$$= P[z \ge -1] = 0.5000 + 0.1587 = 0.6587$$

8.10 Given $\mu = 3,40,000; \sigma = 20,000, n_1 = 50, n_2 = 100$, and $n_3 = 200$

For $n_1 = 50$:

$$z_1 = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{3,42,000 - 3,40,000}{20,000/\sqrt{50}} = 0.88$$

$$z_2 = \frac{3,37,000 - 3,40,000}{20,000/\sqrt{50}} = -0.88$$

$$P(-0.88 \le z \le 0.88) = 0.3106 \times 2 = 0.6212$$

Similar calculations for $n_2 = 100$ and $n_2 = 200$ give

$$P(-1.25 \le z \le 1.25) = 0.3944 \times 2 = 0.7888$$

$$P(-1.76 \le z \le 1.76) = 0.4616 \times 2 = 0.9282$$

8.9 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

There are many situations in which each element of the population can be classified into two mutually exclusive categories such as success or failure, accept or reject, head or tail of a coin, and so on. These and similar situations provide practical examples of binomial experiments, if the sampling procedure has been conducted in an appropriate manner. If a random sample of n elements is selected from the binomial population and x of these possess the specified characteristic, then the sample proportion $\bar{p}$ is the best statistic to use for statistical inferences about the population proportion parameter p. The sample proportion can be defined as:

$$\bar{p} = \frac{\text{Elements of sample having characteristic, } x}{\text{Sample size, } n}$$

With the same logic of sampling distribution of mean, the sampling distribution of sample proportions with mean $\mu_{\bar{p}}$ and standard deviation (also called *standard error*) $\sigma_{\bar{p}}$ is given by

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

If the sample size n is large ($n \geq 30$), the sampling distribution of $\bar{p}$ can be approximated by a normal distribution. The approximation will be adequate if

$$np \geq 5 \text{ and } n(1-p) \geq 5$$

It may be noted that the sampling distribution of the proportion would actually follow binomial distribution because population is binomially distributed.

The mean and standard deviation (error) of the sampling distribution of proportion are valid for a finite population in which sampling is with replacement. However, for finite population in which sampling is done without replacement, we have

$$\mu_{\bar{p}} = p \quad \text{and} \quad \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

Under the same guidelines as mentioned in previous sections, for a large sample size n (≥ 30), the sampling distribution of proportion is closely approximated by a normal distribution with mean and standard deviation as stated above. Hence, to standardize sample proportion $\bar{p}$, the standard normal variable.

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately the standard normal distribution.

8.9.1 Sampling Distribution of the Difference of Two Proportions

Suppose two populations of size N_1 and N_2 are given. For each sample of size n_1 from first population, compute sample proportion $\bar{p}_1$ and standard deviation $\sigma_{\bar{p}_1}$. Similarly, for each sample of size n_2 from second population, compute sample proportion $\bar{p}_2$ and standard deviation $\sigma_{\bar{p}_2}$.

For all combinations of these samples from these populations, we can obtain a sampling distribution of the difference $\bar{p}_1 - \bar{p}_2$ of samples proportions. Such a distribution is called *sampling distribution of difference of two proportions*. The mean and standard deviation of this distribution are given by

$$\mu_{\bar{p}_1 - \bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

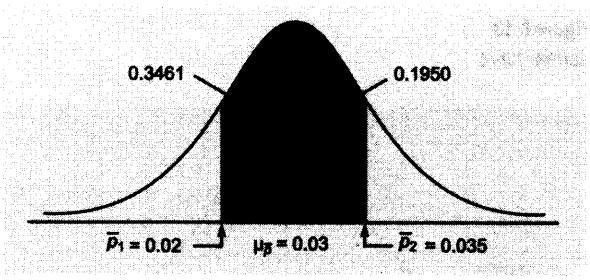
and

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\sigma^2_{\bar{p}_1} + \sigma^2_{\bar{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If sample size n_1 and n_2 are large, that is, $n_1 \geq 30$ and $n_2 \geq 30$, then the sampling distribution of difference of proportions is closely approximated by a normal distribution.

Example 8.10: A manufacturer of watches has determined from experience that 3 per cent of the watches he produces are defective. If a random sample of 300 watches is examined, what is the probability that the proportion defective is between 0.02 and 0.035? *[Delhi Univ., MBA, 1990]*

Figure 8.12 Normal curve



Solution: We are given the following information

$$\mu_{\bar{p}} = p = 0.03,\ \bar{p}_1 = 0.02,\ \bar{p}_2 = 0.035 \text{ and } n = 300$$

Thus standard error of proportion is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.03 \times 0.97}{300}}$$

$$= \sqrt{0.000097} = 0.0098$$

For calculating the desired probability, we apply the following formula

$$P[0.02 \leq \bar{p} \leq 0.035] = P\left[\frac{\bar{p}_1 - p}{\sigma_{\bar{p}}} \leq z \leq \frac{\bar{p}_2 - p}{\sigma_{\bar{p}}}\right]$$

$$= P\left[\frac{0.02 - 0.03}{0.0098} \leq z \leq \frac{0.035 - 0.03}{0.0098}\right]$$

$$= P[-1.02 \leq z \leq 0.51]$$

$$= P(z \geq -1.02) + P(z \leq 0.51) = 0.3461 + 0.1950 = 0.5411$$

Hence the probability that the proportion of defectives will lie between 0.02 and 0.035 is 0.5411.

Example 8.11: Few years back, a policy was introduced to give loan to unemployed engineers to start their own business. Out of 1,00,000 unemployed engineers, 60,000 accept the policy and got the loan. A sample of 100 unemployed engineers is taken at the time of allotment of loan. What is the probability that sample proportion would have exceeded 50 per cent acceptance?

Solution: We are given the following information

$$\mu_{\bar{p}} = p = 0.60,\ N = 1,00,000 \text{ and } n = 100$$

Thus the standard error of proportion in a finite population of size 1,00,000 is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.60 \times 0.40}{100}}\sqrt{\frac{1,00,000-100}{1,00,000-1}}$$

$$= \sqrt{0.0024}\sqrt{0.9990} = 0.0489 \times 0.9995 = 0.0488$$

The probability that sample proportion would have exceeded 50 per cent acceptance is given by

$$P(x \geq 0.50) = P\left[z \geq \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] = P\left[z \geq \frac{0.50 - 0.60}{0.0489}\right]$$

$$= P[z \geq -2.04] = 0.5000 + 0.4793 = 0.9793$$

Example 8.12: Ten per cent of machines produced by company A are defective and five per cent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02?

[South Gujrat Univ, MBA; Delhi Univ., MBA, 1999]

Solution: We are given the following information

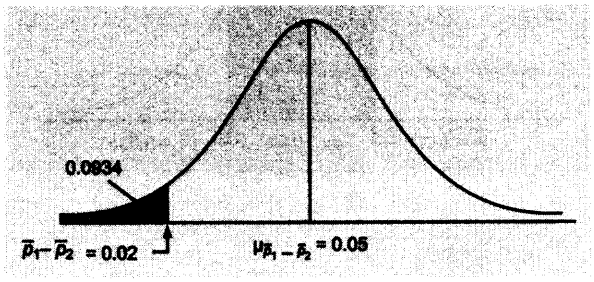
$$\mu_{\bar{p}_1-\bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2 = 0.10 - 0.05 = 0.05;\ n_1 = 250 \text{ and } n_2 = 300$$

Thus standard error of the difference in a simple proportions is given by

$$\mu_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.10\times0.90}{250} + \frac{0.05\times0.95}{300}}$$

$$= \sqrt{\frac{0.90}{250} + \frac{0.0475}{300}} = \sqrt{0.00052} = 0.0228$$

Figure 8.13
Normal curve



The desired probability of difference in sample proportions is given by

$$P[(\bar{p}_1 - \bar{p}_2) \le 0.02] = P\left[z \le \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}\right]$$

$$= P\left[z \le \frac{0.02 - 0.05}{0.0228}\right] = P[z \le -1.32]$$

$$= 0.5000 - 0.4066 = 0.0934$$

Hence the desired probability for the difference in sample proportions is 0.0934.

Self Practice Problems 8B

8.11 Assume that 2 per cent of the items produced in an assembly line operation are defective, but that the firm's production manager is not aware of this situation. What is the probability that in a lot of 400 such items, 3 per cent or more will be defective?

8.12 If a coin is tossed 20 times and the coin falls on head after any toss, it is a success. Suppose the probability of success is 0.5. What is the probability that the number of successes is less than or equal to 12?

8.13 The quality control department of a paints manufacturing company, at the time of despatch of decorative paints, discovered that 30 per cent of the containers are defective. If a random sample of 500 containers is drawn with replacement from the population, what is the probability that the sample proportion will be less than or equal to 25 per cent defective?

8.14 A manufacturer of screws has found that on an average 0.04 of the screws produced are defective. A random sample of 400 screws is examined for the proportion of defective screws. Find the probability that the proportion of defective screws in the sample is between 0.02 and 0.05.

8.15 A manager in the billing section of a mobile phone company checks on the proportion of customers who are paying their bills late. Company policy dictates that this proportion should not exceed 20 per cent. Suppose that the proportion of all invoices that were paid late is 20 per cent. In a random sample of 140 invoices, determine the probability that more than 28 per cent invoices were paid late.

Hints and Answers

8.11 $\mu_{\bar{p}} = np = 400 \times 0.02 = 8$;

$\sigma_p = \sqrt{npq} = \sqrt{400\times0.02\times0.98} = 2.8$

and 3% of 400 = 12 defective items. Thus

$$P(\bar{p} \ge 12) = P\left[z \ge \frac{\bar{p} - np}{\sigma_p}\right] = P\left[z \ge \frac{12-8}{2.8}\right]$$

$$= P(z \ge 1.42) = 0.5000 - 0.4222$$

$$= 0.0778$$

8.12 Given $\mu_{\bar{p}} = np = 20 \times 0.50 = 10$;

$$\sigma_{\bar{p}} = \sqrt{npq} = \sqrt{20\times0.50\times0.50} = 2.24$$

$$P(\bar{p} \le 12) = P\left[z \le \frac{\bar{p} - np}{\sigma_{\bar{p}}}\right] = P\left[z \le \frac{12-10}{2.24}\right]$$

$$= P(z \le 0.89) = 0.8133$$