8.13 Given $\mu_{\bar{p}} = p = 0.30, n = 500;$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30 \times 0.70}{500}} = 0.0205.$$

$$P(\bar{p} \le 0.25) = P\left[z \le \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right] = \left[z \le \frac{0.25 - 0.30}{0.0205}\right]$$

$$= P[z \le -2.43] = 0.5000 - 0.4927$$
$$= 0.0083$$

8.14 Given $\mu_{\bar{p}} = p = 0.04, n = 400;$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.04 \times 0.96}{400}} = 0.009$$

$$P[0.02 \le \bar{p} \le 0.05] = P\left[\frac{\bar{p}_1 - p}{\sigma_{\bar{p}}} \le z \le \frac{\bar{p}_2 - p}{\sigma_{\bar{p}}}\right]$$

$$= P\left[\frac{0.02 - 0.04}{0.009} \le z \le \frac{0.05 - 0.04}{0.009}\right]$$
$$= P[-2.22 \le z \le 2.22]$$
$$= P[z \ge -2.22] + P[z \le 2.22]$$
$$= 0.4861 + 0.4861 = 0.9722$$

8.15 Given $\mu_{\bar{p}} = p = 0.20, n = 140;$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20 \times 0.80}{140}} = 0.033$$

$$P[\bar{p} \ge 0.28] = P\left[z \ge \frac{\bar{p} - p}{\sigma_{\bar{p}}}\right]$$

$$= P\left[z \ge \frac{0.28 - 0.20}{0.033}\right] = P[z \ge 2.42]$$
$$= 0.5000 - 0.4918 = 0.0082$$

Formulae Used

1. **Standard deviation (or standard error) of sampling distribution of mean, $\bar{x}$**

- **Infinite Population:** $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
- **Finite Population:** $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

where $n < 0.5N$; n, N = size of sample and population, respectively.

2. **Estimate of $\sigma_{\bar{x}}$ when population standard deviation is not known**

- **Infinite Population:** $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$
- **Finite Population:** $\sigma_{\bar{x}} = \dfrac{s}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$

3. **Standard deviation of sampling distribution of sample means**

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

4. **Standard deviation (or standard error) of sampling distribution of proportion**

- **Infinite Population:** $\sigma_{\bar{p}} = \sqrt{\dfrac{p(1-p)}{n}}$; $q = 1 - p$
- **Finite Population:** $\sigma_{\bar{p}} = \sqrt{\dfrac{p(1-p)}{n}}\sqrt{\dfrac{N-n}{N-1}}$

5. **Standard deviation of sampling distribution of sample proportions**

$$\sigma_{p_1 - p_2} = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}};$$

$$q_1 = 1 - p_1; \quad q_2 = 1 - p_2$$

Chapter Concepts Quiz

True or False

1. The sampling distribution provides the basis for statistical inference when sample results are analysed. (T/F)
2. The sampling distribution of mean is the probability density function that describes the distribution of the possible values of a sample mean. (T/F)
3. The expected value (mean) is equal to the population mean from which the sample is chosen. (T/F)
4. As the sample size is increased, the sampling distribution of the mean approaches the normal distribution regardless of the population distribution. (T/F)
5. A sample size of $n \ge 30$ is considered large enough to apply the central limit theorem. (T/F)
6. Standard error of the mean is the standard deviation of the sampling distribution of the mean. (T/F)
7. The finite correction factor may be omitted of $n < 0.5$ N. (T/F)
8. The principles of the 'inertia of large number' and

'statistical regularity' govern random sampling. (T/F)

9. Every member of the population is tested in a sample survey. (T/F)

10. Simple random sampling is non-probability sampling method. (T/F)

11. Expected value (mean) of samples drawn randomly from a population are always same. (T/F)

12. The standard error becomes stable with an increase in sample size. (T/F)

13. The principle of 'inertia of large number' is a corollary of the principle of 'statistical regularity'. (T/F)

14. Cluster sampling is a non-random sampling method. (T/F)

15. Quota sampling method is used when the population is widely scattered. (T/F)

Multiple Choice

16. Which of the following is the principle on which theory of sampling is based?
(a) Statistical regularity (b) Inertia of large numbers
(c) Both (a) and (b) (d) None of the above

17. Which of the following is the non-random method of selecting samples from a population?
(a) Multistage sampling
(b) Cluster sampling
(c) Quota sampling (d) All of the above

18. The sampling distribution of proportion is approximated by a normal distribution when
(a) $np \geq 5$ (b) $pq \geq 5$
(c) both (a) and (b) (d) none of the above

19. The finite correction factor is applied when the sampling proportion is
(a) greater then 0.05 (b) greater than 0.50
(c) less than 0.50 (d) none of these

20. The sampling distribution of mean is normal provided
(a) population distribution is normal
(b) population distribution is not normal but the sample size is large ($n \geq 30$)
(c) population distribution is approximately the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$
(d) all of the above

21. To reduce the standard error of proportion by 50%, the sample size in increased by a factor of
(a) 2 (b) 4
(c) 8 (d) none of the above

22. The precision of proportion as an estimator of population proportion is worse when it is
(a) 0.25 (b) 0.50
(c) 0.75 (d) none of the above

23. The central limit theorem assures that the sampling distribution of the mean
(a) is always normal
(b) is always normal for large sample sizes
(c) approches normal distribution as sample size increases
(d) none of the above

24. For a normally distributed population, the sampling distribution of the mean
(a) is also normally distributed
(b) has a mean equal to the population mean
(c) both (a) and (b)
(d) none of the above

25. A significant difference between the statistic and parametric value implies that
(a) statistic values used to approximate parameter
(b) sample statistic is representative of the population
(c) the difference is real
(d) none of the above

26. If the population is finite, then standard error of mean is given by
(a) $\dfrac{\sigma}{\sqrt{n}}$ (b) $\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$
(c) $\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-1}{N-n}}$ (d) $\dfrac{\sigma}{\sqrt{n}}\sqrt{1-\dfrac{n}{N}}$

27. For a sampling distribution of mean the middle 95 per cent observations under the curve are covered in the range
(a) $\bar{x} \pm 1.64\sigma_{\bar{x}}$ (b) $\bar{x} \pm 1.96\sigma_{\bar{x}}$
(c) $\bar{x} \pm 2.34\sigma_{\bar{x}}$ (d) none of the above

28. For a sampling distribution of proportion, the most probable limits for p are
(a) $\bar{p} \pm 3\sigma_{\bar{p}}$ (b) $p \pm 3\sigma_p$
(c) $\bar{x} \pm 3\sigma_{\bar{x}}$ (d) none of the above

29. If the population proportion is 0.5 and standard error of sample proportion is 0.01, then the required sample size is
(a) 250 (b) 2500
(c) 3000 (d) 3500

30. The sampling distribution is approximately the t-distribution with degrees of freedom
(a) n (b) $n - 1$
(c) $n - 2$ (d) $2n - 1$

31. The application of Centeral Limit Theorem ensures that the sampling distribution of the mean
(a) is normal
(b) is normal for large sample size
(c) approximate normal as sample size increases
(d) all of these

32. Use of Central Limit Theorem relates to
(a) the shape of the sampling distribution
(b) the sample statistics to estimate population parameters
(c) large sample size more than 30 observations
(d) all of these

33. There is no need to use finite population multiplier when
(a) $p \geq 0.05$ (b) $pq \leq 0.05$
(c) $p + q = 1$ (d) none of these

34. Sampling distribution of mean and standard deviation is very close to the standard normal distribution, provided

(a) population distribution is normal
(b) population distribution is not normal but sample size is large
(c) both (a) and (b)
(d) either (a) or (b)

35. Which of the following is the probability method of selecting samples from a population?
(a) Quota sampling (b) Purposive sampling
(c) Judgment sampling (d) none of these

Concepts Quiz Answers

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

Review Self-Practice Problems

8.16 An auditor takes a random sample of size $n = 36$ from a population of 1000 accounts receivable. The mean value of the accounts receivable for the population is Rs 260 with the population standard deviation Rs 45. What is the probability that the sample mean will be less than Rs 250?

8.17 A marketing research analyst selects a random sample of 100 customers out of the 400 who purchased a particular item from central store. The 100 customers spent an average of Rs 250 with a standard deviation of Rs 70. For a middle 95 per cent customers, determine the mean purchase amount for all 400 customers.

8.18 In a particular coal mine, 5000 employees on an average are of 58 years of age with a standard deviation of 8 years. If a random sample of 50 employees is taken, what is the probability that the sample will have an average age of less than 60 years?

8.19 A simple random sample of 50 ball bearings taken from a large number being manufactured has a mean weight of 1.5 kg per bearing with a standard deviation of 0.1 kg.

(a) Estimate the value of the standard error of the mean
(b) If the sample of 50 ball bearings is taken from a particular production run that includes just 150 bearings as the total population, then estimate the standard error of the mean and compare it with the result of part (a).

8.20 A population proportion is 0.40. A simple random sample of size 200 will be taken and the sample proportion will be used to estimate the population proportion, what is the probability that the sample proportion will be within ±0.03 of the population proportion?

8.21 A sales manager of a firm believes that 30 per cent of the firm's orders come from first time customers. A simple random sample of 100 orders will be used to estimate the proportion of first-time customers. Assume that the sales manager is correct and proportion is 0.30.

(a) Justify sampling distribution of proportion for this case
(b) What is probability that the sample proportion will be between 0.20 and 0.40?

8.22 The diameter of a steel pipe manufactured at a large factory is expected to be approximately normally distributed with a mean of 1.30 inches and a standard deviation of 0.04 inch.

(a) If a random sample of 16 pipes is selected, then what is the probability that randomly selected pipe will have a diameter between 1.28 and 1.30 inches?
(b) Between what two values will 60 per cent of the pipes fall in terms of the diameter?

Hints and Answers

8.16 Given $\mu_{\bar{x}} = \mu = 260; n = 36;$

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 45/\sqrt{36} = 7.5$$

$$\begin{aligned} P[\bar{x} \le 250] &= P\left[z \le \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}\right] = P\left[z \le \frac{250-260}{7.5}\right] \\ &= P[z \le -1.33] \\ &= 0.5000 - 0.4082 = 0.0918 \end{aligned}$$

8.17 Given $s = 70$, $n = 100$, $\bar{x} = 250$, $z = 1.96$ at 95% confidence. Thus

$$\begin{aligned} s_{\bar{x}} &= \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{70}{\sqrt{100}}\sqrt{\frac{400-100}{400-1}} \\ &= 7(0.867) = 11.33 \\ \bar{x} \pm z\, s_{\bar{x}} &= 250 \pm 1.96(11.33) \\ &= \text{Rs } 227.80 \text{ to Rs } 272.20 \end{aligned}$$

8.18 Given $n = 50$, N = 5000, $\mu = 58$, and $\sigma = 8$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{50}}\sqrt{\frac{5000-50}{5000-1}}$$

$$= 1.131 \times 0.995 = 1.125$$

$$P(\bar{x} \le 60) = P\left[z \le \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}\right] = P\left[z \le \frac{58-60}{1.125}\right]$$

$$= P[z \le -1.77]$$

$$= 0.5000 - 0.4616 = 0.0384$$

8.19 Given $\mu_{\bar{x}} = \mu = 1.5, n = 50$, N = 150 and $s = 0.1$.

(a) $s_{\bar{x}} = s/\sqrt{n} = 0.1/\sqrt{50} = 0.014$ kg

(b) $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}} = 0.014\sqrt{\dfrac{150-50}{150-1}} = 0.011$ kg.

It is less than the value in part (a) due to finite correction factor.

8.20 Given $\mu_{\bar{p}} = \bar{p} = 0.40, n = 200$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40 \times 0.60}{200}} = 0.0346$$

$$P(-0.03 \le \bar{p} \le 0.03) = 2P\left[z \le \frac{\bar{p}-p}{\sigma_{\bar{p}}}\right]$$

$$= 2P\left[z \le \frac{0.03}{0.0346}\right]$$

$$= 2P(z \le 0.87)$$

$$= 2 \times 0.3078 = 0.6156$$

8.21 Given $p = 0.30, n = 100$. Thus

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30 \times 0.70}{100}} = 0.0458$$

(a) Since both $np = 100\,(0.30) = 30$ and $nq = n(1-p) = 100\,(0.70) = 70$ are greater than 5, the normal distribution is appropriate to use

(b) $P(0.20 \le \bar{p} \le 0.40)$

$$= P\left[\frac{0.20-0.30}{0.0458} \le z \le \frac{0.40-0.30}{0.0458}\right]$$

$$= P[-2.18 \le z \le 2.18]$$

$$= 2P(z \le 2.18) = 2 \times 0.4854 = 0.9708$$

8.22 Given $\mu_{\bar{x}} = \mu = 1.30, \sigma = 0.04$ and $n = 16$,

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.04/\sqrt{16} = 0.01$$

(a) $P(1.28 \le \bar{x} \le 1.30) = P\left[\dfrac{\bar{x}_1-\mu}{\sigma_{\bar{x}}} \le z \le \dfrac{\bar{x}_2-\mu}{\sigma_{\bar{x}}}\right]$

$$= P\left[\frac{1.28-1.30}{0.01} \le z \le \frac{1.30-1.30}{0.01}\right]$$

$$= P[2 \le z \le 0] = 0.5000 - 0.4772$$

$$= 0.0228$$

(b) $\bar{x} \pm z\sigma_{\bar{x}} = 1.30 \pm 0.84\,(0.01) = 1.30 \pm 0.0084$

$= 1.2916$ to 1.3084

Chapter 9

Lest men suspect your tale untrue, keep probability in view.

—John Gay

When in doubt use a bigger hammer.

—Thomas L. Martin Jr.

Estimation and Confidence Intervals

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand the concept of a confidence interval.
- compute the margin of error associated with a sample mean and a proportion.
- compute and interpret confidence interval estimates to know the precision of the estimate of a population mean and proportion.
- identify factors that affect the precision of confidence intervals.
- determine the sample sizes for different levels of precision.

9.1 INTRODUCTION

In Chapter 8, we have learnt how to use a sample data to draw statistical inference about the unknown value of a population or process parameter of interest. On many occasions we do not have enough information to calculate an exact value of a population parameter (such as μ, σ, and p) and therefore make the best estimate of this value from the corresponding sample statistic (such as $\bar{x}$, s, and $\bar{p}$). The need to use the sample statistic to draw conclusions about the population characteristic is one of the fundamental applications of statistical inference in business and economics. A few applications are given below:

- A production manager needs to determine the proportion of items being manufactured that do not match with quality standards.
- A telephone service company may be interested to know the average length of a long distance telephone call and its standard deviation.
- A company needs to understand consumer awareness of its product.
- Any service centre needs to determine the average amount of time a customer spends in queue.

In all such cases, a decision-maker needs to examine the following two concepts that are useful for drawing statistical inference about an unknown value of population or process parameters based upon random samples:

(i) ***Estimation***—a sample statistic to estimate most likely value of its population parameter

Estimation The method to estimate the value of a population parameter from the value of the corresponding sample statistic.

(ii) ***Hypothesis testing***—a claim or belief about an unknown parameter value

In this chapter we shall discuss methods to estimate unknown value of a population parameter and then to determine the range of values (confidence interval) likely to contain the parameter value. There are two types of estimates for the value of population parameter:

Point estimate The value of sample statistic that estimate population parameter.

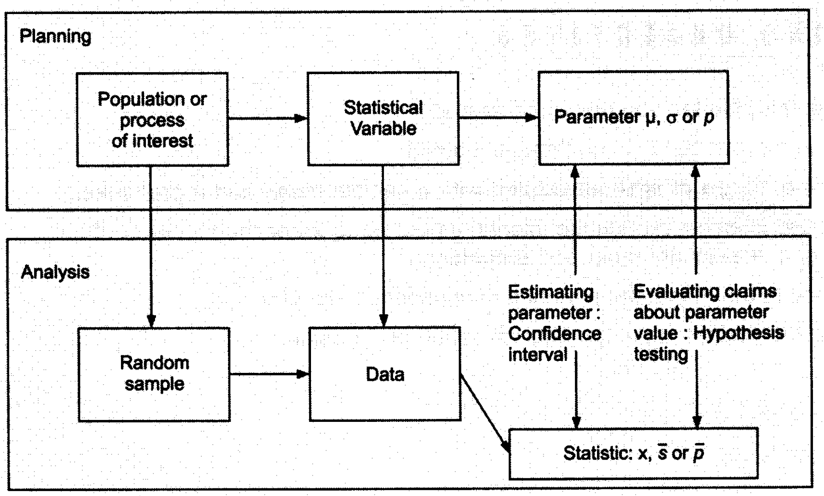
(i) ***Point estimate*** It is the value of sample statictic that is used to estimate most likely value of the unknown population parameter

(ii) ***Confidence interval estimate*** It is the range of values that is likely to have population parameter value with a specified level of confidence

For estimating a parameter value, it is important to know, (i) a point estimate, (ii) the amount of possible error in the point estimate, that is, an interval likely to contain the parameter value, and (iii) the statement or degree of confidence that the interval contains the parameter value. The knowledge of such information is called *confidence interval* or *interval estimation.*

The use of confidence interval and test of hypothesis to draw inference about population parameters μ ,σ and p with the help of sample statistics $\bar{x}$, s and $\bar{p}$ respectively is illustrated in Fig 9.1

Figure 9.1
Process to Draw Inference about Population Parameters


Planning
Population or process of interest
Statistical Variable
Parameter μ, σ or p
Analysis
Estimating parameter : Confidence interval
Evaluating claims about parameter value : Hypothesis testing
Random sample
Data
Statistic: x, s̄ or p̄

9.2 POINT ESTIMATION

A sample statistic (such as $\bar{x}$, s, or $\bar{p}$) that is calculated using sample data to estimate most likely value of the corresponding unknown population parameter (such as μ, σ or p) is termed as *point estimator,* and the numerical value of the estimator is termed as *point estimate.* For example, if we calculate that 10 per cent of the items in a random sample taken from a day's production are defective, then the result '10 per cent' is a point estimate of the percentage of items in the whole lot that are defective. Thus, until the next sample of items is not drawn and examined, we may proceed on manufacturing with the assumption that any day's production contains 10 per cent defective items.

9.2.1 Properties of a Point Estimator

For a statistical point estimate, the sampling distribution of the estimator provides information about the best estimator. Before any statistical inference is drawn, it is essential to resolve following two important issues:

(i) Selection of an appropriate statistic to serve as the best estimator of a population parameter.

(ii) The nature of the sampling distribution of this selected statistic. Since the sample statistic value varies from sample to sample, the accuracy of a given estimator also varies from sample to sample. This means that there is no certainty of the accuracy achieved for the sample one happens to draw. Although in practice only one sample is selected at any given time, we should judge the accuracy of an estimator based on its *average value* over all possible samples of equal size. Hence, we prefer to choose the estimator whose 'average accuracy' is close to the value of population parameter being estimated. The criteria for selecting an estimator are:

- Unbiasedness
- Consistency
- Efficiency

As different sample statistics can be used as point estimators of different population parameters, the following general notations will be used in this section:

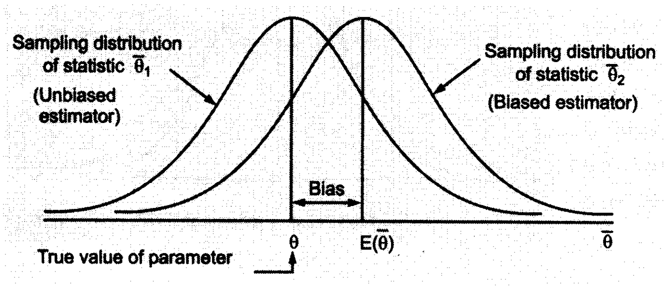
θ = population parameter (such as μ, σ, p) of interest being estimated

$\bar{\theta}$ = sample statistic (such as $\bar{x}$, s, $\bar{p}$) or point estimator of θ

Here, θ (theta) is the Greek letter and $\bar{\theta}$ is read as 'theta hat'.

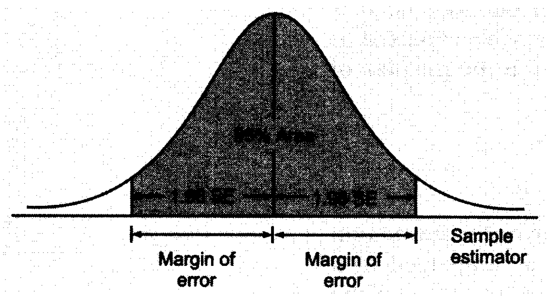
Unbiasedness The value of a statistic measured from a given sample is likely to be above or below the actual value of population parameter of interest due to sampling error. Thus it is desirable that the mean of the sampling distribution of sample means taken from a population is equal to the population mean. If it is true, then the sample statistic is said to be an *unbiased estimator* of the population parameter. Hence, the *sample statistic* $\bar{\theta}$ *is said to be an unbiased estimator of the population parameter, provided* $E(\bar{\theta}) = \theta$ where $E(\bar{\theta})$ = expected value or mean of the sample statistic $\bar{\theta}$.

Figure 9.2
Sampling Distributions for Biased and Unbiased Estimators



In a sampling distribution of sample mean and sample proportion, we have $E(\bar{x}) = \mu$ and $E(\bar{p}) = p$ respectively, therefore both $\bar{x}$ and $\bar{p}$ are unbiased estimators of the corresponding population parameters μ and p. In Chapter 8, we used $n - 1$ rather than n in the denominator, regardless of the sample size, to calculate the sample variance s^2. This was done to make the sample variance an unbiased estimator of the population variance σ^2, that is $E(s^2) = \sigma^2$. However, sample standard deviation s is not an unbiased estimator of σ. But this bias reduces as sample size increases.

Figure 9.3
Sampling Distribution of an Unbiased Estimator



If $E(\bar{\theta}) \neq \theta$ for the sampling distribution of $\bar{\theta}$, then $\bar{\theta}$ is said to be a *biased* estimator. Figure 9.2 illustrates this point by showing sampling distributions of two possible sample statistics, $\bar{\theta}_1$ and $\bar{\theta}_2$, for estimating a population parameter θ. Even if both $\bar{\theta}_1$ and $\bar{\theta}_2$ have the same standard error, the statistic $\bar{\theta}_1$ yields an estimate closer to the parameter θ than $\bar{\theta}_2$, because the parameter θ is located at the mean of the sampling distribution, where $E(\bar{\theta}_1) = \theta$. The distance between the estimate and true value of parameter is called the *error of estimation*. The amount of the bias is shown in Fig. 9.2.

For any point estimator with a normal distribution, it has been proved that approximately 95 per cent of all point estimates will lie within 2(or more-exactly 1.96) standard deviations of the mean of that distribution. This implies that for the unbiased estimators, the difference between the point estimator and the true value of the parameter will be less than 1.96 standard deviations (or standard error). This quantity is called the **margin of error** and which provides a uppor bound for the error of estimation as shown in Fig. 9.3. In other words,

Margin of error The value added or subtracted from a point estimate in order to develop an interval estimate of a population parameter.

$$\text{Margin of error} = 1.96 \times \text{Standard error (SE) of estimator} = 1.96 \frac{\sigma}{\sqrt{n}}$$

If σ is unknown and sample size $n \geq 30$, or large, the sample standard deviation s can be used to approximate σ.

Consistency A point estimator is said to be consistent if its value $\bar{\theta}$ tends to become closer to the population parameter θ as the sample size increases. For example, the standard error of sampling distribution of the mean, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, tends to become smaller as sample size n increases. Thus the sample mean $\bar{x}$ is a consistent estimator of the population mean μ.

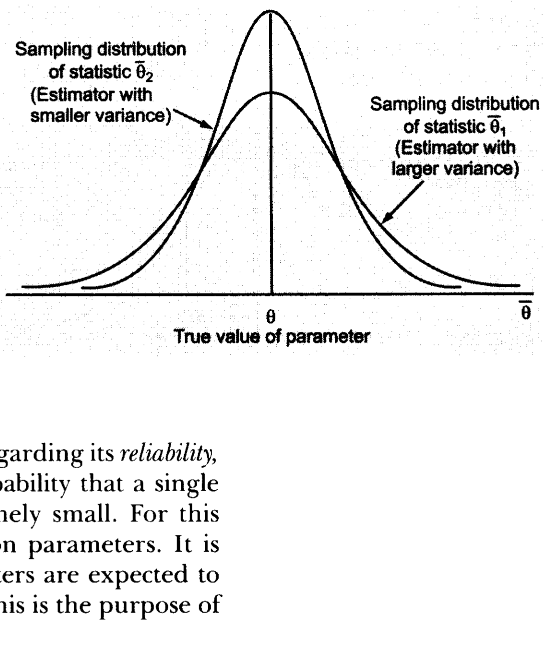
Similarly, the sample proportion $\bar{p}$ is a consistent estimator of the population proportion p because $\sigma_{\bar{p}} = \sigma/\sqrt{n}$.

Efficiency For the same population, out of two unbiased point estimators, the desirable characteristic of an unbiased estimator is that the spread (as measured by the variance) of the sumpling distribution should be as small as possible) Such unbiased estimator is said to be efficient because an individual estimate will fall close to the true value of population parameter with high probability. It is because of this reason that there is less variation in the sampling distribution of the statistic. For example, for a simple random sample of size n, if $\bar{\theta}_1$ and $\bar{\theta}_2$ are two unbiased point estimators of the population parameter θ, then relative efficiency of $\bar{\theta}_2$ to $\bar{\theta}_1$ is given by

$$\text{Relative efficiency} = \frac{\sigma(\bar{\theta}_1)}{\sigma(\bar{\theta}_2)}$$

Figure 9.4 shows the sampling distributions of two unbiased estimators $\bar{\theta}_1$ and $\bar{\theta}_2$ which are being considered for estimation of the population parameter θ. Since standard deviation (or error) of statistic $\bar{\theta}_2$ is less than that of $\bar{\theta}_1$, therefore value of $\bar{\theta}_2$ is more likely to provide an estimate that tend to lie closer to the true value of the parameter θ for a given sample. The statistic $\bar{\theta}_1$ tends to produce a larger estimation error both above and below the parameter θ. Thus, the estimates obtained from statistic $\bar{\theta}_2$ are more consistently close to θ than those of $\bar{\theta}_1$.

Figure 9.4
Sampling Distributions of Two Unbiased Statistics of Population Parameter



9.2.2 Drawback of Point Estimates

The draw back of a point estimate is that no information is available regarding its *reliability*, i.e. how close it is to its true population parameter. In fact, the probability that a single sample statistic actually equals the population parameter is extremely small. For this reason, point estimates are rarely used alone to estimate population parameters. It is better to offer a *range of values* within which the population parameters are expected to fall so that reliability (probability) of the estimate can be measured. This is the purpose of **interval estimation**

9.3 CONFIDENCE INTERVAL ESTIMATION

As pointed out earlier, in most of the cases, a point estimate does not provide information about 'how close is the estimate' to the population parameter unless accompanied by a statement of possible sampling error involved based on the sampling distribution of the statistic. It is therefore important to know the precision of an estimate before depending on it to make a decision. Thus, decision-makers prefer to use an *interval estimate* (i.e. the range of values defined around a sample statistic) that is likely to contain the population parameter value. An interval estimation is a rule for calculating two numerical values, say a and b that create an interval that contains the population parameter of interest. This interval is therefore commonly referred to as a *confidence coefficient* and denoted by $(1 - \alpha)$. However, it is also important to state 'how confident' one should be that the interval estimate contains the parameter value. Hence an interval estimate of a population parameter is a *confidence interval* with a statement of confidence that the interval contains the parameter value. In other words, a confidence interval estimation is an interval of values computed from sample data that is likely to contain the true population parameter value.

Interval estimate The interval calculated from a sample expected to include the corresponding population parameter.

Confidence interval The interval within which the population parameter is expected to lie.

The **confidence interval** estimate of a population parameter is obtained by applying the formula:

$$\text{Point estimate} \pm \text{Margin of error}$$

where Margin of error = z_c × Standard error of a particular statistic

z_c = critical value of standard normal variable that represents confidence level (probability of being correct) such as 0.90, 0.95, and so on.

9.4 INTERVAL ESTIMATION OF POPULATION MEAN (σ KNOWN)

Suppose the population mean μ is unknown and the true population standard deviation σ is known. Then for a large sample size $n(\geq 30)$, the sample mean $\bar{x}$ is the best point estimator for the population mean μ. Since sampling distribution is approximately normal, it can be used to compute confidence interval of population mean μ as follows:

$$\bar{x} \pm z_{\alpha/2}\, \sigma_{\bar{x}} \quad \text{or} \quad \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z-value representing an area $\alpha/2$ in the right tail of the standard normal probability distribution, and $(1 - \alpha)$ is the *level of confidence* as shown in Fig. 9.5.

Confidence limits The boundaries (both upper and lower) of a confidence interval.

Alternative Approch A $(1 - \alpha)$ 100% large sample confidence interval for a population mean μ can also be find by using the statistic

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

which has a standard normal distribution.

If $z_{\alpha/2}$ is the z-value with an area $\alpha/2$ in the right tail of normal curve, then we can write

$$P\left[-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha$$

or

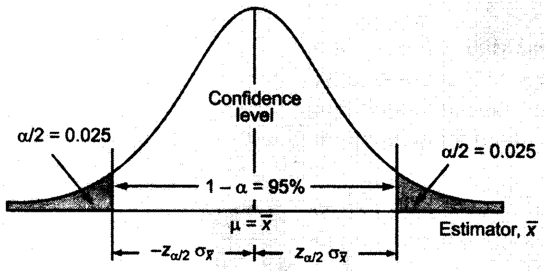
$$-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$-\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

so that

$$P\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Figure 9.5
Sampling Distribution of Mean x



Both lower limit $\bar{x} - z_{\alpha/2}(\sigma / \sqrt{n})$ and upper limit $\bar{x} + z_{\alpha/2}(\sigma / \sqrt{n})$ depend on the sample mean $\bar{x}$. Thus, in repeated sampling the interval $\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$ will contain the population mean μ with probability $1 - \alpha$.

The value of z that has 'tail area' $\alpha/2$ to its right and left is called its critical value and is represented by $z_{\alpha/2}$ and $-z_{\alpha/2}$ respectively. The area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is the confidence coefficient $(1 - \alpha)$.

Confidence level The confidence associated with an interval estimate. It is expressed in terms of probability that the true population parameter is included in the confidence interval.

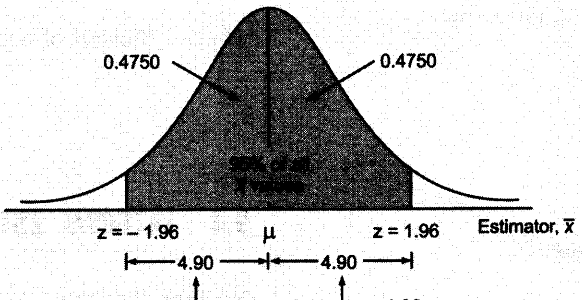
For example, if a 95 per cent level of confidence is desired to estimate the mean, then 95 per cent of the area under the normal curve would be divided equally, leaving an area equal to 47.5 per cent between each limit and population mean μ as shown in Fig. 9.6.

If $n = 100$ and $\sigma = 25$, then $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 25/\sqrt{100} = 2.5$. Using a table of areas for the standard normal probability distribution, 95 per cent of the values of a normally distributed population are within ± $1.96\sigma_{\bar{x}}$ or 1.96 (2.5) = ± 4.90 range. Hence 95 per cent of the sample means will be within ± 4.90 of the population mean μ. In other words, there is a 0.95 probability that the sample mean will provide a **sampling error** equal to $|\ \bar{x} - \mu\ |$ = 4.90 or less. The value 4.90 also provides an upper limit on the sampling error, also called *margin of error*. The value 0.95 is called *confidence coefficient* and the interval estimate $\bar{x}$ ± 4.90 is called a 95 per cent confidence interval.

In general, a 95 per cent confidence interval estimate implies that if all possible samples of the same size were drawn, then it would contain the true population mean μ in the interval $\bar{x} \pm z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$ and 5 per cent area under the curve would not contain value of μ. The values for $z_{\alpha/2}$ for the most commonly-used as well as the other confidence levels can be seen from standard normal probability table as shown in Table 9.1.

Figure 9.6

Sampling Distribution of Mean $\bar{x}$



Sampling error The difference between the value of an unbiased point estimator $\bar{x}$ or $\bar{p}$ and the value of the population parameter μ or p.

Table 9.1: Values of Standard Normal Probability

Confidence Level, (1 – α) (%)	*Acceptable Error Level, α*	*α/2*	$z_{\alpha/2}$
90%	0.10	0.050	±1.64
95%	0.05	0.025	±1.96
98%	0.02	0.010	±2.33
99%	0.01	0.005	±2.58

Example 9.1: The average monthly electricity consumption for a sample of 100 families is 1250 units. Assuming the standard deviation of electric consumption of all families is 150 units, construct a 95 per cent confidence interval estimate of the actual mean electric consumption.

Solution: The information given is: $\bar{x} = 1250$, $\sigma = 150$, $n = 100$ and confidence level $(1 - \alpha)$ = 95 per cent. Using the 'Standard Normal Curve' we find that the half of 0.95 yields a confidence coefficient $z_{\alpha/2}$ = 1.96. Thus confidence limits with $z_{\alpha/2}$ = ± 1.96 for 95 per cent confidence are given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1250 \pm 1.96 \frac{150}{\sqrt{100}} = 1250 \pm 29.40 \text{ units}$$

Thus for 95 per cent level of confidence, the population mean μ is likely to fall between 1220.60 units and 1274.40 units, that is, $1220.60 \leq \mu \leq 1274.40$.

Example 9.2: A survey conducted by a shopping mall group showed that a family in a metro-city spends an average of Rs 500 on cloths every month. Suppose a sample of 81 families resulted in a sample mean of Rs 540 per month and a sample standard deviation of Rs 150, develop a 95 per cent confidence interval estimator of the mean amount spent per month by a family

Solution: The information given is: $\bar{x}$ = 540, σ = 150, n = 81 and $z_{\alpha/2}$ = 1.96 for 95 per cent confidence level, Thus confidence limits with $z_{\alpha/2}$ = ± 1.96 are given by

$$\bar{x} \pm z_{\alpha/2}\, \sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 540 \pm 1.96 \frac{150}{\sqrt{81}}$$

$$= 540 \pm 40.67 \text{ or Rs } 499.33 \text{ and Rs } 580.67$$

Hence for 95 per cent level of confidence, the population mean μ is likely to fall between Rs 499.33 and Rs 580.67, i.e. $499.33 \leq \mu \leq 580.67$.

Example 9.3: The quality control manager at a factory manufacturing light bulbs is interested to estimate the average life of a large shipment of light bulbs. The standard deviation is known to be 100 hours. A random sample of 50 light bulbs gave a sample average life of 350 hours.

(a) Setup a 95 per cent confidence interval estimate of the true average life of light bulbs in the shipment.

(b) Does the population of light bulb life have to be normally distributed ? Explain.

Solution: The following information is given:

$$\bar{x} = 350, \sigma = 100, n = 50, \text{ and confidence level, } (1 - \alpha) = 95 \text{ per cent}$$

(a) Using the 'Standard Normal Curve', we have $z_{\alpha/2} = \pm\ 1.96$ for 95 per cent confidence level. Thus confidence limits are given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 350 \pm 1.96 \frac{100}{\sqrt{50}} = 350 \pm 27.72$$

Hence for 95 per cent level of confidence the population mean μ is likely to fall between 322.28 hours to 377.72 hours, that is, $322.28 \le \mu \le 377.72$.

(b) No, since σ is known and $n = 50$, from the central limit theorem we may assume that $\bar{x}$ is normally distributed.

9.4.1 Interval Estimation for Difference of Two Means

If all possible samples of large size n_1 and n_2 are drawn from two different populations, then sampling distribution of the difference between two means $\bar{x}_1$ and $\bar{x}_2$ is approximately normal with mean $(\mu_1 - \mu_2)$ and standard deviation

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For a desired confidence level, the confidence interval limits for the population mean $(\mu_1 - \mu_2)$ are given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\, \sigma_{\bar{x}_1 - \bar{x}_2}$$

Example 9.4: The strength of the wire produced by company A has a mean of 4500 kg and a standard deviation of 200 kg. Company B has a mean of 4000 kg and a standard deviation of 300 kg. A sample of 50 wires of company A and 100 wires of company B are selected at random for testing the strength. Find 99 per cent confidence limits on the difference in the average strength of the populations of wires produced by the two companies.

Solution: The following information is given:

Company A: $\bar{x}_1 = 4500, \quad \sigma_1 = 200, \quad n_1 = 50$

Company B: $\bar{x}_2 = 4000, \quad \sigma_2 = 300, \quad n_2 = 100$

Therefore $\bar{x}_1 - \bar{x}_2 = 4500 - 4000 = 500$ and $z_{\alpha/2} = 2.576$ and

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40{,}000}{50} + \frac{90{,}000}{100}} = 41.23$$

The required 99 per cent confidence interval limits are given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\, \sigma_{\bar{x}_1 - \bar{x}_2} = 500 \pm 2.576\ (41.23) = 500 \pm 106.20$$

Hence, the 99 per cent confidence limits on the difference in the average strength of wires produced by the two companies are likely to fall in the interval $393.80 \le \mu \le 606.20$.

9.5 INTERVAL ESTIMATION OF POPULATION MEAN (σ UNKNOWN)

If the standard deviation σ of a population is not known, then it can be approximated by the sample standard deviation, s when the sample size, n (≥ 30) is large. So, the interval estimator of a population mean μ for a large sample $n(\ge 30)$ with confidence coefficient $1 - \alpha$ is given by

$$\bar{x} \pm z_{\alpha/2}\, s_{\bar{x}} = \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

When the population standard deviation is not known and the sample size is small, the procedure of interval estimation of population mean is based on a probability distribution known as the *t-distribution*. This distribution is very similar to the normal distribution. However, the t-distribution has more area in the tails and less in the center than does normal distribution. The t-distribution depends on a parameter known as *degrees of freedom*. As the number of degrees of freedom increases, t-distribution gradually approaches the normal distribution, and the sample standard deviation s becomes a better estimate of population standard deviation σ.

The interval estimate of a population mean when the sample size is small ($n \leq 30$) with confidence coefficient $(1 - \alpha)$, is given by

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the critical value of t-test statistic providing an area $\alpha/2$ in the right tail of the t-distribution with $n - 1$ degrees of freedom, and

$$s = \sqrt{\frac{\Sigma\,(x_i - \bar{x})^2}{n - 1}}$$

The critical values of t for the given degrees of freedom can be obtained from the table of t-distribution (See appendix).

The procedure of the confidence interval estimation of population mean μ when population standard deviation is unknown and sample size is large or small, is summarized in Table 9.2.

Table 9.2: Confidence Interval for μ

Sample size	*Interval Estimate of Population Mean μ*
Large	
• σ is known	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
• σ estimated by s	$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$
Small	
• σ is known	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
• σ estimated by s	$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

Example 9.5: A random sample of 50 sales invoices was taken from a large population of sales invoices. The average value was found to be Rs 2000 with a standard deviation of Rs 540. Find a 90 per cent confidence interval for the true mean value of all the sales.

Solution: The information given is: $\bar{x}_1 = 2000$, $s = 540$, $n = 64$, and $\alpha = 10$ per cent. Therefore

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{540}{\sqrt{64}} = 67.50 \text{ and}$$

$$z_{\alpha/2} = 1.64 \text{ (from Normal table)}$$

The required confidence interval of population mean μ is given by

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 2000 \pm 1.64\ (67.50) = 2000 \pm 110.70$$

Thus the mean of the sales invoices for the whole population is likely to fall between Rs 1889.30 and Rs 2110.70, that is, $1889.30 \le \mu \le 2110.70$.

Example 9.6: The personnel department of an organization would like to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of 10 employees reveals the following family dental expenses (in thousand Rs) in the previous year: 11, 37, 25, 62, 51, 21, 18, 43, 32, 20.

Set up a 99 per cent confidence interval of the average family dental expenses for the employees of this organization.

Solution: The calculations for sample mean $\bar{x}$ and standard deviation are shown in Table 9.3.

Table 9.3: Calculation for x and s

Variable, x	$(x - \bar{x}) = (x - 32)$	$(x - \bar{x})^2$
11	−21	441
37	05	25
25	−07	49
62	30	900
51	19	361
21	−11	121
18	−14	196
43	11	121
32	00	00
20	−12	144
320	0	2358

From the data in Table 9.3, the sample mean $\bar{x} = \Sigma x/n = 320/10 =$ Rs 32, and the sample standard deviation $s = \sqrt{\Sigma (x - \bar{x})^2/n - 1} = \sqrt{2358/9} =$ Rs 5.11. Using this information and $t_{\alpha/2} = 1.833$ at $df = 9$, we have

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 32 \pm 1.833 \frac{5.11}{\sqrt{10}} = 32 \pm 2.962$$

Hence the mean expenses per family are likely to fall between Rs 29.038 and Rs 34.962, that is, $29.038 \le \mu \le 34.962$.

Self-Practice Problems 9A

9.1 In an effort to estimate the mean amount spend per customer for dinner at a city hotel, data were collected for a sample of 49 customers. Assume a population standard deviation of Rs 25.

(a) At 95 per cent confidence, what is the margin of error?

(b) If the sample mean is Rs 124, what is the 95 per cent confidence interval for the population mean?

9.2 The following data have been collected for a sample from a normal population: 5, 10, 8, 11, 12, 6, 15, 13,

(a) What is the point estimate of population mean and standard deviation?

(b) What is the confidence interval for population mean at 95 per cent confidence interval?

9.3 The quality control department of a wire manufacturing company periodically selects a sample of wire specimens in order to test for breaking strengths. Past experience has shown that the breaking strengths of a certain type of wire are normally distributed with standard deviation of 200 kg. A random sample of 64 specimens gave a mean of 6200 kg. Determine a 95 per cent confidence interval for the mean breaking strength of the population to suggest to the quality control supervisor.

9.4 A machine is producing ball bearings with a diameter of 0.5 inches. It is known that the standard deviation of the ball bearings is 0.005 inch. A sample of 100 ball bearings is selected and their average diameter is found to be 0.48 inch. Determine the 99 per cent confidence interval.

9.5 Suppose a wholesaler of paints wants to estimate the actual amount of paint contained in 10 kg cans purchased from a paint manufacturing company. It is known from the manufacturer's specifications that the standard deviation of the amount of paint is equal to 0.02 kg. A random sample of 50 cans is selected, and the average amount of paint per 10 kg can is 0.995 kg

(a) Setup a 99 per cent confidence interval estimate of the true population average amount of paint included in a 10 kg can.

(b) Based on your results, do you think that the wholesaler has a right to complain to the manufacture? Why?

9.6 Repeated tests on the determination of human blood composition during a laboratory analysis are known to be normally distributed. Ten tests on a given sample of blood yielded the values

1.002, 0.958, 1.014, 1.009, 1.041, 1.058, 1.024, 1.019, 1.020

Find a 99 per cent confidence interval for true composition of the blood in repeated tests of the sample.

9.7 The HRD department of a company developed an aptitude test for screening potential employees. The person who devised the test asserted that the mean mark attained would be 100. The following results were obtained with a random sample of applicants: $\bar{x} = 96$, $s = 5.2$, and $n = 13$. Calculate a 95 per cent confidence interval for the mean mark for all candidates and use it to see if the mean rank could be 100.

Hints and Answers

9.1 Given, $\bar{x} = 124$, $n = 49$, $\sigma = 250$, $z_{\alpha/2} = 1.96$ at 95 per cent confidence coefficient

(a) Margin of error $= z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = 1.96\dfrac{25}{\sqrt{49}} = 6.99$

(b) $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = 124 \pm 6.99$ or $117.01 \le \mu \le 130.99$

9.2 (a) $\bar{x} = \Sigma x/n = 80/8 = 10$

(b) $s = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{(n-1)}} = \sqrt{\dfrac{84}{(8-1)}} = 3.46$

(c) Since sample size $n = 8$ is small, degrees of freedom $n - 1 = 7$. At 7 degrees of freedom and 95 per cent confidence coefficient, $t = 2.365$, we have

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 10 \pm 2.365\frac{3.46}{\sqrt{8}}$$

$$= 10 \pm 2.90;\ 7.10 \le \mu \le 12.90$$

9.3 Given, $\bar{x} = 6200$, $\sigma = 200$, $n = 64$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence coefficient

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6200 \pm 1.96\frac{200}{\sqrt{64}}$$

$$= 6200 \pm 49;\ 6151 \le \mu \le 6249$$

9.4 Given $\bar{x} = 0.498$, $\sigma = 0.005$, $n = 100$, $z_{\alpha/2} = 2.58$ at 99 per cent confidence coefficient

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.498 \pm 2.58\frac{0.005}{\sqrt{100}}$$

$$= 0.498 \pm 0.0129;\ 0.4967 \le \mu \le 0.4993$$

9.5 Given $\bar{x} = 0.995$, $\sigma = 0.02$, $n = 50$ and $z_{\alpha/2} = 2.58$ at 99 per cent confidence coefficient

(a) $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = 0.995 \pm 2.58\dfrac{0.02}{\sqrt{100}}$

$$= 0.995 \pm 0.0072;\ 0.9878 \le \mu \le 1.0022$$

(b) Since the value 1.0 is included in the interval, there is no need to believe that the average is below 1.0.

9.6 $\bar{x} = \dfrac{\sum x}{n} = \dfrac{10.107}{10} = 1.010, s = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{(n-1)}} = 0.025;$ $n = 10$

$t_{\alpha/2} = 3.250$ at $df = 9$ and 99 per cent confidence coefficient

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 1.010 \pm 3.250\frac{0.025}{\sqrt{100}}$$

$$= 1.010 \pm 0.025;\quad 0.985 \le \mu \le 1.035$$

9.7 Given $\bar{x} = 96$, $s = 5.2$; $n = 13$; $z_{\alpha/2} = 1.96$ at 95 per cent confidence coefficient

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 96 \pm 1.96\frac{5.2}{\sqrt{13}}$$

$$= 96 \pm 2.83;\quad 93.17 \le \mu \le 98.83.$$

9.6 INTERVAL ESTIMATION FOR POPULATION PROPORTION

Recall that when $np \ge 5$ and $nq = n(1-p) \ge 5$, the central limit theorem for sample proportions give us the formula

$$z = \frac{\bar{p} - p}{\sqrt{pq/n}}$$

When the sample size is large, the sample proportion $\bar{p} = x/n$ is the best point estimator for the population proportion p. Since the sampling distribution of sample proportion

$\bar{p}$ is approximately normal with mean $\mu_{\bar{p}}$ and standard error $\sqrt{pq/n}$, the confidence interval for a population proportion at 1 – α confidence coefficient is given by

$$\bar{p} \pm z_{\alpha/2}\, \mu_{\bar{p}} = \bar{p} \pm z_{\alpha/2} \sqrt{pq/n}$$

or

$$\bar{p} - z_{\alpha/2}\, \sigma_{\bar{p}} \le p \le \bar{p} + z_{\alpha/2}\, \sigma_{\bar{p}}$$

where $q = 1 - p$ and $z_{\alpha/2}$ is the z-value corresponding to an area of α/2 in the right tail of the standard normal probability distribution and the quantity $z_{\alpha/2}\, \sigma_{\bar{p}}$ is the margin of error (or error of the estimation). Since p and q are unknown, they are estimated using the point estimator: $\bar{p}$ and $\bar{q}$. Thus for a *sample proportion*, the *standard error* denoted by SE ($\bar{p}$) or $\sigma_{\bar{p}}$ is given by

$$\sigma_{\bar{p}} = \sqrt{\bar{p}\,\bar{q}/n} = \sqrt{\bar{p}\,(1-\bar{p})/n}\,.$$

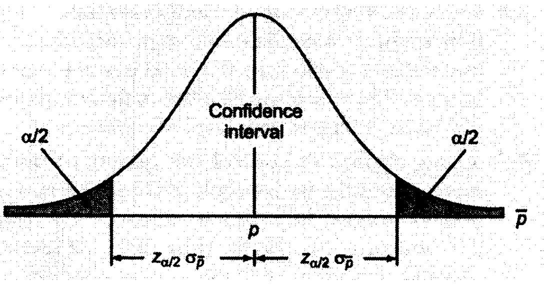


Figure 9.7
Sampling Distribution of Proportion $\bar{p}$

Example 9.7: Suppose we want to estimate the proportion of families in a town which have two or more children. A random sample of 144 families shows that 48 families have two or more children. Setup a 95 per cent confidence interval estimate of the population proportion of families having two or more children. [*HP Univ., MBA, 1999*]

Solution: The sample proportion is: $\bar{p} = x/n = 48/144 = 1/3$. Using the information, $n = 144$, $\bar{p} = 1/3$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence coefficient, we have

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}} = \frac{1}{3} \pm 1.96 \sqrt{\frac{(1/3)\,(2/3)}{144}} = 0.333 \pm 0.077$$

Hence the population proportion of families who have two or more children is likely to be between 25.6 to 41 per cent, that is, $0.256 \le p \le 0.410$.

Example 9.8: An auditor for an insurance company would like to determine the proportion of claims settled by the company within 2 months of the receipt of the claim. A random sample of 200 claims is selected, and it is determined that 80 were paid the money within 2 months of the receipt of the claim. Setup a 99 per cent confidence interval estimate of the population proportion of the claims paid within 2 months.

Solution: The sample proportion is: $\bar{p} = x/n = 80/200 = 0.40$. Using the information, $n = 200$, $\bar{p} = 0.40$ and $z_{\alpha/2} = 2.576$ at 95 per cent confidence coefficient, we have

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}} = 0.40 \pm 2.576 \sqrt{\frac{0.40 \times 0.60}{200}} = 0.40 \pm 0.088$$

Hence the population proportion of claims settled by the company within 2 months is likely to be between 30.8 and 48.8 per cent, that is, $0.308 \le p \le 0.488$.

Example 9.9: A shoe manufacturing company is producing 50,000 pairs of shoes daily. From a sample of 500 pairs, 2 per cent are found to be of substandard quality. Estimate at 95 per cent level of confidence the number of pairs of shoes that are reasonably expected to be spoiled in the daily production.

Solution: The sample proportion is, $\bar{p} = 0.02$, $\bar{q} = 0.98$. Using the information, $n = 500$, N = 50,000 and $z_{\alpha/2} = 1.96$ at 95 per cent confidence, we have

$$\begin{aligned}\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}\,(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}} &= 0.02 \pm 1.96 \sqrt{\frac{0.02 \times 0.98}{500}} \sqrt{\frac{50{,}000-500}{50{,}000-1}} \\ &= 0.02 \pm 1.96\ (0.0063)\ (0.9949) \\ &= 0.02 \pm 0.0122 \text{ or } 0.0078 \text{ and } 0.0322\end{aligned}$$

Hence, the expected number of shoes that are expected to be spoiled in the daily production is given by

$$50{,}000\ (0.0078 \le p \le 0.0322) \text{ or } 390 \le p \le 1610$$

Self-Practice Problems 9B

9.8 Out of 20,000 customer's ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing wherein 45 mistakes were found. Assign limits within which the number of defective cases can be expected at 95 per cent level of confidence.

9.9 In an attempt to control the quality of output for a manufactured part, a sample of parts is chosen randomly and examined in order to estimate the population proportion of parts that are defective. The manufacturing process operated continuously unless it must be stopped for inspection or adjustment. In the latest sample of 90 parts, 15 defectives are found. Determine a point estimate and interval estimate at 98 per cent confidence of population proportion defective.

9.10 A survey of 672 audited tax returns showed that 448 resulted in additional payments. Construct a 95 per cent confidence interval for the true percentage of all audited tax returns that resulted in additional payments.

9.11 In a survey carried out in a large city, 170 households out of a random sample of 250 owned at least one pet. Find the 95 per cent confidence interval for the percentage of households in the city who own at least one pet. Does the result support a pet food manufacturer's claim that 75 per cent of all households have atleast one pet?

9.12 A TV channel conducted a survey on political stability in the country. Out of 814 adults a total of 562 responded 'yes' to the question: Do you feel things are going well in the country these days?

(a) Determine the margin of error at 90 per cent confidence level.

(b) What is the 90 per cent confidence interval for the proportion of the adult population that feels things are going well in the country?

9.13 A ball pen manufacturer makes a lot of 10,000 refills. The procedure desires some control over these lots so that no lot will contain an excessive number of defective refills. He decides to take a random sample of 400 refills for inspection from a lot of 10,000 and finds 9 defectives. Obtain a 90 per cent confidence interval for the number of defectives in the entire lot.

Hints and Answers

9.8 Given $\bar{p}$ = proportion of mistakes, 45/600 = 0.075; $q = 1 - p = 0.925$; $n = 660$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$\bar{p} \pm z_{\alpha/2}\sqrt{pq/n} = 0.075 \pm 1.96\sqrt{\frac{0.075 \times 0.925}{660}}$$
$$= 0.075 \pm 1.96\ (0.011)$$
$$= 0.075 \pm 0.021;\ 0.053 \le p \le 0.097$$

Out of 20,000 cases expected number of defective cases would be between 20,000 × 0.053 = 1060 and 20,000 × 0.097 = 1940.

9.9 Given $\bar{p} = 15/90 = 0.167$, $\bar{q} = 1 - \bar{p} = 0.833$; $n = 90$ and $z_{\alpha/2} = 2.33$ at 98 per cent confidence

(a) Point estimate of population proportion = np = 16 approx.

(b) Interval estimate = $p \pm z_{\alpha/2}\sqrt{pq/n}$ = 0.167 ± 2.33 (0.0393) = 0.167 ± 0.092; $0.075 \le p \le 0.025$

9.10 Given $\bar{p}$ = 448/672 = 0.666; $\bar{q} = 1 - \bar{p} = 0.334$, $n = 672$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$\bar{p} \pm z_{\alpha/2}\sqrt{pq/n} = 0.666 \pm 1.96\sqrt{\frac{0.666 \times 0.334}{672}}$$
$$= 0.666 \pm 1.96\ (0.018)$$
$$= 0.666 \pm 0.036;\ 0.63 \le p \le 0.702$$

9.11 Given $\bar{p}$ = 170/250 = 0.68, $\bar{q} = 1 - \bar{p} = 0.32$, $n = 250$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$\bar{p} \pm z_{\alpha/2}\sqrt{pq/n} = 0.66 \pm 1.96\sqrt{0.68 \times \frac{0.32}{250}}$$
$$= 0.66 \pm 1.96\ (0.029)$$
$$= 0.66 \pm 0.059;\ 0.601 \le p \le 0.719$$

Since upper limit of confidence interval covers only 71.9 per cent households, it does not support the pet food manufacturer's claim.

9.12 Given $n = 814$; $\bar{p} = 562/814 = 0.69$; $\bar{q} = 1 - \bar{p} = 0.31$ and $z_{\alpha/2} = 1.645$ at 90 per cent confidence

(a) Margin of error, $\sigma_{\bar{p}} = z_{\alpha/2}\sqrt{pq/n}$

$$= 1.645\sqrt{\frac{0.69 \times 0.31}{814}}$$
$$= 1.645(0.0162) = 0.027$$

(b) $\bar{p} \pm z_{\alpha/2}\sqrt{pq/n} = 0.69 \pm 1.645\ (0.0162)$

$$= 0.69 \pm 0.027;$$
$$0.663 \le p \le 0.717$$

9.13 Given $p = 9/400 = 0.0225$, $q = 0.9775$; $n = 400$ and $z_{\alpha 2} = 1.645$ at 90 per cent confidence

$$\bar{p} \pm z_{\alpha/2}\sqrt{pq/n} = 0.0225 \pm 1.645\sqrt{\frac{0.0225 \times 0.9775}{400}}$$
$$= 0.0225 \pm 1.645\ (0.0074)$$
$$= 0.0225 \pm 0.122;\ 0.0103 \le p \le 0.0347$$

9.7 ESTIMATING SAMPLE SIZE

For sample statistics to infer about population, it is important to estimate the size of the sample. Since standard error $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $\sigma_{\bar{p}} = \sqrt{pq/n}$ of sampling distribution of sample statistic $\bar{x}$ and $\bar{p}$ are both inversely proportional to the sample size n, which is also related to the width of the confidence intervals $\bar{x} \pm z_{\alpha/2}\, \sigma_{\bar{x}}$ and $\bar{p} \pm z_{\alpha/2}\, \sigma_{\bar{p}}$, the width or range of the confidence interval can be decreased by increasing the sample size n.

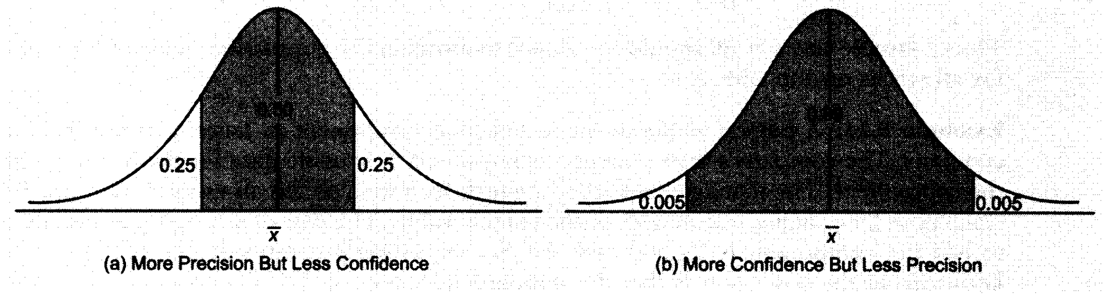
Precision of Confidence Interval The precision with which a confidence interval estimates the true population parameter is determined by the width of the confidence interval. *Narrow* the confidence interval, *more precise* the estimate and vice-versa. The width of confidence interval is influenced by

- Specified level of confidence
- Sample size
- Population standard deviation

To gain more precision, or confidence, or both, the sample size needs to be increased provided there is very less variability in the population itself. For certain reasons, if the sample size cannot be increased, the investigator can not afford the increased cost of sampling. So with the same sample size, the only way to maintain the desired level of precision is to lower the confidence level to an extent so as so predict that the estimate may become close to the target μ. A trade-off between precision and confidence levels is illustrated in Fig 9.8(*a*) and *b*. Figure 9.8(*a*) indicates that 50 per cent of the time, the true population mean μ will fall within the narrow range and the 25 per cent in each tail representing nonconfidence or the probability of making errors in estimation on either side. Figure 9.8(*b*) indicates that 99 per cent of the time it may be expected that the true mean μ will fall within the much wider range and there is only a 0.005 per cent chance that there could be an error in this estimation.

The decision regarding the appropriate size of sample depends on (i) Precision level needed in estimating the characteristics of a population, i.e. what is the *margin* of error to make? (ii) Confidence level needed, i.e. how much *chance* to make error in estimating a population parameter? (iii) Extent of *variability* in the population investigated? (iv) *Cost–benefit* analysis of increasing the sample size?

For example, an insurance company wants to estimate the proportion of claims settled within 2 months of the receipt of claim. For this purpose, the company must decide how much error it is willing to allow in estimating the population proportion of claims settled in a particular financial year. This means, whether accuracy is required to be within ± 80 claims, ± 100 claims, and so on. Also, the company needs to determine in advance the level of confidence for estimating the true population parameter. Hence for determining the sample size for estimating population mean or proportion, such requirements must be kept in mind along with information regarding standard deviation.



(a) More Precision But Less Confidence

(b) More Confidence But Less Precision

Figure 9.8 Trade-off between Precision and Confidence.

9.7.1 Sample Size for Estimating Population Mean

When the distribution of sample mean $\bar{x}$ is normal, the standard normal variable z is given as

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

The value of z can be seen from 'standard normal table' for a specified confidence coefficient $1 - \alpha$.

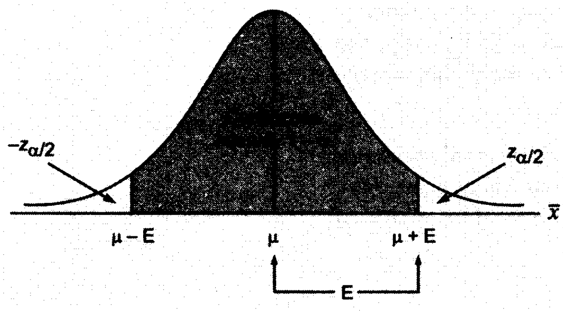


Figure 9.9
Sampling Distribution of Sample Mean

The value of z in the above equation will be positive or negative, depending on whether the sample mean $\bar{x}$ is larger or smaller than the population mean μ, as shown in Fig. 9.9. This difference between $\bar{x}$ and μ is called the *sampling error* or *margin of error* E. Thus for estimating the population mean μ with a condition that the error in its estimation should not exceed a fixed value, say E, we require that the sample mean $\bar{x}$ should fall within the range, $\mu \pm E$ with a specified probability. Thus the margin of error *acceptable* (i.e., maximum tolerable difference between unknown population mean μ and the sample estimate at a particular level of confidence) can be written as:

$$\bar{x} - \mu = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Solving for n yields a formula that can be used to determine sample size

$$\sqrt{n} = \frac{\sigma z_{\alpha/2}}{E} \quad \text{or} \quad n = \left(\frac{\sigma z_{\alpha/2}}{E}\right)^2$$

This formula for sample size n will provide the tolerable margin of error E, at the chosen confidence level $1 - \alpha$ (which determines the critical value of z from the normal table) with known or estimated population standard deviation σ.

Remarks

1. Another formula to determine the sample size is as follows:

$$n = \left(\frac{s^2}{\sigma_{\bar{x}}^2}\right) + 1$$

2. If population standard deviation σ is not known, then sample standard deviation s can be used to determine the sample size n.

Example 9.10: Suppose the sample standard deviation of P/E ratios for stocks listed on the Mumbai Stock Exchange (BSE) is $s = 7.8$. Assume that we are interested in estimating the population mean of P/E ratio for all stocks listed on BSE with 95 per cent confidence. How many stocks should be included in the sample if we desire a margin of error of 2?

Solution: The information given is: $E = 2$, $s = 7.8$, $z_{\alpha/2} = 1.96$ at 95 per cent level of confidence

Using the formula for n and substituting the given values, we have

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2} = \frac{(1.96)^2 (7.8)^2}{(2)^2} = \frac{3.84 \times 60.84}{4} = 59 \text{ approx.}$$

Thus a sample size $n = 59$ should be chosen to estimate the population mean of P/E ratio for all stocks on the BSE.

Example 9.11: A person wants to buy a machine component in large quantity from a company. The company's sales manager is requested to provide data for the mean life of the component. The manager considers it worth Rs 800 to obtain an estimate that has 19 chances in 20 of being within 0.05 of the correct value. The cost of setting up equipment to test the component is Rs 500 and the cost of testing a component is Rs 2.50. It is known from the past records that the standard deviation of the life of the component is 0.80. Will the manager be able to obtain the required estimate? What is the minimum cost of obtaining the necessary estimate?

Solution: We know that the money required to obtain an estimate is Rs 800; money required for setting up equipment to test is Rs 500, and the cost of testing a component is Rs 2.50. Thus

$$800 = 500 + 2.5n \quad \text{or} \quad n = (800 - 500)/2.5 = 120 \text{ (sample size)}$$

Also given that, sample standard deviation, $s = 0.8$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence level. Thus, the confidence interval for population mean μ is:

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm 1.96\,\frac{0.8}{\sqrt{120}}$$

$$= \bar{x} \pm 1.96(0.073) = \bar{x} \pm 0.143$$

Thus the confidence interval to estimate the population mean μ value is:

$$\bar{x} - 0.143 \leq \mu \leq \bar{x} + 0.143.$$

Since the difference in estimate is required to be 0.05, new sample size required is

$$E = z_{\alpha/2}\frac{s}{\sqrt{n}} \text{ or } 0.05 = 1.96\,\frac{0.8}{\sqrt{n}}, \text{ i.e. } n = 984 \text{ units (approx.)}$$

Thus the new cost for testing the component in the sample of 984 units will be 2.5 × 984 = Rs 7560.

Remark: The general rule used in determining sample size is always to rounded off to the nearest integer value in order to slightly oversatisfy the desire of estimation.

9.7.2 Sample Size for Estimating Population Proportion

The method for determining a sample size for estimating the population proportion is similar to that used in the previous section. The procedure begins with z formula for sample proportions:

$$z = \frac{\bar{p} - p}{\sqrt{pq/n}}, q = 1 - p$$

Since sample proportion $\bar{p}$ will rarely equal the population proportion p, resulting in an error of estimation, which is written as $E = \bar{p} - p$. We require that the population proportion p should fall within the range $\bar{p} \pm$ E, with a specified probability

$$z = \frac{E}{\sqrt{pq/n}} \quad \text{or} \quad E = z_{\alpha/2}\sqrt{\frac{pq}{n}} \;; q = 1 - p$$

$$E^2 = (z_{\alpha/2})^2\,\frac{pq}{n}, \text{ i.e. } n = \frac{(z_{\alpha/2})^2 pq}{E^2}$$

The value of z can be calculated from 'Standard normal table' for a specified confidence coefficient.

This formula for n will provide the desired *margin of error* E at the chosen confidence level $1 - \alpha$ (which determines the critical value of z) with known or estimated population proportion p.

Example 9.12: A car manufacturing company received a shipment of petrol filters. These filters are to be sampled to estimate the proportion that is unusable. From past experience, the proportion of unusable filter is estimated to be 10 per cent. How large a random sample should be taken to estimate the true proportion of unusable filters to within 0.07 with 99 per cent confidence.

Solution: The information given is: E = 0.07, p = 0.10, and $z_{\alpha/2} = 2.576$ at 99 per cent confidence level.

Using the formula for n and substituting the given values, we have

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(2.576)^2(0.10 \times 0.90)}{(0.07)^2} = 121.88$$

Therefore a slightly larger sample size of $n = 122$ filters should be taken.

9.7.4 Sample Size Determination for Finite Population

In Chapter 8, we introduced the concept of *finite population correction factor* when samples are drawn without replacement from a finite population of size N. The use of such a factor reduces the standard error by a value equal to $\sqrt{(N-n)/(N-1)}$. For example, for deciding sample size n for estimating the population mean μ, the desired margin of error is given by

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Similarly, when estimating the proportion, the desired margin of error is given by

$$\sigma_{\bar{p}} \text{ or } E = z_{\alpha/2}\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

Let n_0 be the size for estimating population mean without using correction factor. Then

$$n_0 = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

The revised sample size, taking into consideration the size of the population, is given by

$$n = \frac{n_0 N}{n_0 + (N-1)}$$

Example 9.13: For a population of 1000, what should be the sampling size necessary to estimate the population mean at 95 per cent confidence with a sampling error of 5 and the standard deviation equal to 20?

Solution: We have E = 5, σ = 20, $z_{\alpha/2}$ = 1.96 at 95 per cent confidence level, and N = 1000. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(20)^2}{(5)^2} = 61.456$$

Since the population size is finite, the revised sample size obtained by using the correction factor

$$n = \frac{n_0 N}{n_0 + (N-1)} = \frac{(61.456)(1000)}{61.456 + (1000-1)}$$

$$= \frac{61456}{1060.456} = 57.952$$

Conceptual Questions 9A

Thus a slightly larger sample size of n = 58 should be taken.

1. Why is estimation important? What are the advantages of using interval estimates rather than point estimates for statistical inference?
2. Is there is any relationship between estimation and descriptive statistics? Explain.
3. Distinguish between the point estimation and interval estimation. Explain how an interval estimate is better than a point estimate.
4. How does sampling without replacement from a finite population affect the confidence interval estimate and the sample size necessary?
5. When is the student's t distribution used in developing the confidence interval estimate for the mean?
6. For a given sample size, an increase in confidence level is achieved by widening (or making less precise) the confidence interval obtained. Explain
7. Explain the concept of 'margin of error' in deciding the size of a sample.
8. Prove that the mean of a simple random sample from a given population is an unbiased estimator of the population mean.
9. Under what circumstances can the normal distribution be used to construct a confidence interval estimate of the population mean?
10. What are the properties of a good estimator? Explain, how these properties are essential for estimating the population characteristic of interest.
11. Distinguish between statistic and parameter and explain the meaning of confidence interval of a population parameter.
12. Explain the following terms with an example
    (a) Point estimate (b) Interval estimate
    (c) Confidence interval (d) Confidence limits
    (e) Confidence coefficients or critical values.

13. Describe the effect of sample size on the margin of sampling error of point estimate of the proportion mean. Does this error depends on the sample size in the same way?

14. In determining the accuracy of an estimator, what must be known and why is this important?

15. Based on the knowledge about the desirable qualities of estimators, for what reasons might $\bar{x}$ be considered the best estimator of the ture population mean?

16. Why is the size of a statistic's standard error important in its use as an estimator? To which characteristic of estimator does this relate?

Self-Practice Problems 9C

9.14 Given a population with a standard deviation of 8.6. What sample size is needed to estimate the mean of population within ±0.5 with 99 per cent confidence?

[*Delhi Univ., MCom, 1999*]

9.15 A cigarette manufacturer wishes to use random sampling to estimate the average nicotine content. The sampling error should not be more than one milligram above or below the true mean with a 99 per cent confidence coefficient. The population standard deviation is 4 milligrams. What sample size should the company use in order to satisfy these requirements?

9.16 A firm wishes to estimate with a maximum allowable error of 0.05 and a 95 per cent level of confidence, the proportion of consumers who prefer its product. How large a sample will be required in order to make such an estimate if the preliminary sales reports indicate that 25 per cent of all consumers prefer the firm's product?

9.17 A agency responsible for electricity distribution would like to estimate the average electric bills for a particular month for single-family homes in a large city. Based on studies conducted in other cities, the standard deviation is assumed to be Rs 40. The agency would like to estimate the average bill for that month to within ±Rs 10 of the true average. If 95 per cent confidence is desired, then what sample size is necessary?

9.18 A private TV channel would like to estimate the proportion of voters who vote for a particular political party's candidate in the next general election for the Lok Sabha. If he wants to have 95 per cent confidence that his prediction is within ±0.08 of the population proportion, then what sample size is needed, considering sampling error of ±0.03?

9.19 If a decision-maker wants 95 per cent confidence with a sampling error of 5 and the standard deviation equal to 10 to estimate of the mean μ for a population of 2000, what sample size would be required?

Hints and Answers

9.14 Given, $\sigma = 8.6$, $E = \pm\, 0.5$ and $z_{\alpha/2} = 2.576$ at 99 per cent confidence

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = (2.576)^2\,\frac{(8.6)^2}{(0.5)^2}$$

$$= \left(\frac{6.635 \times 73.96}{0.25}\right) \cong 1963$$

9.15 Given $\sigma = 4$, $E = 1$, and $z_{\alpha/2} = 2.576$ at 99 per cent confidence

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(2.576)^2(4)^2}{(2)^2} \cong 107$$

9.16 Given $p = 0.25$, $E = 0.05$, and $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(1.96)^2\,(0.25 \times 0.75)}{(0.05)^2} \cong 407$$

9.17 Given $\sigma = 40$, $E = \pm\, 10$ and $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(40)^2}{100} \cong 62$$

9.18 Given $p = 0.08$, $q = 0.92$, $E = \pm\, 0.03$, $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2}$$

$$= \frac{(1.96)^2(0.08 \times 0.92)}{(0.03)^2} = 315.$$

9.19 Given $\sigma = 10$, $E = 5$, $N = 2000$, $z_{\alpha/2} = 1.96$ at 95 per cent confidence

$$n_0 = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2 100}{25} = 15.366$$

Using the correction factor, we have

$$n = \frac{n_0\,N}{n_0 + (N - 1)}$$

$$= \frac{15.366\,(2000)}{15.366 + (2000 - 1)}$$

$$= \frac{30732.80}{2014.366} \cong 16$$

Formulae Used

1. 100 (1 – α)% confidence interval to estimate μ

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

2. Confidence interval to estimate population mean μ: small sample size ($n \le 30$) and population standard deviation σ is unknown

(a) $\mu = \bar{x} \pm t_{\sigma/2}\, \sigma_{\bar{x}}$, when $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$

- $\bar{x} - t_{\sigma/2,\, n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\sigma/2,\, n-1}\frac{s}{\sqrt{n}}$

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

3. Confidence interval to estimate μ: large sample size ($n > 30$) and population standard deviation σ is known

- $\mu = \bar{x} \pm z_{\alpha/2}\, \sigma_{\bar{x}}$, where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

4. Confidence interval to estimate μ using finite correction factor: $\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

5. Confidence interval to estimate population proportion large sample size $n \ge 30$

$$p = \bar{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

6. Confidence interval to estimate the difference between the means of two normally distributed populations

- When standard deviations are known

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- When standard deviations are not known

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7. Sample size when estimating:

- Population mean, μ

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

- Population proportion, p

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2};\ q = 1 - p$$

8. Sample size when estimating finite population mean, μ with size N

$$n = \frac{n_0 N}{n_0 + (N - 1)},\ \text{where } n_0 = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

Chapter Concepts Quiz

True or False

1. Interval estimates for the parameters μ and p are based on a random sample of observations from the population or process of interest. (T/F)
2. The confirmatory analysis is used to assess the insights resulting from exploratory analysis. (T/F)
3. While estimating a parameter, it is not important to provide the amount of possible error in the point estimate. (T/F)
4. The sampling distribution of $\bar{x}$ is not necessarily normal with unknown mean and standard error. (T/F)
5. The error in the point estimate is known as margin of sampling error associated with a given level of confidence. (T/F)
6. Confidence limits are the range of values used to estimate the shape of the population distribution. (T/F)
7. Student's t-distribution is always used when the standard deviation of the population in not known. (T/F)
8. The width of a confidence interval is twice the margin of the sampling error. (T/F)
9. The level of confidence represents the probability with which population parameter is expected to fall within a given interval estimate. (T/F)
10. The degrees of freedom used in t-distribution estimation are equal to the sample size. (T/F)
11. There is a different critical value for each level of confidence.
12. The population is assumed to be normally distributed before using t-distribution for interval estimation. (T/F)
13. The shape of the t-distribution tends to become flatter as sample size increases. (T/F)
14. If a random variable x is normally distributed, the statistic $(\bar{x} - \mu)/(s/\sqrt{n})$ has a t-distribution with $n - 1$ degrees of freedom. (T/F)
15. The t-distribution has more area in the tails and less in the center than does the normal distribution. (T/F)

Multiple Choice

16. When the value of population standard deviation σ is unknown, the values of t in the t-distribution are:
(a) more variable than for z (b) less variable than for z
(c) equal to z (d) none of these

17. Which of the following is a necessary condition for using a t-distribution?
(a) small sample size
(b) unknown population standard deviation
(c) both (a) and (b)
(*d*) infinite population

18. The interval estimate of a population mean with large sample size and known standard deviation, is given by:
(a) $\bar{x} \pm z_{\alpha/2}\, \sigma_{\bar{x}}$ (b) $\bar{x} \pm z_{\alpha/2}\, s_{\bar{x}}$
(c) $\bar{x} \pm t_{\alpha/2}\, \sigma_{\bar{x}}$ (d) $\bar{x} \pm t_{\alpha/2}\, s_{\bar{x}}$

19. Which part of the area under the normal curve is represented by the coefficient $z_{\alpha/2}$?
(a) left tail (b) right tail
(c) both tails (d) none of these

20. If $\bar{x} = 85$, $\sigma = 8$, and $n = 64$, then standard error of sample mean is equal to:
(a) 1 (b) 1.96
(c) 2.576 (d) none of these

21. If $\bar{x} = 25$, $\sigma = 5$, and $n = 25$, the margin of sampling error at 95 per cent confidence is:
(a) 1 (b) 1.96
(c) 2.576 (d) none of these

22. Sampling distribution is usually the distribution of:
(a) parameter (b) statistic
(c) mean (c) variance

23. The criteria for the best estimator are
(a) consistency and efficiency
(b) unbiasedness and sufficiency
(c) consistency and sufficiency
(d) all of the above

24. An unbiased estimator is necessarily
(a) consistent (b) not consistent
(c) efficient (d) none of these

25. $\bar{\theta}_n$ is consistent estimator of θ if $n \to \infty$:
(a) $\text{var}(\bar{\theta}_n) \to \theta$ (b) $\text{var}(\bar{\theta}_n) = \theta$
(c) $\text{var}(\bar{\theta}_n) = \infty$ (d) $\text{var}(\bar{\theta}_n) \to 0$

26. If $E(\theta_n) < \theta$, the estimator is called:
(a) unbiased (b) positively biased
(c) negatively biased (d) none of these

27. Maximum likelyhood estimator μ in normal population with
(a) sample variance (b) sample mean
(c) sample median (d) none of these

28. Sample mean $\bar{x}$ of a sample size n from $N(\mu, 1)$ is distributed as:
(a) $N(0, 1)$ (b) $N(n\mu, 1/n)$
(c) $N(\mu, 1/n)$ (d) none of these

29. The standard deviation of the sampling distribution of a statistic is referred as
(a) sampling error (b) standard error
(c) mean error (d) none of these

30. Maximum likelihood estimator of μ of normal distribution is:
(a) $\bar{x}/n$ (b) $\bar{x}$
(c) $n\bar{x}$ (d) none of these

31. The degrees of freedom used in a t-distribution are equal to
(a) sample size n (b) sample size $n - 1$
(c) sample size $n + 1$ (d) (a) or (b) but not (c)

32. As the width of a confidence interval increases, the confidence level associated with the interval
(a) tend to increase (b) tend of decrease
(c) remains same (d) none of these

33. In which probability distribution, 100 per cent of the population mean lies within ±3 standard deviations
(a) Binomial (b) Poisson
(c) Normal (d) Exponential

34. If a sample statistic either underestimates or overestimates population , then it is called
(a) consistent (b) unbiased
(c) efficient (d) none of these

35. If a normally distributed population has standard deviation, $\sigma = 1$, then the total width of the 95 per cent confidence interval for the population mean is
(a) 1.28 (b) 1.64
(c) 1.96 (d) None of these

Concepts Quiz Answers

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

Review Self-Practice Problems

9.20 A hospital wants an estimate of the mean time that a doctor spends with each patient in the OPD. How large a sample should be taken if the desired margin of error is 2 minutes at a 95 per cent level of confidence, assuming a population standard deviation of 8 minutes?

9.21 A bank is interested in estimating the proportion of credit cardholders who carry a non-zero balance at the end of the month and incur an interest charge. Assume that the desired margin of error is 0.03 at 99 per cent confidence. How large a sample should if it be is anticipated that roughly 80% of the cardholders carry a non-zero balance at the end of the month?

9.22 In a survey, 220 people out of 400, when asked to identify their major source of news information, stated that their major source was television news.

(a) Construct a 95 per cent confidence internal for the proportion of people in the population who consider television their major source of news information.

(b) How large a sample would be necessary to estimate the population proportion with a margin of error of 0.05 at 95 per cent confidence.

9.23 The weight of cement in packed bags in distributed normally with a standard deviation of 0.2 kg. A sample of 35 bags is picked up at random and the mean weight of cement in these bags is only 49.7 kg

(a) Find out the 90 per cent confidence interval for the mean weight of the cement bags

(b) How large should a sample be so that mean weight of cement in filled bags can be estimated with margin of error ±0.05 kg of the true value at 90 per cent confidence level.

9.24 In an auditing firm, the auditors in their annual audit audited 100 entries out of a ledger containing 1000 entries. It showed that the stated total amount received exceeded the correct amount receivable by Rs 120. Determine the 95 per cent confidence limit of the amount by which the reported total receivables in the whole ledger exceed the correct amount. Assume standard deviation as Rs 6 in the population of the book-keeping error.

9.25 A life insurance company has 1500 policies averaging Rs 2,00,000 on lives at age 40. From experience, it is found that of 1,00,000 alive at the age of 30, 99,000 alive at age 31. Find the highest and lowest amount of money that company will have to pay out during the year.

9.26 An automobile association wants to solicit the assistance of some of its many thousands of members. The association wants the members selected to keep detailed records on the cost of maintaining and running their automobiles over a period of years.

Assume that the automobile association has the following requirements: of 95 per cent confidence coefficient, a standard deviation of mean within Rs 50 of the true mean, the sample chosen in an earlier study had a standard deviation of Rs 400.

(a) How many persons should be asked to keep cost records?

(b) With the records of the number of persons recommended by you in part (a), suppose the standard deviation of these costs was computed to be 300, was the sample size recommended in part (a) too large, too small, or exactly correct?

9.27 A leading detergent manufacturer believes that about 30 per cent of all housewives use his product, but he intends to conduct a survey to check whether or not he is justified in his belief.

(a) What must be the size of the sample to enable the manufacturer to estimate the true percentage to within 2 per cent.

(b) After the sample is taken, the manufacturer discovers that the estimated population percentage is 28 per cent. Will a large sample be necessary?

9.28 A TV channel wishes to estimate within a margin of error of 5 per cent, the percentage of viewers in a specific area that prefer a given programme with 95 per cent level of confidence. The TV channel has no information concerning the likely percentage of viewers preferring the programme. Determine the approximate sample size to help the TV channel to attain its objective.

9.29 Machines are used to pack sugar into packets supposedly containing 1.20 kg each. On testing a large number of packets over a long period of time, it was found that the mean weight of the packets was 1.24 kg and the standard deviation was 0.04 kg. A particular machine is selected to check the total weight of each of the 25 packets filled consecutively by the machine. Calculate the limits within which the weight of the packets should lie assuming that the machine is not be classified as faulty.

9.30 A manufacturer of a synthetic fibre takes periodic samples to estimate the mean breaking strength of the fibre. For each periodic sample, an experiment is devised in which the breaking strength (in kgs) are observed for 16 randomly selected samples from the production process. The strengths of the current samples are: 20.8, 20.6, 21.0, 20.9, 19.9, 20.2, 19.8, 19.6, 20.9, 21.1, 20.4, 19.7, 19.6, 20.3, 16.6, and 20.7. Determine 95 per cent confidence interval for the average breaking strength of the fibre, assuming that breaking strength is adequately approximated by a normal distribution.

9.31 A vendor claims that no more than 8 per cent of parts shipped to a manufacturer fails to meet specifications. The manufacturer selects at random 200 parts from a large batch just received from the vendor and finds 19 defective parts. Determine the extent to which the current sample contradicts the vendor's claim.

9.32 A cable operator needs to estimate the percentage of households in a city which would take cable connection if it was made available. Based on responses in other cities, the percentage certainty is somewhere between 10 and 20 per cent. What sample size should be selected if the promoter wants the estimate to be correct within two percentage points and with 99 per cent confidence?

9.33 Mr Mohan is interested in purchasing a Maruti-800 used car, selected 64 sale ads and found that the average price of a car in this sample was Rs 1,62,500. He knows that the standard deviation of used-car prices is Rs 30,750.

Determine an interval estimate for the average price of a car so that he can be 68.3 per cent certain that the population mean lies within this interval.

9.34 The department of drug control recently raided premises of drug dealers in the city. Out of 12,000 drug dealers only 720 have been caught. The mean rupee value of drugs found on these 700 dealers is Rs 1,12,50,000. The standard deviation of the rupee value of drugs is Rs 18,45,000. Construct a 90 per cent confidence interval for the mean rupee value of drugs possessed by these drug dealers.

9.35 For a year and half, sales have been falling constently in all 300 franchises of a fast-food chain all over the country. It has been discovered that 30 per cent of a sample of 35 indicate clean signs of mis-management. Construct 98 per cent confidence interval for this proportion.

9.36 Given a sample mean of 8, a population standard deviation of 2.6 and a sample size of 32, find the confidence level associated with the interval (17.613, 8.386).

9.37 An agriculture research scientist discovered that a certain strain of corn will always produce between 80 and 140 bushels per acre. For a confidence level of 90 per cent, how many 1-acre samples must be taken in order to estimate the average production per acre to within ±5 bushels per acre.

Hints and Answers

9.20 GivenE = 2, $z_{\alpha/2}$ = 1.96 at 95 per cent confidence and $\sigma = 8$

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = (1.96)^2\frac{(8)^2}{(2)^2} \cong 62$$

9.21 GivenE = 0.03, $z_{\alpha/2}$ = 2.576 at 99 per cent confidence; $p = 0.80, 8 = 1-p = 0.20$

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(2.576)^2(0.80)(0.20)}{(0.03)^2}$$

$$= \frac{1.0617}{0.0009} = 1180$$

9.22 Given$p$ = 220/400 = 0.55, q = 0.45, $z_{\alpha/2}$ = 1.96 at 95 per cent confidence, E = 0.05

$$\text{(a) } \bar{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} = 0.55 \pm 1.96\sqrt{\frac{0.55\times 0.45}{400}}$$

$$= 0.55 \pm 0.049;\ 0.501 \le p \le 0.599$$

$$\text{(b) } n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(1.96)^2(0.55)(0.45)}{(0.05)^2}$$

$$= \frac{0.9507}{0.0025} = 381$$

9.23 Given $\bar{x}$ = 49.7, σ = 0.20, n = 35, E = 0.05, $z_{\alpha/2}$ = 1.645 at 90 per cent confidence

$$\text{(a) } \bar{x} \pm z_{\alpha/2}\,\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 49.7 \pm 1.645\frac{0.20}{\sqrt{35}}$$

$$= 49.7 \pm 0.0556;\ 49.644 \le \mu \le 49.755$$

$$\text{(b) } n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.645)^2(0.20)^2}{(0.05)^2} = \frac{2.706\times 0.04}{0.0025}$$

= 44 bags

9.24 Given $\bar{x}$ = 120/100 = 1.20, σ = 6, n = 100, N = 1000, $z_{\alpha/2}$ = 1.645 at 90 per cent confidence

Expected number of entries exceeded

$$= N\left(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= 1000\left(1.20 + 1.645\frac{\sigma}{\sqrt{n}}\right) = 2187$$

9.25 Given$p$= chance that a person aged 30 would not alive upto an age 31

$$= 1 - \frac{99{,}000}{1{,}00{,}000} = \frac{1}{100};\ q = 1 - p = \frac{99}{100}$$

Expected number of deaths in 1500 policies

$$= 1500 \times \frac{1}{100} = 15$$

$$\text{Standard error} = \sqrt{npq} = \sqrt{1500\times\left(\frac{1}{100}\right)\left(\frac{99}{100}\right)}$$

$$= 3.85$$

Interval estimate of actual number of deaths at 99.97 per cent is:

$$p \pm z_{\alpha/2}\sqrt{npq} = 15 \pm 3\,(3.85);\ 3.45 \text{ to } 26.55$$

Thus expected amount (both lower and upper limit) of money to be paid by the company will be:

2,00,000(3.45) to 2,00,000(26.55)

or Rs 6,90,000 to Rs 53,10,000

9.26 (a) Given σ = 400, n = 246 and $z_{\alpha/2}$ = 1.96 at 95 per cent confidence

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96\frac{400}{\sqrt{246}} = \bar{x} \pm 49.987$$

or $\bar{x} \pm 50$

$$\text{(b) } \bar{x} \pm 1.96\frac{300}{\sqrt{n}} = \bar{x} \pm 50 \text{ or } 1.96\frac{300}{\sqrt{n}} = 50 \text{ or } n$$

= 139 which is too less than n

= 246 in part (a).

9.27 Given p = 0.30, q = 0.70, E = 0.02; $z_{\alpha/2}$ = 1.96 at 95 per cent confidence

$$\text{(a) } n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(1.96)^2(0.30)(0.70)}{(0.02)^2} = 202$$

(b) Since p = 0.28 < 0.30, no increase in sample is required

9.28 Let$p$= 0.50 and q = 0.50; $z_{\alpha/2}$ = 1.645 at 90 per cent confidence, E = 0.05

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(1.645)^2(0.50)(0.50)}{(0.05)^2} = 271$$

9.29 As sample size n = 25 (< 30) is small, t-distribution is used to calculate confidence interval.

Given s = 0.04, n = 25, $df = n - 1$ = 24, $t_{\alpha/2}$ = 2.064 at 95 per cent confidence and degrees of freedom, df = 24. Thus, limits are

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 1.24 \pm 2.064\,\frac{0.04}{\sqrt{25}}$$

$$= 1.24 \pm 0.0165;\ 1.2235 \le \mu \le 1.2565$$

9.30 From the given data, we have $\bar{x} = \sum x/n = 20.381, s = 0.523, n = 16; t_{\alpha/2} = 2.131$ at 95 per cent confidence and degrees of freedom = 15

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 20.381 \pm 2.131\,\frac{0.523}{\sqrt{16}}$$

$$= 20.381 \pm 0.278;\ 20.103 \le 20.659$$

9.31 Given $p = 19/400 = 0.0475$, $q = 0.9525$, $n = 400$, $z_{\alpha/2} = 1.34$ at 92 per cent confidence

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} = 0.0475 \pm 1.34\sqrt{\frac{0.0475 \times 0.9525}{400}}$$

$$= 0.0475 \pm 0.0144;\ 0.0331 \le p \le 0.0619$$

that is, 3.31 per cent to 6.19 per cent may fail to meet specifications.

9.32 Given $p = 0.20$, $q = 0.80$, $E = 0.02$, $z_{\alpha/2} = 2.58$ at 99 per cent confidence

$$n = \frac{(z_{\alpha/2})^2 pq}{E^2} = \frac{(2.58)^2(0.20)(0.80)}{(0.02)^2} = 2663$$

9.33 Given $\bar{x} = 1{,}62{,}500$, $\sigma = 30{,}750$ and $n = 64$. So $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 30{,}750/\sqrt{64} = 3843.75$ for 68.3 per cent confidence interval

$$\bar{x} \pm \sigma_{\bar{x}} = 1{,}62{,}500 \pm 3843.75$$

$$= (\text{Rs } 1{,}58{,}656.25,\ \text{Rs } 1{,}66{,}343.75)$$

9.34 $N = 12{,}000, n = 720, \bar{x} = 1{,}12{,}50{,}000, s = 18{,}45{,}000$; $n/N = 720/12000 = 0.06 > 0.05$

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{18{,}45{,}000}{\sqrt{720}}\sqrt{\frac{12000-720}{12000-1}}$$

$$= 2{,}10{,}701.82$$

9.35 Given $N = 300, n = 35, n/N = 35/300 = 0.117, \bar{p} = 0.30$

$$\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.30 \times 0.70}{35}}\sqrt{\frac{300-35}{300-1}}$$

$$= 0.077 \times 0.941 = 0.072$$

98 per cent confidence interval: $\bar{p} \pm 2.33\,\sigma_{\bar{p}}$

$$= 0.30 \pm 2.33 \times 0.072 = (0.132 \text{ to } 0.468)$$

9.36 Given $\bar{x} = 32, \sigma = 2.6$ and $n = 32$. So

$$\frac{8.386 - 7.613}{2} = \frac{z\sigma}{\sqrt{n}} \text{ or } 0.386 = \frac{2.6\sigma}{\sqrt{32}}$$

or $$z = \frac{0.386\sqrt{32}}{2.6} = 0.84$$

$P(|z| \le 0.84) = 2(0.2995) = 0.5990$. It is a 60 per cent interval.

9.37 Given $6\sigma = 140 - 80$, or $\sigma = 10$; So

$$5 = 1.64\frac{\sigma}{\sqrt{n}} = \frac{1.64 \times 10}{\sqrt{n}}$$

or $$n = \left(\frac{1.64 \times 10}{5}\right)^2 = 10.75 \cong 11$$

Chapter 10

If there is an opportunity to make a mistake, sooner or later the mistake will be made.

—Edmond C Berekely

Hypothesis Testing

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- explain why hypothesis testing is important
- know how to establish null and alternative hypotheses about a population parameter
- develop hypothesis testing methodology for accepting or rejecting null hypothesis.
- use the test statistic *z*, *t*, and *F* to test the validly of a claim or assertion about the true value of any population parameter.
- understand Type I and Type II errors and its implications in making a decision.
- compute and interpret p-values
- interpret the confidence level, the signifience level, and the power of a test

10.1 INTRODUCTION

Inferential statistics is concerned with estimating the true value of population parameters using sample statistics. Chapter 9 contains three techniques of inferential statistics, namely, a (i) *point estimate*, (ii) *confidence interval that is likely to contain the true parameter value, and* (iii) *degree of confidence associated with a parameter value which lies within an interval values.* This information helps decision-makers in determining an interval estimate of a population parameter value with the help of sample statistic along with certain level of confidence of the interval containing the parameter value. Such an estimate is helpful for drawing statistical inference about the characteristic (or feature) of the population of interest.

Another way of estimating the true value of population parameters is to test the validity of the claim (assertion or statement) made about this true value using sample statistics.

10.2 HYPOTHESIS AND HYPOTHESIS TESTING

A *statistical hypothesis* is a claim (assertion, statement, belief or assumption) about an unknown population parameter value. For example (i) a judge assumes that a person charged with a crime is innocent and subject this assumption (hypothesis) to a verification by reviewing the evidence and hearing testimoney before reaching to a verdict (ii) a phamaceutical company claims the efficacy of a medicine against a disease that 95 per cent of all persons suffering from the said disease get cured (iii) an investment company claims that the average return across all its investments is 20 per cent, and so on. To test such claims or assertions statistically, sample data are collected and analysed. On the basis of sample findings the hypothesized value of the population parameter is either accepted or rejected. *The process that enables a decision maker to test the validity (or significance) of his claim by analysing the difference between the value of sample statistic and the corresponding hypothesized population parameter value, is called hypothesis testing.*

10.2.1 Formats of Hypothesis

As stated earlied, a hypothesis is a statement to be tested about the true value of population parameter using sample statistics. A hypothesis whether there exists any significant difference between two or more populations with respect to any of their common parameter can also be tested. To examine whether any difference exists or not, a hypothesis can be stated in the form of *if-then* statement. Consider, for instance, the nature of following statements:

- If inflation rate has decreased, then wholesale price index will also decrease.
- If employees are healthy, then they will take sick leave less frequently.

If terms such as 'positive,' 'negative,' 'more than,' 'less than,' etc are used to make a statement, then such a hypothesis is called *directional hypothesis* because it indicates the direction of the relationship between two or more populations under study with respect two a parameter value as illustrated below:

- Side effects were experienced by less than 20 per cent of people who take a particular medicine.
- Greater the stress experienced in the job, lower the job satisfaction to employees.

The *nondirectional hypothesis* indicates a relationship (or difference), but offer no indication of the direction of relationships (or differences). In other words, though it may be obvious that there would be a significant relationship between two populations with respect to a parameter, we may not be able to say whether the relationship would be positive or negative. Similarly, even if we consider that two populations differ with respect to a parameter, it will not be easy to say which population will be more or less. Following examples illustrate non-directional hypotheses.

- There is a relationship between age and job satisfaction.
- There is a difference between average pulse rates of men and women.

10.3 THE RATIONALE FOR HYPOTHESIS TESTING

The inferential statistics is concerned with estimating the unknown population parameter by using sample statistics. If a claim or assumption is made about the specific value of population parameter, then it is expected that the corresponding sample statistic is close to the hypothesized parameter value. It is possible only if hypothesized parameter value is correct and the sample statistic turns out to be a good estimator of the parameter. This approach to test a hypothesis is called a *test statistic*.

Since sample statistics are random variables, therefore their sampling distributions show the tendency of variation. Consequently we do not expect the sample statistic value to be equal to the hypothesized parameter value. The difference, if any, is due to chance and/or sampling error. But if the value of the sample statistic differs significantly from the hypothesized parameter value, then the question arises whether the hypothesized parameter value is correct or not. The greater the difference between the value of the sample statistic and hypothesized parameter, the more doubt is there about the correctness of the hypothesis.

In statistical analysis, difference between the value of the sample statistic and hypothesized parameter is specified in terms of the given level of probability whether the particular level of difference is significant or not when the hypothesized parameter value is correct. The probability that a particular level of deviation occurs by chance can be calculated from the known sampling distribution of the test statistic.

The probability level at which the decision-maker concludes that observed difference between the value of the test statistic and hypothesized parameter value cannot be due to chance is called the *level of significance* of the test.

10.4 GENERAL PROCEDURE FOR HYPOTHESIS TESTING

As mentioned before, to test the validity of the claim or assumption about the population parameter, a sample is drawn from the population and analysed. The results of the analysis are used to decide whether the claim is true or not. The steps of general procedure for any hypothesis testing are summarized below:

Step 1: State the Null Hypothesis (H_0) and Alternative Hypothesis (H_1)

The **null hypothesis** H_0 (read as H_0 nsub-zero) represents the claim or statement made about the value or range of values of the population parameter. The capital letter H stands for hypothesis and the subscript *'zero' implies 'no difference'* between sample statistic and the parameter value. Thus hypothesis testing requires that the null hypothesis be considered *true* (*status quo* or *no difference*) until it is proved false on the basis of results observed from the sample data. The null hypothesis is always expressed in the form of mathematical statement which includes the sign (≤, = ≥) making a claim regarding the specific value of the population parameter. That is:

$$H_0 : \mu \ (\leq, =, \geq) \ \mu_0$$

where μ is population mean and μ_0 represents a hypothesized value of μ. Only one sign out of ≤, = and ≥ will appear at a time when stating the null hypothesis

Null hypothesis: The hypothesis which is initially assumed to be true, although it may in fact be either true or false based on the sample data.

An **alternative hypothesis**, H_1, is the counter claim (statement) made against the value of the particular population parameter. That is, an alternative hypothesis must be true when the null hypothesis is found to be false. In other words, the alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis and is written as:

Alternative hypothesis: The hypothesis concluded to be true if the null hypothesis is rejected.

$$H_1 : \mu \neq \mu_0$$

Consequently $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$

Each of the following statements is an example of a null hypothesis and alternative hypothesis:

- $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$
- $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$
- $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$

(a) **Directional hypothesis**

- H_0 : There is no difference between the average pulse rates of men and women
  H_1 : Men have lower average pulse rates than women do
- H_0 : There is no relationship between exercise intensity and the resulting aerobic benefit
  H_1 : Increasing exercise intensity increases the resulting aerobic benefit
- H_0 : The defendent is innocent
  H_1 : The defendent is guilty

(b) **Non-Directional hypothesis**

- H_0 : Men and Women have same verbal abilities
  H_1 : Men and women have different verbal abilities
- H_0 : The average monthly salary for management graduates with a 4-year experience is Rs 75,000.
  H_1 : The average monthly salary is not Rs 75,000.

- H_0 : Older workers are more loyal to a company
  H_1 : Older workers may not be loyal to a company

Step 2: State the Level of Significance, α (alpha)

The level of significance, usually denoted by α (alpha), is specified before the samples are drawn, so that the results obtained should not influence the choice of the decision-maker. It is specified in terms of the probability of null hypothesis H_0 being wrong. In other words, the level of significance defines the likelihood of rejecting a null hypothesis when it is true, i.e. it is *the risk a decision-maker takes of rejecting the null hypothesis when it is really true*. The guide provided by the staistical theory is that this probability must be 'small'. Traditionally $\alpha = 0.05$ is selected for consumer research projects, $\alpha = 0.01$ for quality assurance and $\alpha = 0.10$ for political polling.

Step 3: Establish Critical or Rejection Region

The area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions(areas) as shown in Fig. 10.1. These regions are called the *acceptance region* and the *rejection* (or *critical*) *region*.

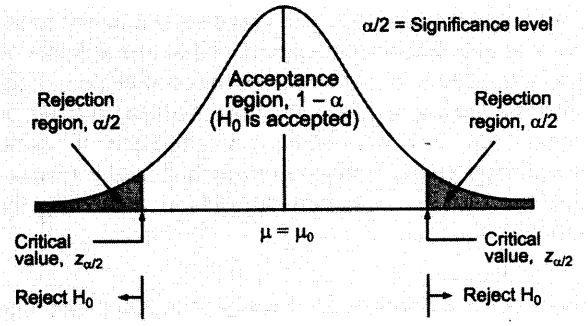


Figure 10.1
Areas of Acceptance and Rejection of H_0 (Two-Tailed Test)

The acceptance region is a *range of values* of the sample statistic spread around the *null hypothesized population parameter*. If values of the sample statistic fall within the limits of acceptance region, the null hypothesis is accepted, otherwise it is rejected.

Rejection region: The range of values that will lead to the rejection of a null hypothesis.

The **rejection region** is the *range of sample statistic values* within which if values of the sample statistic falls (i.e. outside the limits of the acceptance region), then null hypothesis is rejected.

The value of the sample statistic that separates the regions of acceptance and rejection is called **critical value.**

Critical value: A table value with which a test statistic is compared to determine whether a null hypothesis should be rejected or not.

The size of the rejection region is directly related to the level of precision to make decisions about a population parameter. Decision rules concerning null hypothesis are as follows:

- If prob (H_0 is true) $\leq \alpha$, then reject H_0
- If prob (H_0 is true) $> \alpha$, then accept H_0

In other words, if probability of H_0 being true is less than or equal to the significance level, α then reject H_0, otherwise accept H_0, i.e. the *level of significance α is used as the cut-off point which separates the area of acceptance from the area of rejection.*

Step 4: Select the Suitable Test of Significance or Test Statistic

The tests of significance or test statistic are classified into two categories: *parametric and nonparametric tests*. Parametric tests are more powerful because their data are derived from interval and ratio measurements. Nonparametric tests are used to test hypotheses with nominal and ordinal data. Parametric techniques are the tests of choice provided certain assumptions are met. Assumptions for parametric tests are as follows:

(i) The selection of any element (or member) from the population should not affect the chance for any other to be included in the sample to be drawn from the population.

(ii) The samples should be drawn from normally distributed propulations.

(iii) Populations under study should have equal variances.

Nonparametric tests have few assumptions and do not specify normally distributed populations or homogeneity of variance.

Selection of a test. For choosing a particular test of significance following three factors are considered:

(a) Whether the test involves one sample, two samples, or *k* samples?

(b) Whether two or more samples used are independent or related?

(c) Is the measurement scale nominal, ordinal, interval, or ratio?

Further, it is also important to know: (i) sample size, (ii) the number of samples, and their size, (iii) whether data have been weighted. Such questions help in selecting an appropriate test statistic.

One- sample tests are used for single sample and to test the hypothesis that it comes from a specified population. The following questions need to be answered before using one sample tests:

- Is there a difference between observed frequencies and the expected frequencies based on a statistical theory?
- Is there a difference between observed and expected proportions?
- Is it reasonable to conclude that a sample is drawn from a population with some specified distribution (normal, Poisson, and so on)?
- Is there a significant difference betweeen some measures of central tendency and its population parameter?

The value of test statistic is calculated from the distribution of sample statistic by using the following formula

$$\text{Test statistic} = \frac{\text{Value of sample statistic} - \text{Value of hypothesized population parameter}}{\text{Standard error of the sample statistic}}$$

The choice of a probability distribution of a sample statistic is guided by the sample size n and the value of population standard deviation σ as shown in Table 10.1 and Fig 10.2.

Table 10.1: Choice of Probability Distribution

Sample Size n	*Population Standard Deviation σ*	
	Known	*Unknown*
• $n > 30$	Normal distribution	Normal distribution
• $n \leq 30$, population being assumed normal	Normal distribution	t-distribution

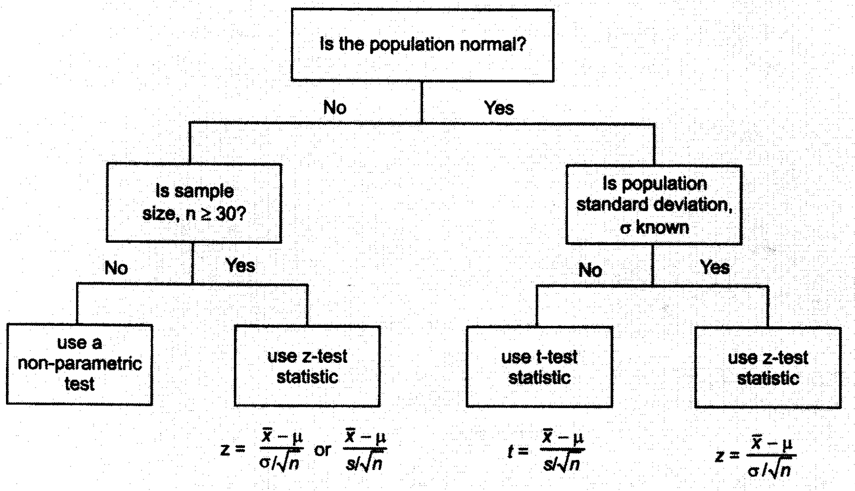


Figure 10.2
Choice of the Test Statistic

Often, when the null hypothesis is false, another alternative value of the population mean, μ is unknown. So for each of the possible values of the population mean μ, the probability of committing Type II error for several possible values of μ is required to be calculated.

Suppose a sample of size $n = 50$ is drawn from the given population to compute the probability of committing a Type II error for a specific alternative value of the population mean, μ. Let sample mean so obtained be $\bar{x} = 71$ with a standard deviation, $s = 21$. For significance level, $\alpha = 0.05$ and a two-tailed test, the table value of $z_{0.05} = \pm 1.96$. But the deserved value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

Since $z_{\text{cal}} = -3.03$ value falls in the rejection region, the null hypothesis H_0 is rejected. The rejection of null hypothesis, leads to either make a correct decision or commit a Type II error. If the population mean is actually 75 instead of 80, then the probability of commiting a Type II error is determined by computing a critical region for the mean $\bar{x}_c$. This value is used as the cutoff point between the area of acceptance and the area of rejection. If for any sample mean so obtained is less than (or greater than for right-tail rejection region), $\bar{x}_c$, then the null hypothesis is rejected. Solving for the critical value of mean gives

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } \pm 1.96 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

$$\bar{x}_c = 80 \pm 5.82 \text{ or } 74.18 \text{ to } 85.82$$

If $\mu = 75$, then probability of accepting the false null hypothesis $H_0 : \mu = 80$ when critical value is falling in the range $\bar{x}_c = 74.18$ to 85.82 is calculated as follows:

$$z_1 = \frac{74.18 - 75}{21/\sqrt{50}} = -0.276$$

The area under normal curve for $z_1 = -0.276$ is 0.1064.

$$z_2 = \frac{85.82 - 75}{21\sqrt{50}} = 3.643$$

The area under normal curve for $z_2 = 3.643$ is 0.4995

Thus the probability of committing a Type II error (β) falls in the region:

$$\beta = P(74.18 < \bar{x}_c < 85.82) = 0.1064 + 0.4995 = 0.6059$$

The total probability 0.6059 of committing a Type II error (β) is the area to the right of $\bar{x}_c = 74.18$ in the distribution. Hence the power of the test is $1 - \beta = 1 - 0.6059 = 0.3941$ as shown in Fig. 10.4(b).

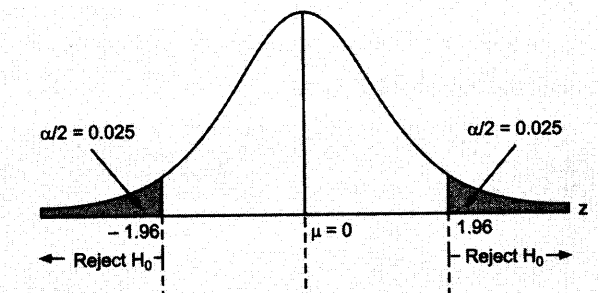


Figure 10.4 (a) Sampling distribution with $H_0 : \mu = 80$

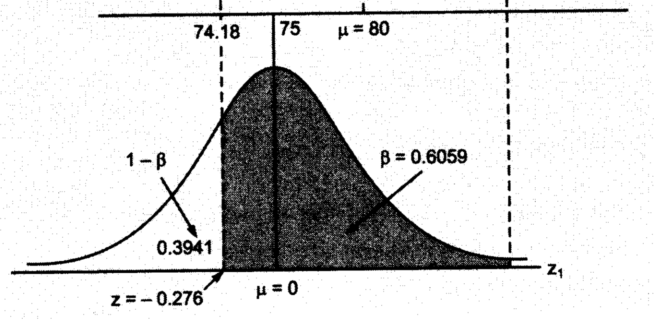


Figure 10.4 (b) Sampling distribution with $H_0 : \mu = 75$

To keep α or β low depends on which type of error is more costly. However, if both types of errors are costly, then to keep both α and β low, then inferences can be made more reliable by reducing the variability of observations. It is preferred to have large sample size and a low α value.

Few relations between two errors α and β, the power of a test $1 - \beta$, and the sample size n are stated below:

(i) If α (the sum of the two tail areas in the curve) is increased, the shadded area corresponding to β gets smaller, and vice versa.

(ii) The β value can be increased for a fixed α, by increasing the sample size n.

Special Case: Suppose hypotheses are defined as:

$$H_0 : \mu = 80 \text{ and } H_1 : \mu < 80$$

Given $n = 50$, $s = 21$ and $\bar{x} = 71$. For $\alpha = 0.05$ and left-tailed test, the table value $z_{0.05} = -1.645$. The observed z value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

The critical value of the sample mean $\bar{x}_c$ for a given population mean $\mu = 80$ is given by:

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \quad \text{or} \quad -1.645 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

$$\bar{x}_c = 75.115$$

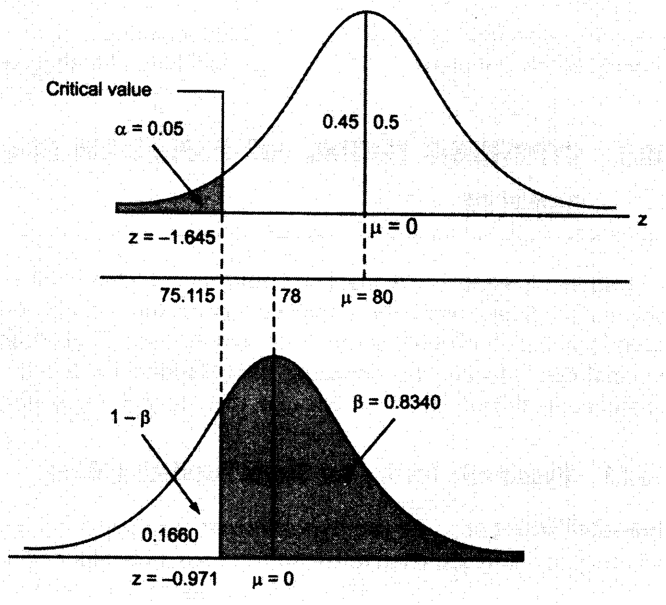


Figure 10.5 (a) Sampling distribution with $H_0 : \mu = 80$

Figure 10.5 (b) Sampling distribution with $H_0 : \mu = 78$

Fig. 10.5 (a) shows that the distribution of values that contains critical value of mean $\bar{x}_c = 75.115$ and below which H_0 will be rejected. Fig 10.5(b) shows the distribution of values when the alternative population mean value $\mu = 78$ is true. If H_0 is false, it is not possible to reject null hypothesis H_0 whenever sample mean is in the acceptance region, $\bar{x} \geq 75.151$. Thus critical value is computed by extending it and solved for the area to the right of $\bar{x}_c$ as follows:

$$z_1 = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} = \frac{75.115 - 78}{21/\sqrt{50}} = -0.971$$

This value of z yields an area of 0.3340 under the normal curve. Thus the probability $= 0.3340 + 0.5000 = 0.8340$ of committing a Type II error is all the area to right of $\bar{x}_c = 75.115$.

Remark In general, if alternative value of population mean μ is relatively more than its hypothesized value, then probability of committing a Type II error is smaller compared to the case when the alternative value is close to the hypothesized value. The probability of committing a Type II error decreases as alternative values are greater than the hypothesized mean of the population.

Conceptual Questions 10A

1. Discuss the difference in purpose between the estimation of parameters and the testing of statistical hypothesis.
2. Describe the various steps involved in testing of hypothesis. What is the role of standard error in testing of hypothesis? *[Delhi Univ., M.Com, 1999]*
3. What do you understand by null hypothesis and level of significance? Explain with the help of one example. *[HP Univ., MBA, 1996]*
4. What is a test statistic? How is it used in hypothesis testing?
5. Define the term 'level of significance'. How is it related to the probability of committing a Type I error?
   (a) Explain the general steps needed to carry out a test of any hypothesis.
   (b) Explain clearly the procedure of testing hypothesis. Also point out the assumptions in hypothesis testing in large samples. *[Kurukshetra Univ., MPWL, 1997]*
6. This is always a trade-off between Type I and Type II errors. Discuss. *[Delhi Univ., MCom, 1999]*
7. When should a one-tailed alternative hypothesis be used? Under what circumstances is each type of test used?
8. Explain the general procedure for determining a critical value needed to perform a test of a hypothesis.
9. What is meant by the terms hypothesis and a test of a hypothesis?
10. Define the standard error of a statistic. How it is helpful in testing of hypothesis and decision-making?
11. Define the terms 'decision rule' and 'critical value'. What is the relationship between these terms?
12. (a) How is power related to the probability of making a Type II error?
    (b) What is the power of a hypothesis test? Why is it important.
    (c) How can the power of a hypothesis test be increased without increasing the sample size?
13. Write short notes on the following:
    (a) Acceptance and rejection regions
    (b) Type I and Type II errors
    (c) Null and alternative hypotheses
    (d) One-tailed and two-tailed tests
14. When planning a hypothesis test, what should be done if probabilities of both Type I and Type II are to be small

10.7 HYPOTHESIS TESTING FOR POPULATION PARAMETERS WITH LARGE SAMPLES

Hypothesis testing involving large samples ($n>30$) is based on the assumption that the population from which the sample is drawn has a normal distribution. Consequently the sampling distribution of mean $\bar{x}$ is also normal. Even if the population does not have a normal distribution, the sampling distribution of mean $\bar{x}$ is assumed to be normal due to the central limit theorem because the sample size is large.

10.7.1 Hypothesis Testing for Single Population Mean

Two-tailed Test Let μ_0 be the hypothesized value of the population mean to be tested. For this the null and alternative hypotheses for two-tailed test are defined as:

$$H_0 : \mu = \mu_0 \quad \text{or} \quad \mu - \mu_0 = 0$$

and $$H_1 : \mu \neq \mu_0$$

If standard deviation σ of the population is known, then based on the central limit theorem, the sampling distribution of mean $\bar{x}$ would follow the standard normal distribution for a large sample size. The z-test statistic is given by

$$\text{Test-statistic: } z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In this formula, the numerator $\bar{x} - \mu$, measures how far (in an absolute sense) the observed sample mean $\bar{x}$ is from the hypothesized mean μ. The denominator $\sigma_{\bar{x}}$ is the *standard error of the mean*, so the z-test statistic represents how many standard errors $\bar{x}$ is from μ.

If the population standard deviation σ is not known, then a sample standard deviation s is used to estimate σ. The value of the z-test statistic is given by

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

The two rejection areas in two-tailed test are determined so that half the level of significance, $\alpha/2$ appears in each tail of the distribution of mean. Hence $z_{a/2}$ represents the standardized normal variate corresponding $\alpha/2$ in both the tails of normal curve as shown in Fig 10.1. The decision rule based on sample mean for the two-tailed test takes the form:

- Reject H_0 if $z_{cal} \leq -z_{\alpha/2}$ or $z_{cal} \geq z_{\alpha/2}$
- Accept H_0 if $-z_{\alpha/2} < z < z_{\alpha/2}$

where $z_{\alpha/2}$ is the table value (also called CV, critical value) of z at a chosen level of significance α.

Left-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a left-tailed test is of the form

$$H_0 : \mu \geq \mu_0 \quad \text{and} \quad H_1 : \mu < \mu_0$$

Test statistic: $z = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Decision rule:
- Reject H_0 if $z_{cal} \leq -z_\alpha$ (Table value of z at α)
- Accept H_0 if $z_{\alpha/2}\, z > -z_\alpha$

Right-tailed Test Large sample ($n > 30$) hypothesis testing about a population mean for a right-tailed test is of the form

$$H_0 : \mu \leq \mu_0 \text{ and } H_1 : \mu > \mu_0 \text{ (Right-tailed test)}$$

Test statistic: $z = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Decision rule:
- Reject H_0 if $z_{cal} \geq z_\alpha$ (Table value of z at α)
- Accept H_0 if $z_{cal} < z_\alpha$

10.7.2 Relationship between Interval Estimation and Hypothesis Testing

Consider following statements of null and alternative hypothesis:

- $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ (Two-tailed test)
- $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (Right-tailed test)
- $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (Left-tailed test)

The following are the confidence intervals in all above three cases where hypothesized value μ_0 of population mean, μ is likely to fall. Accordingly, the decision to accept or reject the null hypothesis will be taken.

Two-tailed test Two critical values CV_1 and CV_2 one for each tail of the sampling distribution is computed as follows:

Two-tailed test: The test of a null hypothesis which can be rejected when the sample statistic is in either extreme end of the sampling distribution.

(a) Known σ

Normal population	: Any sample size, n
Any population	: Large sample size n
$CV_1 = \mu_0 - z_{\alpha/2}\,\sigma_{\bar{x}}$	
$CV_2 = \mu_0 + z_{\alpha/2}\,\sigma_{\bar{x}}$	

where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

(b) Unknown σ

Any population	: Large sample size, n
$CV_1 = \mu_0 - z_{\alpha/2}\,s_{\bar{x}}$	
$CV_2 = \mu_0 + z_{\alpha/2}\,s_{\bar{x}}$	

$s_{\bar{x}} = s/\sqrt{n}$

Decision rule:
- Reject H_0 when $\bar{x} \leq CV_1$ or $\bar{x} \geq CV_2$.
- Accept H_0 when $CV_1 < \bar{x} < CV_2$

Left-tailed test The critical value for left tail of the sampling distribution is computed as follows:

(a) Known σ		(b) Unknown σ	
Normal population :	Any sample size, n	Any population :	Large sample size, n
Any population :	Large sample size, n		
$CV = \mu_0 - z_\alpha\, \sigma_{\bar{x}}$		$CV = \mu_0 - z_\alpha s_{\bar{x}}$	

Decision rule:
- Reject H_0 when $\bar{x} \leq CV$
- Accept H_0 when $\bar{x} > CV$

Right-tailed test The critical value for right tail of the sampling distribution is computed as follows:

(a) Known σ		(b) Unknown σ	
Normal population :	Any sample size, n	Any population :	Large sample size, n
Any population :	Large sample size, n		
$CV = \mu_0 + z_\alpha \sigma_{\bar{x}}$		$CV = \mu_0 + z_\alpha s_{\bar{x}}$	

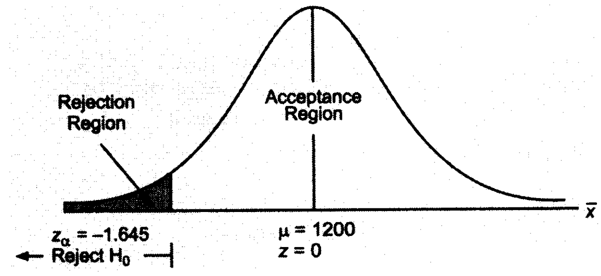
Decision rule :
- Reject H_0 when $\bar{x} \geq CV$
- Accept H_0 when $\bar{x} < CV$

Example 10.1: Individual filing of income tax returns prior to 30 June had an average refund of Rs 1200. Consider the population of 'last minute' filers who file their returns during the last week of June. For a random sample of 400 individuals who filed a return between 25 and 30 June, the sample mean refund was Rs 1054 and the sample standard deviation was Rs 1600. Using 5 per cent level of significance, test the belief that the individuals who wait until the last week of June to file their returns to get a higher refund than early the filers.

Solution: Since population standard deviation is not given, the standard error must be estimated with $s_{\bar{x}}$. Let us take the null hypothesis H_0 that the individuals who wait until the last week of June to file their returns get a higher return than the early filers, that is,

$$H_0 : \mu \geq 1200 \quad \text{and} \quad H_1 : \mu < 1200 \quad \text{(Left-tailed test)}$$

Figure 10.6



Given, $n = 400$, $s = 1600$, $\bar{x} = 1054$, $\alpha = 5\%$. Thus using the z-test statistic

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1054 - 1200}{1600/\sqrt{400}} = -\frac{146}{80} = -1.825$$

Since the calculated value $z_{cal} = -1.825$ is less than its critical value $z_\alpha = -1.645$ at $\alpha = 0.05$ level of significance, the null hypothesis, H_0 is rejected, as shown in Fig. 10.6. Hence, we conclude that individuals who wait until the last week of June are likely to receive a refund of less than Rs 1200.

Alternative approach: $CV = \mu_0 - z_\alpha \sigma_{\bar{x}} = 1200 - 1.645 \times (1600/\sqrt{400})$
$= 1200 - 131.6 = 1068.4$

Since $\bar{x}(= 1054) < CV(= 1068.4)$, the null hypothesis H_0 is rejected