# 12 ESTIMATION

## 12.1 STATISTICAL INFERENCE

*Statistical inference* is a field concerned with drawing conclusions about distributions by using observed values of random variables which are governed by these distributions.

Statistical inferences are the conclusions made about the unknown value of the parameter of a population using a limited information contained in an observed sample taken from it at random. The two most important types of statistical inferences are

(*i*) Estimation of parameters

(*ii*) Testing of hypotheses

## 12.2 STATISTICAL ESTIMATION

The *statistical estimation* is a procedure of making judgment about the unknown value of a population parameter by using the sample observations.

Population parameters are estimated from sample data because it is impracticable to examine the entire population in order to make such an exact determination. Statistical estimation procedures provide estimates of population parameters with a desired degree of confidence. This degree of confidence can be controlled, in part, by the size of the sample (the larger the sample, the greater the accuracy of the estimate) and by the type of the estimate made. The statistical estimation of population parameters is further divided into two types

(*i*) Point estimation

(*ii*) Interval estimation

## 12.3 POINT ESTIMATION OF A PARAMETER

The object of *point estimation* is to obtain a single number from the sample that is intended for estimating the unknown true value of a population parameter.

**12.3.1 Point Estimator.** A *point estimator* is a sample statistic that is used to estimate the unknown true value of a population parameter.

An estimator is always a statistic which is both a function and random variable with a probability distribution. An estimator is denoted by a capital letter ( *e. g.*, $T$, $U$, $\cdots$ ).

Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from a population with probability mass function or probability density function $f(x\,;\,\theta)$, then the estimator $T$ intended to estimate $\theta$ is a function given by

$$T = g(X_1, X_2, \cdots, X_n)$$

**12.3.2 Point Estimate.** A *point estimate* is a specific value of an estimator computed from the sample data after the sample has been observed. When a random sample becomes available from

the population and the estimator $T$ is computed from the sample data, the numerical value obtained is an estimate of population parameter $\theta$ from the particular sample. An estimate is denoted by a small letter ( *e. g.*, $t$, $u$, $\cdots$ ).

Suppose that $x_1, x_2, \cdots, x_n$ is an observed random sample from a population with probability mass function or probability density function $f(x; \theta)$, then the particular value of an estimator $T$ intended to estimate $\theta$ is a given by

$$t = g(x_1, x_2, \cdots, x_n)$$

***Example* 12.1** *A random sample selected from a normal population with mean $\mu$ and variance $\sigma^2$ gave the values* 25, 31, 23, 33, 28, 36, 22, 26. *Give the point estimators for $\mu$ and $\sigma^2$ and find their point estimates.*

**Solution.** We have

| $x_i$ | 25 | 31 | 23 | 33 | 28 | 36 | 22 | 26 | $\Sigma x_i = 224$ |
|---|---|---|---|---|---|---|---|---|---|
| $x_i - \bar{x}$ | $-3$ | 3 | $-5$ | 5 | 0 | 8 | $-6$ | $-2$ | |
| $(x_i - \bar{x})^2$ | 9 | 9 | 25 | 25 | 0 | 64 | 36 | 4 | $\Sigma(x_i - \bar{x})^2 = 172$ |

The point estimator of population mean $\mu$ is $\quad \bar{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

The point estimate of population mean $\mu$ is $\quad \bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{224}{8} = 28$

The point estimators of population variance $\sigma^2$ are

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \qquad \hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The point estimates of population variance $\sigma^2$ are

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{172}{8} = 21.5, \qquad \hat{s}^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{172}{8-1} = 24.57$$

## 12.4 UNBIASEDNESS

The distribution of an estimator should be centred in some sense at the value of the parameter to be estimated. Because expected value is a measure of the centre of a distribution, a reasonable requirement for an estimator $T$ may be $E(T) = \theta$. This property is called as unbiasedness of the estimator $T$. It refers to the desirability of the sampling distribution of an estimator being centred at the parameter to be estimated.

**12.4.1 Unbiased Estimator.** An estimator is *unbiased* if the mean of its sampling distribution is equal to the population parameter to be estimated.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution $f(x; \theta)$. An estimator $T = g(X_1, X_2, \cdots, X_n)$ is said to be *unbiased* for parameter $\theta$ if

$$E(T) = \theta$$

**12.4.2 Biased Estimator.** An estimator $T$ of a population parameter $\theta$ is said to be biased if:

$$E(T) \neq \theta$$

**12.4.3 Bias.** If an estimator $T$ of a population parameter $\theta$ is biased, the amount of its bias is

$$Bias = E(T) - \theta$$

If $T$ is an unbiased estimator, it will tend to give estimates nearer to $\theta$ and if $T$ is a biased estimator, it will tend to give estimates far from $\theta$.

***Example* 12.2** *A population consists of five numbers* 2, 4, 6, 8, *and* 10. *Consider all possible samples of size* 2 *which can be drawn with replacement from this population. By forming the sampling distributions, show that*

(*i*) *The sample variance* $S^2 = \sum(X_i - \bar{X})^2/n$ *is a biased estimator of the population variance* $\sigma^2$.

(*ii*) *The sample variance* $\hat{S}^2 = \sum(X_i - \bar{X})^2/(n-1)$ *is an unbiased estimator of the population variance* $\sigma^2$.

**Solution.** Population: 2, 4, 6, 8, 10 Population size: $N = 5$ Sample size : $n = 2$

The mean and variance of the population are

| $x_j$ | 2 | 4 | 6 | 8 | 10 | $\sum x_j = 30$ |
|---|---|---|---|---|---|---|
| $x_j^2$ | 4 | 16 | 36 | 64 | 100 | $\sum x_j^2 = 220$ |

$$\mu = \frac{\sum x_j}{N} = \frac{30}{5} = 6$$

$$\sigma^2 = \frac{\sum x_j^2}{N} - \mu^2 = \frac{220}{5} - (6)^2 = 8$$

Number of possible samples $= N \times N = 5 \times 5 = 25$

All possible samples:

| | | | | |
|---|---|---|---|---|
| (2, 2) | (2, 4) | (2, 6) | (2, 8) | (2, 10) |
| (4, 2) | (4, 4) | (4, 6) | (4, 8) | (4, 10) |
| (6, 2) | (6, 4) | (6, 6) | (6, 8) | (6, 10) |
| (8, 2) | (8, 4) | (8, 6) | (8, 8) | (8, 10) |
| (10, 2) | (10, 4) | (10, 6) | (10, 8) | (10, 10) |

(*i*) All possible sample variances: $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n} = \dfrac{(x_1 - x_2)^2}{4}$ when $n = 2$

| | | | | |
|---|---|---|---|---|
| 0 | 1 | 4 | 9 | 16 |
| 1 | 0 | 1 | 4 | 9 |
| 4 | 1 | 0 | 1 | 4 |
| 9 | 4 | 1 | 0 | 1 |
| 16 | 9 | 4 | 1 | 0 |

The sampling distribution $S^2$ and its mean are

| Value of $S^2$ $s^2$ | Number of occurrences $f$ | Probability $p(s^2) = f/\Sigma f$ | $s^2\ p(s^2)$ |
|---|---|---|---|
| 0 | 5 | 5/ 25 | 0 |
| 1 | 8 | 8/ 25 | 8/ 25 |
| 4 | 6 | 6/ 25 | 24/ 25 |
| 9 | 4 | 4/ 25 | 36/ 25 |
| 16 | 2 | 2/ 25 | 32/ 25 |
| | $\Sigma f = 25$ | 1 | $\Sigma s^2 p(s^2) = 100/25$ |

$$E(S^2) = \Sigma s^2\, p(s^2) = \frac{100}{25} = 4$$

Since $4 = E(S^2) \neq \sigma^2 = 8$, therefore $S^2$ is a biased estimator of $\sigma^2$.

*(ii)* All possible sample variances: $\hat{s}^2 = \dfrac{\Sigma(x_i - \bar{x})^2}{n-1} = \dfrac{(x_1 - x_2)^2}{2}$ when $n = 2$

| | | | | |
|---|---|---|---|---|
| 0 | 2 | 8 | 18 | 32 |
| 2 | 0 | 2 | 8 | 18 |
| 8 | 2 | 0 | 2 | 8 |
| 18 | 8 | 2 | 0 | 2 |
| 32 | 18 | 8 | 2 | 0 |

The sampling distribution of $\hat{S}^2$ and its mean are

| Value of $\hat{S}^2$ $\hat{s}^2$ | Number of occurrences $f$ | Probability $p(\hat{s}^2) = f/\Sigma f$ | $\hat{s}^2\ p(\hat{s}^2)$ |
|---|---|---|---|
| 0 | 5 | 5/ 25 | 0 |
| 2 | 8 | 8/ 25 | 16/ 25 |
| 8 | 6 | 6/ 25 | 48/ 25 |
| 18 | 4 | 4/ 25 | 72/ 25 |
| 32 | 2 | 2/ 25 | 64/ 25 |
| | $\Sigma f = 25$ | 1 | $\Sigma \hat{s}^2 p(\hat{s}^2) = 200/25$ |

$$E(\hat{S}^2) = \Sigma \hat{s}^2\, p(\hat{s}^2) = \frac{200}{25} = 8$$

Since $8 = E(\hat{S}^2) = \sigma^2 = 8$, therefore $\hat{S}^2$ is an unbiased estimator of $\sigma^2$.

## 12.5 BEST ESTIMATOR

Let $X_1, X_2, \cdots, X_n$ be a random sample of size $n$ from the distribution $f(x; \theta)$. Among the class $U$ of all unbiased estimators $T = g(X_1, X_2, \cdots, X_n)$ for a given parameter $\theta$, the estimator $T^*$ is said to be a *best* or *minimum variance* estimator if among the class $U$ of all unbiased estimators, none has a smaller variance than $T^*$.

**12.5.1 Best Estimators of the Population Mean and Variance.** Let $X_1, X_2, \cdots, X_n$ be a random sample of size $n$ from a population with unknown mean $\mu$ and unknown variance $\sigma^2$, then the best estimators of $\mu$ and $\sigma^2$ are

$$\bar{X} = \frac{\Sigma X_i}{n} \qquad \text{and} \qquad \hat{S}^2 = \frac{\Sigma (X_i - \bar{X})^2}{n - 1}$$

respectively.

***Example* 12.3** *Obtain the best unbiased estimates of the population mean $\mu$ and variance $\sigma^2$ from which the following sample is drawn:*

$$n = 8, \qquad \Sigma x_i = 120, \qquad \Sigma (x_i - \bar{x})^2 = 302$$

**Solution.** The best estimate of the population mean $\mu$ is the sample mean

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{120}{8} = 15$$

The best estimate of the population variance $\sigma^2$ is the sample variance

$$\hat{s}^2 = \frac{\Sigma (x_i - \bar{x})^2}{n - 1} = \frac{302}{8 - 1} = 43.14$$

**12.5.2 Best estimator of the population proportion.** From a population which has unknown proportion of successes $\pi$, we take a random sample of size $n$ with $X$ as the number of successes in the sample, then the best estimator of $\pi$ is

$$P = \frac{X}{n}$$

***Example* 12.4** *A random sample of 50 children from a large school is chosen and the number who are left handed is noted. It is found that 6 are left handed. Obtain an unbiased estimate of the proportion of children in the school who are left handed.*

**Solution.** We have $n = 50 \qquad x = 6$

Sample proportion: $p = \frac{x}{n} = \frac{6}{50} = 0.12$

## 12.6 POOLED ESTIMATORS FROM TWO SAMPLES

Estimates of the population mean, variance, proportion, *etc.*, may be obtained by pooling observations from two random samples.

**12.6.1 Pooled Estimator of Population Mean.** Let $X_{11}, X_{21}, \cdots, X_{n_1 1}$ and $X_{12}, X_{22}, \cdots, X_{n_2 2}$ be two random samples of sizes $n_1$ and $n_2$ from a population with unknown mean $\mu$, then the pooled estimator $\overline{X}_p$ of $\mu$ is

$$\overline{X}_p = \frac{\sum_{i=1}^{n_1} X_{i1} + \sum_{i=1}^{n_2} X_{i2}}{n_1 + n_2} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the unbiased estimators of $\mu$, based on the first and the second sample, respectively.

**12.6.2 Pooled Estimator of Population Variance.** Let $X_{11}, X_{21}, \cdots, X_{n_1 1}$ and $X_{12}, X_{22}, \cdots, X_{n_2 2}$ be two random samples of sizes $n_1$ and $n_2$ from a population with unknown variance $\sigma^2$, then the pooled estimator $S_p^2$ of $\sigma^2$ is

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{i1} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(X_{i2} - \overline{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\,\hat{S}_1^2 + (n_2 - 1)\,\hat{S}_2^2}{n_1 + n_2 - 2}$$

where $\hat{S}_1^2$ and $\hat{S}_2^2$ are the unbiased estimators of $\sigma^2$, based on the first and the second sample, respectively.

***Example* 12.5** *Two samples of sizes* 40 *and* 50, *respectively, are taken from a population with unknown mean* $\mu$ *and unknown variance* $\sigma^2$.

Sample I: $n_1 = 40, \quad \Sigma f x_1 = 807, \quad \Sigma f x_1^2 = 16329$

Sample II: $n_2 = 50, \quad \Sigma f x_2 = 977, \quad \Sigma f x_2^2 = 19177$

*Using the data from the two samples, obtain the best estimates of* $\mu$ *and* $\sigma^2$.

**Solution.** The best estimates of $\mu$ and $\sigma^2$ are

$$\bar{x}_p = \frac{\Sigma f x_1 + \Sigma f x_2}{n_1 + n_2} = \frac{807 + 977}{40 + 50} = 19.82$$

$$\Sigma f(x_1 - \bar{x}_1)^2 = \Sigma f x_1^2 - \frac{(\Sigma f x_1)^2}{n_1} = 16329 - \frac{(807)^2}{40} = 47.775$$

$$\Sigma f(x_2 - \bar{x}_2)^2 = \Sigma f x_2^2 - \frac{(\Sigma f x_2)^2}{n_2} = 19177 - \frac{(977)^2}{50} = 86.42$$

$$s_p^2 = \frac{\Sigma f(x_1 - \bar{x}_1)^2 + \Sigma f(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{47.775 + 86.42}{40 + 50 - 2} = 1.525$$

**12.6.3 Pooled Estimator of Population Proportion.** From a population which has unknown proportion of successes $\pi$, we take two random samples of sizes $n_1$ and $n_2$ with $X_1$ and $X_2$ as the number of successes in the respective sample, then the pooled estimator $\hat{\pi}$ of $\pi$ is

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

where $P_1$ and $P_2$ are the unbiased estimators of $\pi$, based on the first and second sample, respectively.

***Example* 12.6** *A random sample of* 600 *people from a certain district were questioned and the results indicated that* 30% *used a particular product. In a second random sample of* 300 *people,* 96 *used the product. Using the data from the two samples, find the best estimate of the proportion of people in the district who used the product.*

**Solution.** The best estimate of the population proportion is

$$n_1 = 600 \qquad p_1 = 0.30$$

$$n_2 = 300 \qquad x_2 = 96 \qquad \Rightarrow \qquad p_2 = \frac{x_2}{n_2} = \frac{96}{300} = 0.32$$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{600(0.30) + 300(0.32)}{600 + 300} = 0.307$$

## Exercise 12.1

**1.** (*a*) Explain what is meant by statistical inference?

(*b*) What is meant by estimation? Differentiate between estimator and estimate?

**2.** (*a*) Specify the estimator and the estimate in each of the following:

(*i*) A sample of 35 students gave an average height of 62 inches.

(*ii*) A sample of 50 households having television sets showed that 85 percent of them liked a particular programme.

(*iii*) A sample of 25 bolts produced by a company showed that 20 of them were according to specifications.

(*iv*) A sample of 30 houses showed an average consumption of electricity as 65 units.

{ (*i*) Sample mean height $\overline{X}$; $\bar{x} = 62$ inches.
(*ii*) Sample proportion of households who liked particular programme $P$; $p = 0.85$.
(*iii*) Sample number of bolts according to specification $X$; $x = 20$, or sample proportion of bolts according to specification $P$; $p = 0.80$.
(*iv*) Sample mean consumption of electricity $\overline{X}$ ; $\bar{x} = 65$ units }

(*b*) Suppose I choose a random sample of three observations from a population and obtain the values 2, 5, 3. From these values I estimate the centre of the population by ranking the observations and taking the middle one. What estimator am I using and what is my estimate?

{ Sample median $X_{0.5} = X_{((n+1)/2)}$; $x_{0.5} = x_{((n+1)/2)} = 3$ }

**3.** (*a*) Is an estimator a random variable? Why or why not?
( Yes. An estimator is a random variable having its own probability distribution. )

(*b*) Why we call the standard deviation of a sample statistic as standard error of the statistic.
{ In the context of estimation, the deviation of a sample statistic $T$ from its target $\theta$ ( parameter to be estimated ) must be considered an error. So the standard deviation of a sample statistic is commonly called the standard error of the sample statistic. }

**4.** (*a*) What is meant by unbiasedness? Differentiate between an unbiased and a biased estimator.

(*b*) A finite population consists of the numbers 3, 5, 7 and 9. Take all possible samples of size 2 which can be drawn with replacement form this population. By forming the sampling distributions of $\overline{X}$ and $S^2$ show that

(*i*) the sample mean $\overline{X} = \sum X_i / n$ is an unbiased estimator of the population mean $\mu$.

(*ii*) the sample variance $S^2 = \sum (X_i - \overline{X})^2 / n$ is a biased estimator of the population variance $\sigma^2$.

{ (*i*) $\mu = 6$, $E(\overline{X}) = 6$, $E(\overline{X}) = \mu$, (*ii*) $\sigma^2 = 5$, $E(S^2) = 2.5$, $E(S^2) \neq \sigma^2$ }

(*c*) Draw all possible samples of size 3 taken without replacement from the population 7, 10, 13 and 16. By forming the sampling distributions, show that both sample mean $\overline{X}$ and sample median $X_{0.5}$ are unbiased estimators.

**5.** (*a*) A finite population consists of the numbers 1, 3, 5, 7 and 9. Consider all possible samples of size two which can be drawn with replacement form this population. By forming the sampling distributions of $\overline{X}$, $S^2$ and $\hat{S}^2$ show that

(*i*) the sample mean $\overline{X} = \sum X_i / n$ is an unbiased estimator of the population mean $\mu$.

(*ii*) the sample variance $S^2 = \sum (X_i - \overline{X})^2 / n$ is a biased estimator of the population variance $\sigma^2$.

(*iii*) the sample variance $\hat{S}^2 = \sum (X_i - \overline{X})^2 / (n-1)$ is an unbiased estimator of the population variance $\sigma^2$.

{ (*i*) $\mu = 5$, $E(\overline{X}) = 5$, $E(\overline{X}) = \mu$, (*ii*) $\sigma^2 = 8$, $E(S^2) = 4$, $E(S^2) \neq \sigma^2$
(*iii*) $\sigma^2 = 8$, $E(\hat{S}^2) = 8$, $E(\hat{S}^2) = \sigma^2$ }

(*b*) A finite population consists of the numbers 2, 3, 4, 5, 6 and 8. Find the proportion $P$ of even numbers in all possible random samples of size $n = 2$ that can be drawn with replacement from this population. By forming the sampling distribution of sample proportions show that sample proportion is an unbiased estimator of the population proportion. Also verify the relation

$$Var(P) = \frac{\pi(1-\pi)}{n}$$

where $P$ and $\pi$ are sample and population proportions respectively.

$\{ \pi = 2/3, \mu_P = 2/3, \sigma_P^2 = 1/9 \}$

(*c*) A finite population consists of the numbers 4, 5, 6 and 8. Find the proportion $P$ of even numbers in all possible random samples of size $n = 3$ that can be drawn without replacement from this population. By forming the sampling distribution of sample proportions show that sample proportion is an unbiased estimator of the population proportion. Also verify the relation

$$Var(P) = \frac{\pi(1-\pi)}{n}\left(\frac{N-n}{N-1}\right)$$

where $P$ and $\pi$ are sample and population proportions respectively.

$\{ \pi = 3/4, \mu_P = 3/4, \sigma_P^2 = 1/48 \}$

◆

## 12.7 INTERVAL ESTIMATION

*Interval estimation* is a procedure of constructing an interval from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter.

**12.7.1 Need for Interval Estimation.** Any point estimate has the limitation that it does not provide information about the precision of the estimate *i. e.*, about the magnitude of error due to sampling. Often such information is essential for proper interpretation of the sample result.

A point estimator, calculated from the sample data, provides a single number as an estimate of the parameter. This single number lies in the fore front even though a statement of accuracy in terms of the standard error is attached to it. A point estimator, however efficient it may be, cannot be expected to be exactly equal to the population parameter. Moreover, we cannot assess simply by looking at just only one value ( point estimator ) how close is the estimate to the unknown true value of the parameter being estimated. A point estimate by itself does not supply this information about its precision.

An alternative approach to estimation is to extend the concept of error bound to produce an interval of values that is likely to include the unknown true value of the parameter. This is the concept underlying estimation by confidence intervals.

**12.7.2 Interval Estimate.** An *interval estimate* is an interval calculated from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population with unknown parameter $\theta$. A confidence interval for $\theta$ is an interval $( L , U )$ computed from the sample observations $X_1, X_2, \cdots, X_n$, such that prior to sampling, it includes the unknown true value of $\theta$ with a specified high probability. Let $( 1 - \alpha )$ be a specified high probability and $L$ and $U$ be functions of sample observations $X_1, X_2, \cdots, X_n$ such that

$$P(L < \theta < U) = 1 - \alpha, \qquad \text{for } 0 < \alpha < 1$$

Then the interval $(L, U)$ is called a $100(1-\alpha)\%$ *confidence interval* for the parameter $\theta$, and the probability $(1-\alpha)$ is called the *confidence coefficient* or the *level of confidence*. Note that, $(1-\alpha)$ is the probability that the random interval $(L, U)$ includes the parameter $\theta$ and not the probability that $\theta$ lies in the interval $(L, U)$. The end points $L$ and $U$ that bound the confidence interval, are called the *lower* and *upper confidence limits* for the parameter $\theta$. These limits being the functions of sample observations are random variables. The width $U - L$ of the confidence interval measures the precision of the estimate. The shorter the confidence interval, the more precise the estimate will be. The precision can be increased by

(*i*) decreasing the standard error of the estimate (*i. e.*, increasing the sample size).

(*ii*) decreasing the confidence coefficient.

### 12.7.3 Confidence Coefficient.

***Meaning of Confidence Coefficient.*** From the definition of a confidence interval, we know that, prior to selecting the random sample, the probability is $1 - \alpha$ that the confidence interval we obtain will include the population parameter $\theta$. The particular confidence interval result will be either correct or incorrect, and we do not know for certain which is the case.

***Selecting the Confidence Coefficient.*** We should like the confidence interval to be very precise (*i. e.*, very narrow) and would like to be very confident that it includes $\theta$. Unfortunately, for any fixed sample size, the confidence coefficient can only be increased by increasing the width of the confidence interval. The confidence interval widens rapidly as the confidence coefficient gets near 100 percent.

The choice of $1 - \alpha$ will vary from case to case, depending on how much risk of obtaining an incorrect interval can be taken. The numerical confidence coefficient (*e. g.*, 0.95) is often expressed as a percent (*e. g.*, 95%). Confidence coefficients of 90, 95, 98, and 99 percent are often used in practice.

## 12.8 CONFIDENCE INTERVAL FOR POPULATION MEAN, $\mu$

The interval $(L, U)$ is a $100(1-\alpha)\%$ confidence interval for the population mean $\mu$ if prior to sampling:

$$P(L < \mu < U) = 1 - \alpha$$

This definition simply states that a confidence interval with confidence coefficient $1 - \alpha$ is an interval estimate such that the probability is $1 - \alpha$ that the calculated limits include $\mu$ for any random sample. In other words, in many repeated random samples of size $n$ from a population, $100(1-\alpha)\%$ of the interval estimates will include $\mu$ and therefore will be correct and $\alpha\%$ of the interval estimates will not include $\mu$ and therefore will be incorrect. The choice of method used in constructing a confidence interval for $\mu$ depends upon whether or not the population is normal, whether the population variance $\sigma^2$ is known or unknown, and whether the sample size $n$ is large or small. We discuss these different cases below.

**12.8.1 Normal Population, $\sigma^2$ known.** Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size $n$ is drawn from a normal population with unknown mean $\mu$ and known variance $\sigma^2$. We wish to construct a confidence interval which is likely to include the true unknown value of the population mean $\mu$ with a degree of confidence $1 - \alpha$. We know that the sampling distribution

of $\bar{X}$ will be normal with mean $\mu$ and variance $\sigma^2/n$. Consequently, the distribution of the statistic

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

will be normal with mean 0 and variance 1. Then a two-sided $100(1 - \alpha)\%$ confidence interval for population mean $\mu$ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If $\bar{x}$ is the mean of an observed random sample of size $n$ taken from a normal population with unknown mean $\mu$ and known variance $\sigma^2$, then a $100(1 - \alpha)\%$ confidence interval for $\mu$ is given by

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This can be written $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

Note that, often the word *central* is omitted when considering confidence intervals, but it is assumed that a two-sided interval that is central, or symmetric about the mean is required.

**12.8.2 Interpretation of a Confidence Interval.** A $100(1 - \alpha)\%$ confidence interval for $\mu$ is

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If we identify

$$L = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

the probability statement implies that prior to sampling, the random interval $(L, U)$ will include the parameter $\mu$ with a probability $1 - \alpha$. That is

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

It is to be emphasized that in this expression, $\mu$ is constant and it is the end points which are random variables. To better understand the meaning of a confidence statement, we perform repeated samplings from a normal distribution with mean $\mu$ and standard deviation $\sigma$ and a $100(1 - \alpha)\%$ confidence interval $\bar{x} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ is computed from each random sample, approximately $100(1 - \alpha)\%$ of the intervals derived would contain the true value of $\mu$.

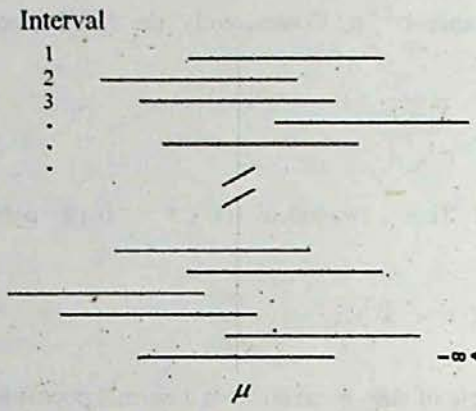


**Fig. 12.1** Repeated forming confidence intervals for $\mu$

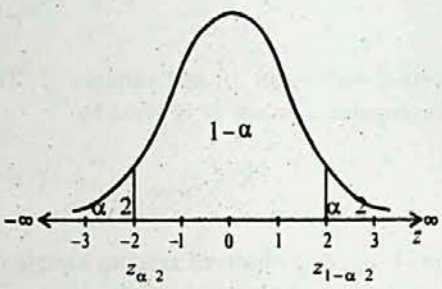


**Fig. 12.2** Two-sided confidence interval for $\mu$

Figure 12.1 shows what would typically happen if a number of samples were drawn from the same population and a confidence interval for $\mu$ were computed for each sample. The true value of $\mu$ is indicated by a vertical line in the figure. Different confidence intervals, resulting from different random samples, are shown as horizontal line segments. Most of the confidence intervals would contain $\mu$, but some of them would not contain $\mu$. If a 95% confidence interval were calculated for each sample then in the long-run, 95% of the confidence intervals that were formed would contain $\mu$. That is not surprising, because the specified probability 0.95 represents the long-run relative frequency of these intervals crossing the vertical line.

Figure 12.2 represents that, before the sample is taken, the probability is $1 - \alpha$ that the quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ will fall in the shaded interval. The interval estimate $\bar{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ will be correct ( *i. e.*, will include $\mu$ ) if $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ does fall in the shaded interval. In effect, the risk $\alpha$ of an incorrect confidence interval is divided equally in the two tails of the standard normal distribution.

**12.8.3 Steps to Follow When Forming a Confidence Interval.** We will follow the following standard format when estimating parameters with confidence intervals:

(*i*) Identify the population of interest, and state the conditions required for the validity of the procedure being used to construct the confidence interval.

(*ii*) Give the procedure ( formula ) that will be used

(*iii*) Construct the confidence interval

(*iv*) Interpret the results.

***Example* 12.7** *A normal population has a variance of* 100. *A random sample of size* 16 *selected from the population has a mean of* 52.5. *Construct the* 90% *confidence interval estimate of the population mean,* $\mu$. *Interpret the result.*

**Solution.** The size and mean of sample and the variance of normal population are

$$n = 16, \qquad \bar{x} = 52.5, \qquad \sigma^2 = 100 \quad \Rightarrow \quad \sigma = 10,$$

*Confidence coefficient*: $1 - \alpha = 0.90$

$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$

$z_{1-\alpha/2} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

The two-sided 90% confidence interval for $\mu$ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$52.5 - (1.645)\frac{10}{\sqrt{16}} < \mu < 52.5 + (1.645)\frac{10}{\sqrt{16}}$$

$$48.4 < \mu < 56.6$$

A two-sided 90% confidence interval for $\mu$ obtained from the observed sample is (48.4, 56.6). We are 90% confident that the interval estimate contains $\mu$.

***Example* 12.8** *Unoccupied seats on flights cause the airlines to lose revenue. Suppose a large airline obtained the 90% confidence interval for the average number of unoccupied seats per flight, on the basis of the records of its randomly selected 225 flights over the past year, as 11.15 to 12.05. Find the value of $\bar{x}$, the mean of the sample and $\sigma$ the standard deviation of the normal population from which the sample was drawn. Estimate the average number of unoccupied seats per flight over the past year with 99% confidence coefficient.*

**Solution.** Sample size $n = 225$

*Confidence coefficient*: $1 - \alpha = 0.90$

$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$

$z_{1-\alpha/2} = z_{0.95} = 1.645$ { From Table 10 (*b*) }

The two-sided $100(1 - \alpha)\%$ confidence interval for $\mu$ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The lower and upper limits of 90% confidence interval for $\mu$ are 11.15 and 12.05. Thus

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 11.15 \Rightarrow \bar{x} - (1.645)\frac{\sigma}{\sqrt{225}} = 11.15 \quad \ldots\ldots(i)$$

$$\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 12.05 \Rightarrow \bar{x} + (1.645)\frac{\sigma}{\sqrt{225}} = 12.05 \quad \ldots\ldots(ii)$$

Adding (*i*) and (*ii*), we get

$2\bar{x} = 23.20 \Rightarrow \bar{x} = 11.6$

Putting the value of $\bar{x}$ in (*ii*), we get

$$11.6 + (1.645)\frac{\sigma}{\sqrt{225}} = 12.05 \Rightarrow \sigma = 4.1$$

*Confidence coefficient*: $1 - \alpha = 0.90$

$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow 1 - \alpha/2 = 0.995$

$z_{1-\alpha/2} = z_{0.995} = 2.576$ { From Table 10 (*b*) }

The two-sided 90% confidence interval for $\mu$ is

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$11.6 - (2.576)\frac{4.1}{\sqrt{225}} < \mu < 11.6 + (2.576)\frac{4.1}{\sqrt{225}}$$

$$10.9 < \mu < 12.3$$

**12.8.4 Any Population, $\sigma^2$ known/unknown, $n$ large.** Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size $n$ is drawn from a population with mean $\mu$ and variance $\sigma^2$. We wish to construct a confidence interval which is likely to trap the true unknown value of the population mean $\mu$ with a degree of confidence $1 - \alpha$. If the population is not normal, and if $\sigma^2$ is either known or unknown, then according to the Central Limit Theorem the sampling distribution of $\bar{X}$ is approximately normal with mean $\mu$ and variance $\sigma^2/n$ (when $\sigma^2$ is known and $\hat{S}^2/n$ when $\sigma^2$ is unknown) if the sample size is sufficiently large, say, $n > 30$. Consequently the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cong \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}}$$

is approximately normal with mean 0 and variance 1. Then a two sided $100(1 - \alpha)\%$ approximate confidence interval for population mean $\mu$ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We now turn to the more realistic situation for which the population variance $\sigma^2$ is unknown. Because $n$ is large, replacing $\sigma$ with its best unbiased estimator $\hat{S}$ does not appreciably affect the probability statement. When $n$ is large and population variance $\sigma^2$ is unknown, a $100(1 - \alpha)\%$ approximate confidence interval for population mean $\mu$ is given by

$$\bar{X} - z_{1-\alpha/2}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\hat{S}}{\sqrt{n}}.$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size $n$ sufficiently large from a population with unknown mean $\mu$ and unknown but finite variance $\sigma^2$, then a $100(1 - \alpha)\%$ approximate confidence interval for $\mu$ is given by

$$\bar{x} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

This can be written $\bar{x} \pm z_{1-\alpha/2} \dfrac{\hat{s}}{\sqrt{n}}$

**12.8.5 Sampling Without Replacement.** When sampling is done without replacement from a finite population of size $N$, the standard error of $\bar{X}$ is given by

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

If the sample size $n$ is greater than 5% of the population size $N$ ( *i. e.*, $n > 0.05\,N$ ), then a $100\,(1-\alpha)\%$ confidence interval for $\mu$ is given by

$$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

However, when $n$ is large and population variance $\sigma^2$ is unknown, a $100\,(1-\alpha)\%$ approximate confidence interval for population mean $\mu$ is given by

$$\bar{X} \pm z_{1-\alpha/2} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

If $\bar{x}$ is the mean of an observed random sample of size $n$ taken from a population with unknown mean $\mu$ and known variance $\sigma^2$, then a $100\,(1-\alpha)\%$ confidence interval for $\mu$ is given by

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size $n$ sufficiently large from a population with unknown mean $\mu$ and unknown but finite variance $\sigma^2$, then a $100\,(1-\alpha)\%$ approximate confidence interval for $\mu$ is given by

$$\bar{x} \pm z_{1-\alpha/2} \frac{\hat{s}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

The finite population correction $(N-n)/(N-1)$ may be ignored when the sample size $n$ is less than 5% of the population size $N$ ( *i. e.*, $n < 0.05\,N$ ).

***Example* 12.9** *A particular component in a transistor circuit has a lifetime which is known to follow a skew distribution. A random sample of* 250 *components from a week's production given an average lifetime of* 840 *hours, and the variance of lifetimes is* 483 ( *hours*$^2$ ). *Find approximately* 95% *confidence limits to the true mean lifetime in the whole population of the product.*

**Solution.** The size, mean and variance of the sample are

$$n = 250, \quad \bar{x} = 840, \quad \hat{s}^2 = 483 \Rightarrow \hat{s} = \sqrt{483} = 21.98$$

*Confidence coefficient* : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 95% approximate confidence interval for $\mu$ is

$$\bar{x} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

$$840 - (1.960)\frac{21.98}{\sqrt{250}} < \mu < 840 + (1.960)\frac{21.98}{\sqrt{250}}$$

$$837.3 < \mu < 842.7$$

***Example* 12.10** *A random sample of size* $n = 200$, *selected without replacement from a finite population of size* $N = 1000$ *with* $\sigma = 1.28$, *showed that* $\bar{x} = 68.6$. *Construct a* 97% *confidence interval for the mean* $\mu$ *of the population.*

**Solution.** The size and mean of sample and the size and a standard deviation of population are

$$n = 200, \quad \bar{x} = 68.6, \quad N = 1000, \quad \sigma = 1.28$$

*Confidence coefficient*: $1 - \alpha = 0.97$

$$1 - \alpha = 0.97 \Rightarrow \alpha = 0.03 \Rightarrow \alpha/2 = 0.015 \Rightarrow 1 - \alpha/2 = 0.985$$

$$z_{1-\alpha/2} = z_{0.985} = 2.17 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 97% approximate confidence interval for $\mu$ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$68.6 \pm (2.17)\frac{1.28}{\sqrt{200}}\sqrt{\frac{1000-200}{1000-1}}$$

$$(68.42, 68.78) \qquad \Rightarrow \qquad 68.42 < \mu < 68.78$$

***Example* 12.11** *An auditor has selected a simple random sample of* 100 *accounts from the* 8042 *accounts receivable of a freight company to estimate the total audit amount of the receivable in the population. The sample mean is* $\bar{x} = 33.19$ *and the sample standard deviation is* $\hat{s} = 34.48$. *Obtain the* 95.44 *percent confidence interval for the mean audit amount in the population.*

**Solution.** The size, mean and standard deviation of the sample and the population size are

$$n = 100, \quad \bar{x} = 33.19, \quad \hat{s} = 34.48, \quad N = 8042$$

*Confidence coefficient*: $1 - \alpha = 0.9544$

$$1 - \alpha = 0.9544 \Rightarrow \alpha = 0.0456 \Rightarrow \alpha/2 = 0.0228 \Rightarrow 1 - \alpha/2 = 0.9772$$

$$z_{1-\alpha/2} = z_{0.9772} = 2 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 95.44% approximate confidence interval for the mean audit amount $\mu$ is

$$\bar{x} \pm z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$33.19 \pm (2)\frac{34.48}{\sqrt{100}}\sqrt{\frac{8042-100}{8042-1}}$$

$$(26.3366, 40.0434) \qquad \Rightarrow \qquad 26.3366 < \mu < 40.0434$$

**12.8.6 Normal Population, $\sigma^2$ unknown, $n$ small.** Suppose that a random sample $X_1, X_2, \cdots, X_n$ of size $n$ is drawn from a normal population with unknown mean $\mu$ and unknown variance $\sigma^2$. We wish to construct a confidence interval which is likely to contain the true unknown value of population mean $\mu$ with a confidence coefficient $1 - \alpha$. However, time and cost restrictions would probably limit the sample size to a small number. Many inferences in business must be made on the basis of very limited information, *i. e., small samples.* When the population is normal, the sampling distribution of the statistic

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

is a $t$-distribution with $\nu = n - 1$ degrees of freedom. Thus, when $n$ is small, population is normal and population variance is unknown, a two-sided $100(1 - \alpha)\%$ confidence interval for population mean $\mu$ is given by

$$\overline{X} - t_{\nu;1-\alpha/2}\frac{\hat{S}}{\sqrt{n}} < \mu < \overline{X} + t_{\nu;1-\alpha/2}\frac{\hat{S}}{\sqrt{n}}$$

If $\bar{x}$ and $\hat{s}$ is the mean and standard deviation of an observed random sample of size $n$ from a normal population with unknown mean $\mu$ and unknown but finite variance $\sigma^2$, then a $100(1 - \alpha)\%$ interval for $\mu$ is given by

$$\bar{x} - t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

This can be written $$\bar{x} \pm t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

For ***large degrees of freedom*** ( e.g., beyond the range of Table 12 ) the ***t-distribution*** can be approximated by a ***standard normal distribution.***

***Example* 12.12** *Ten packets of a particular brand of biscuits are chosen at random and their mass measured in grams. The results are*

$$n = 10, \qquad \sum x_i = 3978.7, \qquad \sum x_i^2 = 1583098.3$$

*Assuming that the sample is taken from a normal population with mean mass $\mu$, calculate the 98% confidence interval for $\mu$.*

**Solution.** The mean and standard deviation of the sample are

$$\bar{x} = \frac{\sum x}{n} = \frac{3978.7}{10} = 397.87$$

$$\hat{s} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{1583098.3 - 10(397.87)^2}{10-1}} = 3.213$$

*Confidence coefficient*: $1 - \alpha = 0.98$

$$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02 \Rightarrow \alpha/2 = 0.01 \Rightarrow 1 - \alpha/2 = 0.99$$

*Degrees of freedom* : $\nu = n - 1 = 10 - 1 = 9$

$t_{\nu;1-\alpha/2} = t_{9;0.99} = 2.821$ ( From Table 12 )

The two-sided 98% confidence interval for $\mu$ is

$$\bar{x} - t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}} < \mu < \bar{x} + t_{\nu;1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}$$

$$397.87 - (2.821)\frac{3.213}{\sqrt{10}} < \mu < 397.87 + (2.821)\frac{3.213}{\sqrt{10}}$$

$$395.004 < \mu < 400.736$$

## Exercise 12.2

**1.** (*a*) What is meant by estimation ? Distinguish between point estimate and interval estimate. Why is an interval estimate more useful ?

(*b*) Distinguish between the following

(*i*) Estimator and estimate,

(*ii*) Point and interval estimation..

**2.** (*a*) Explain what is meant by

(*i*) Confidence interval,

(*ii*) Confidence limits,

(*iii*) Confidence coefficient.

(*b*) When would a confidence interval be preferred over point estimation for a parameter. ( When the reliability of a point estimator is needed, the confidence interval conveniently express the estimator along with its measure of variation. Reliability is reported through the confidence coefficient and the variation is reflected in the length of the interval. )

**3.** (*a*) Find a 90% confidence interval for the mean of a normal distribution with $\sigma = 3$, given the sample as 2.3, −0.2, −0.4, −0.9.
( $-2.268 < \mu < 2.668$ )

(*b*) The standard deviation of the amounts poured into bottles by an automatic filling machine is 1.8 *ml* (*millilitre* ). The amounts of fill in a random sample of bottles, in *ml*, were 481, 479, 482, 480, 477, 478, 481 and 482. Suppose the population of amounts of fill is normal. Construct a 90% confidence interval for the mean amount in all bottles filled by the machine.
( $478.95 < \mu < 481.05$ )

**4.** (*a*) A random sample of size 36 is taken from a normal population with a known variance $\sigma^2 = 25$. If the mean of the sample is 42.6, find 95% confidence limits for the population mean.
( $40.967 < \mu < 44.233$ )

(*b*) A school wishes to estimate the average weight of students in the sixth grade. A random sample of $n = 25$ is selected and the sample mean is found to be $\bar{x} = 100$ *lbs*. The

standard deviation of the population is known to be 15 *lbs*. Compute 90% confidence interval for the population mean.
( $95.065 < \mu < 104.935$ )

**5.** (*a*) Suppose that the weights of 100 male students of a university represent a random sample of weights of 1546 students of the university. Find 99% confidence intervals for the mean weight of the students, given $\bar{x} = 67.45$ and $\hat{s} = 2.93$.
( $66.68 < \mu < 68.22$ )

(*b*) 150 bags of flour of a particular brand are weighed and the mean mass is found to be 748 *g* with standard deviation 3.6 *g*. Find 98% confidence intervals for the mean mass of bags of flour of this brand.
( $747.316 < \mu < 748.684$ )

**6.** (*a*) If the two-sided $100(1 - \alpha)\%$ confidence interval based on random sample taken from $X \sim N(\mu, \sigma^2)$ is $12.18 < \mu < 20.56$, find $\bar{x}$.
( $\bar{x} = 16.37$ )

(*b*) On the basis of the results obtained from a random sample of 100 men from a particular area, the 95% confidence interval for the mean height of the male population of the area was found to be (177.22 *cm*. to 179.18 *cm*). Find the value of $\bar{x}$, the mean of the sample and $\sigma$, the standard deviation of the normal population from which the sample was drawn. Find 98% confidence interval for the mean height.
( $\bar{x} = 178.2$, $\sigma = 5$, $177.04 < \mu < 179.36$ )

**7.** (*a*) Explain what is meant by the statement, 'we are $100(1 - \alpha)\%$ confident that our interval estimate contains $\mu$.'
{ In repeated sampling, $100(1 - \alpha)\%$ of all such confidence intervals contain $\mu$. }

(*b*) Explain what is meant by the statement, "we are 95% confident that our interval contains $\mu$".
( In repeated sampling, 95% of all such confidence intervals contain $\mu$. )

(*c*) If an 85% confidence interval is $27.5 < \mu < 43.8$, what does this statement mean ?
( Intervals so formed would contain $\mu$ 85% of the time. )

(*d*) If $\alpha = 0.10$, how many intervals would be expected to contain $\mu$ ?
( We would expect about 90% of the intervals to contain $\mu$ and 10% to miss $\mu$ in the long-run in repeated sampling.)

**8.** (*a*) What role does the sample mean play in a two-sided confidence interval for $\mu$, based on a random sample from $X \sim N(\mu, \sigma^2)$?
( The sample mean is the midpoint of the confidence interval but has no effect on the length of the interval.)

(*b*) When setting a two-sided $100(1 - \alpha)\%$ confidence interval for $\mu$, based on a random sample of size $n$ from a normal population, how the following changes will affect the length of the confidence interval for $\mu$: (Assume all other quantities remain fixed.)

| | | | |
|---|---|---|---|
| (*i*) | increasing $n$ | (*ii*) | increasing $(1 - \alpha)$ |
| (*iii*) | decreasing $n$ | (*iv*) | decreasing $(1 - \alpha)$ |
| (*v*) | increasing $\hat{s}^2$ | (*vi*) | increasing $\bar{x}$ |
| (*vii*) | increasing $\alpha$ | (*viii*) | decreasing $\hat{s}$ |

( decreased, increased, increased, decreased, increased, no effect, decreased, decreased )

**9.** (*a*) Define Student's *t*-statistic. What assumptions are made about the population where the *t*-distribution is used?

(*b*) The contents of 10 similar containers of a commercial soap are : 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 litres. Find 99% confidence interval for the mean soap content of all such containers, assuming an approximate normal distribution. ( $9.807 < \mu < 10.313$ )

**10.** (*a*) The masses in grams, of thirteen ball bearings taken at random from a batch are 21.4, 23.1, 25.9, 24.7, 23.4, 21.5, 25.0, 22.5, 26.9, 26.4, 25.8, 23.2, 21.9. Calculate a 95% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.
( $22.82 < \mu < 25.14$ )

(*b*) A random sample of seven independent observations of a normal variable gave $\sum x = 35.9$, $\sum x^2 = 186.19$. Calculate a 90% confidence interval for the population mean.
( $4.70 < \mu < 5.56$ )

**11.** (*a*) A random sample of eight observations of a normal variable gave $\sum x = 261.2$, $\sum (x - \bar{x})^2 = 3.22$. Calculate a 95% confidence interval for the population mean. ( $32.08 < \mu < 33.22$ )

(*b*) A sample of 12 measurements of the breaking strength of cotton threads gave a mean $\bar{x} = 209$ grams and a standard deviation $\hat{s} = 35$ grams. Find 95% and 99% confidence limits for the actual mean breaking strength.
( $186.76 < \mu < 231.24$; $177.62 < \mu < 240.38$ )

**12.** (*a*) A random sample of 16 values from a normal population showed a mean of 41.5 inches and a sum of squares of deviations from this mean equal to 135 $(\text{inches})^2$. Show that the 95% confidence limits for this mean are 39.9 and 43.1 inches.

(*b*) Find a 99% confidence interval for the mean of normal distribution with $\sigma = 2.5$ and if a sample of size 7 gave the values 9, 16, 10, 14, 8, 13, 14. What would be the confidence interval if $\sigma$ were unknown.
( $9.566 < \mu < 14.434$; $7.797 < \mu < 16.203$ when $\sigma$ is unknown. )

◆

## 12.9 CONFIDENCE INTERVAL FOR POPULATION PROPORTION OF SUCCESSES, $\pi$

The interval $(L, U)$ is a $100(1 - \alpha)\%$ confidence interval for the population proportion of successes $\pi$ if prior to sampling

$$P(L < \pi < U) = 1 - \alpha$$

This definition simply states that a confidence interval with confidence coefficient $1 - \alpha$ is an interval estimate such that the probability is $1 - \alpha$ that the calculated limits include $\pi$ for any random sampling. In other words, in many repeated random samples of size $n$ from a Bernoulli population, $100(1 - \alpha)\%$ of the interval estimates will include the true population proportion

of successes $\pi$ and therefore will be correct and $100\,\alpha\,\%$ of the interval estimates will not include $\pi$ and therefore will be incorrect.

In many problems we must estimate the population proportion or percentage, for example, the proportion of defectives found in shipment of raw materials upon inspection. In this case it seems to be reasonable that we are sampling from a Bernoulli population; hence our problem is to estimate its parameter $\pi$. The interval is based on the estimator $P = X/n$, the sample fraction of successes. We know that the sampling distribution of $P$ is a binomial distribution. The binomial distribution of the estimator $P$ can be approximated by the normal distribution with a mean of $\mu_P = \pi$ and a standard deviation of $\sigma_P = \sqrt{\pi(1-\pi)/n}$, when $n$ is large and $\pi$ is not too near 0 or 1. Consequently the distribution of the statistic

$$Z = \frac{P - \pi}{\sqrt{P(1-P)/n}}$$

will be approximately normal with mean 0 and variance 1. Then a two-sided confidence interval for population proportion of successes $\pi$ is given by

$$P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} < \pi < P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

If $p = x/n$ is the proportion of successes in an observed random sample of size $n$, then a $100\,(1-\alpha)\%$ confidence interval for $\pi$ is given by

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

This can be written $\quad p \pm z_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

**12.9.1 Sampling Without Replacement.** When sampling is done without replacement from a finite population of size $N$, the standard error of $P$ is given by

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}}$$

which is estimated as

$$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

If the sample size $n$ is greater than 5% of the population size $N$ (*i. e.*, $n > 0.05\,N$), then a $100\,(1-\alpha)\%$ confidence interval for $\pi$ is given

$$P \pm z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}}$$

The finite population correction $(N-n)/(N-1)$ may be ignored when the sample size $n$ is less than 5% of the population size $N$ (*i. e.*, $n < 0.05\,N$).

***Example* 12.13** *In a random sample of* 500 *young persons from a small town* 40 *were found to be unemployed. Compute a* 96% *confidence interval for the rate of unemployment in the town. Interpret the result.*

**Solution.** The sample size, number of successes and proportion of successes in the sample are

$$n = 500, \quad x = 40, \quad p = \frac{x}{n} = \frac{40}{500} = 0.08$$

*Confidence coefficient* : $1 - \alpha = 0.96$

$$1 - \alpha = 0.96 \Rightarrow \alpha = 0.04 \Rightarrow \alpha/2 = 0.02 \Rightarrow 1 - \alpha/2 = 0.98$$

$$z_{1-\alpha/2} = z_{0.98} = 2.054 \qquad \{ \text{From Table 10 } (a) \}$$

The two-sided 96% confidence interval for $\pi$ is

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$0.08 - 2.054\sqrt{\frac{0.08(1-0.08)}{500}} < \pi < 0.08 + 2.054\sqrt{\frac{0.08(1-0.08)}{500}}$$

$$0.055 < \pi < 0.105$$

We are 96% confident that rate of unemployment is between 5.5% to 10.5% because our procedure will produce true statement 96% of the time.

***Example* 12.14** *A poll is taken among the residents of a city and the surrounding country to determine the feasibility of a proposal to construct a civic centre. If* 2400 *of* 5000 *city residents favour it, find almost certain limits for the true fraction favouring the proposal to construct the civic centre.*

**Solution.** The sample size, number of successes and proportion of successes in the sample are

$$n = 5000, \quad x = 2400, \quad p = \frac{x}{n} = \frac{2400}{5000} = 0.48$$

*Confidence coefficient* : $1 - \alpha = 0.999$ ( almost certain is 99.9% confident )

$$1 - \alpha = 0.999 \Rightarrow \alpha = 0.001 \Rightarrow \alpha/2 = 0.0005 \Rightarrow 1 - \alpha/2 = 0.9995$$

$$z_{1-\alpha/2} = z_{0.9995} = 3.291 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 99.9% confidence interval for $\pi$ is

$$p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$0.48 - 3.291\sqrt{\frac{0.48(1-0.48)}{5000}} < \pi < 0.48 + 3.291\sqrt{\frac{0.48(1-0.48)}{5000}}$$

$$0.457 < \pi < 0.503$$

We are almost certain that the true fraction favouring the proposal to construct the civic centre lies between 0.457 and 0.503.

***Example* 12.15** *A random sample of* 250 *from the* 5000 *students in Govt. College, Gujranwala contained* 30 *left-handed students. Give an approximate* 95% *confidence interval for the proportion of left-handed students in the college.*

**Solution.** The sample size, number of successes and proportion of successes in the sample and the population size are

$$n = 250, \quad x = 30, \quad p = \frac{x}{n} = \frac{30}{250} = 0.12, \quad N = 5000$$

*Confidence coefficient* : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table } 10\,(b) \}$$

The two-sided 95% confidence interval for $\pi$ in the finite population is

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$0.12 \pm 1.960 \sqrt{\frac{0.12(1-0.12)}{250}} \sqrt{\frac{5000-250}{5000-1}}$$

$$(0.081, \ 0.159) \qquad \Rightarrow \qquad 0.081 < \pi < 0.159$$

## Exercise 12.3

**1.** (*a*) A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find (*i*) 95% and (*ii*) 99% confidence limits for the proportion of all the voters in favour of this candidate.

{ (*i*) $0.453 < \pi < 0.647$, (*ii*) $0.422 < \pi < 0.678$ }

(*b*) In a random sample of 1000 houses in a certain city, it is found that 228 own colour television sets. Find 98% confidence interval for the proportion of houses in this city that have coloured sets.

$(0.197 < \pi < 0.259)$

**2.** (*a*) In 40 tosses of a coin 24 heads were obtained. Find (*i*) 95% and (*ii*) 99.73% confidence limits for the proportion of heads which would be obtained in an unlimited number of tosses of the coin.

{ (*i*) $0.448 < \pi < 0.752$, (*ii*) $0.368 < \pi < 0.832$ }

(*b*) A random sample of 200 voters in a constituency included 110 who said they would vote for Mr. *A*. Assuming all the 15000 voters in the constituency would vote, give an approximate 95% confidence interval for the proportion who would vote for Mr. *A*.

$(0.4815 < \pi < 0.6185)$

(*c*) A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the percentage of bad pineapples in the consignment almost certainly lies between 8.05 and 17.95.

◆

## 12.10 COMPARATIVE STUDIES

To this point we have been concerned with inferences about parameters of a single population. We now turn our attention to estimation procedures that are important in comparing the parameter values of two populations. To make inferences about two populations; we must obtain two samples — one from each population. There are many methods by which the two samples could be obtained; we will discuss two of them in this text. These methods result in either *independent* or *dependent* samples.

**12.10.1 Independent Samples.** If two samples are selected, one from each of two populations, then the two samples are *independent* if the selection of objects from one population is unrelated to the selection of objects from the other population.

**12.10.2 Dependent Samples.** If two samples are selected, one from each of two populations then the two samples are *dependent*, if for each object selected from one population an object is chosen from the other population to form a pair of similar objects. These samples are also called as matched samples. The set of sample pairs is called a paired samples.

The key to recognizing two independent samples is to realize that they are always two different random samples, whereas the dependent samples always consist of matched, or paired, observations.

## 12.11 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO POPULATION MEANS, $\mu_1 - \mu_2$

In many business and management problems, we wish to estimate the difference between the means of two populations. For instance, we, may want to decide upon the basis of suitable samples to what extent, if any, a fertilizer is more effective than an existing fertilizer; a newly introduced product is more reliable than an existing product, or the degree to which a particular training programme improves worker attitudes or performance.

**12.11.1 Independent Samples:** ***Normal populations, known variances, any sample size.*** If $\bar{X}_1$ and $\bar{X}_2$ are respectively, the means of independent random samples of sizes $n_1$ and $n_2$ taken from two normal populations having means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, then $\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n)$ and that $\bar{X}_1$ is independent of $\bar{X}_2$.

Since any linear combination of independent normal random variables is also normally distributed, then $\bar{X}_1 - \bar{X}_2$ is a random variable having a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Thus the distribution of random variable

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

is a standard normal distribution. Then a two-sided $100\,(1 - \alpha)\%$ confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means of the two independent observed samples, then a $100\,(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

***Example* 12.16** *Apex's current packing machinery is known to pour ground coffee into "1-pound cans" with a standard deviation of* 0.6 *ounce. Apex is considering using a new packing machine which is said to pour coffee into "1-pound cans" more accurately, with a standard deviation of* 0.3 *ounce. Both machines pour ground coffee according to a normal distribution. Before deciding to invest, Apex wishes to evaluate the performance of the new machine against that of the old machine. A sample was taken on each machine against that of mean weight of the contents of the "1-pound cans" yielding the following result.*

| Sample | Size | Mean |
|---|---|---|
| Using Old Machine: I | $n_1 = 25$ | $\bar{x}_1 = 16.7$ |
| Using New Machine: II | $n_2 = 36$ | $\bar{x}_2 = 15.8$ |

*Construct a* 95% *confidence interval for the difference in the average weight of the contents poured by the old versus the new machine.*

**Solution.** The sizes and means of two samples and the standard deviations of two populations are

$$n_1 = 25, \qquad \bar{x}_1 = 16.7, \qquad \sigma_1 = 0.6$$

$$n_2 = 36, \qquad \bar{x}_2 = 15.8, \qquad \sigma_2 = 0.3$$

*Confidence coefficient* : $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{\text{From Table 10 } (b)\}$$

The two-sided 95% confidence limits for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(16.7 - 15.8) \pm 1.960\sqrt{\frac{(0.6)^2}{25} + \frac{(0.3)^2}{36}}$$

$$(0.65,\ 1.15) \qquad \Rightarrow \qquad 0.65 < \mu_1 - \mu_2 < 1.15$$

**12.11.2 Independent Samples:** ***Any populations, variances known/unknown, large samples.*** When both sample sizes are large ( say greater than 30 ) the assumptions regarding small samples can be greatly relaxed. It is no longer necessary to assume that the parent distributions are normal, because the Central Limit Theorem assures that $\bar{X}_1$ is approximately normally distributed with mean $\mu_1$ and variance $\sigma_1^2/n_1$, and that $\bar{X}_2$ is also approximately normally distributed with mean $\mu_2$ and variance $\sigma_2^2/n_2$, and that $\bar{X}_1$ is independent of $\bar{X}_2$, then $(\bar{X}_1 - \bar{X}_2)$ is approximately normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

Thus the random variable

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

has an approximately standard normal distribution. Because $n_1$ and $n_2$ are both large, the approximation remains valid of if $\sigma_1^2$ and $\sigma_2^2$ are replaced by their sample variances $\hat{S}_1^2$ and $\hat{S}_2^2$. The assumption of equal variance is not required in inferences derived from large samples. We can modify the previous result to obtain a confidence interval by substituting the sample variances $\hat{S}_1^2$ for $\sigma_1^2$ and $\hat{S}_2^2$ for $\sigma_2^2$ as long as both samples are large enough ($n_1 > 30$, $n_2 > 30$) for the Central Limit Theorem to be invoked. Hence the distribution of random variable $\bar{X}_1 - \bar{X}_2$ approaches a normal distribution with mean $\mu_1 - \mu_2$ and variance $\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2$.

Then the distribution of random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

approaches the standard normal distribution. Then a two-sided $100(1 - \alpha)\%$ approximate confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means and $\hat{s}_1^2$ and $\hat{s}_2^2$ are the variances of the two independent observed random samples, then a $100(1 - \alpha)\%$ approximate confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$$

***Example* 12.17** *Rural and urban students are to be compared on the basis of their scores on a nation wide musical aptitude test. Two random samples of sizes* 90 *and* 100 *are selected from rural and urban seventh class students. The summary statistics from the test scores are*

| Sample | Size | Mean | Standard deviation |
|---|---|---|---|
| Rural: I | $n_1 = 90$ | $\bar{x}_1 = 76.4$ | $\hat{s}_1 = 8.2$ |
| Urban: II | $n_2 = 100$ | $\bar{x}_2 = 81.2$ | $\hat{s}_2 = 7.6$ |

*Establish a* 98% *confidence interval for the difference in population mean scores between urban and rural students.*

**Solution.** *Confidence coefficient*: $1 - \alpha = 0.98$

$$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02 \Rightarrow \alpha/2 = 0.0 \Rightarrow 1 - \alpha/2 = 0.99$$

$$z_{1-\alpha/2} = z_{0.99} = 2.326 \quad \{ \text{From Table 10 } (b) \}$$

The two-sided 98% approximate confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{x}_2 - \bar{x}_1) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}$$

$$(81.2 - 76.4) \pm 2.326 \sqrt{\frac{(7.6)^2}{100} + \frac{(8.2)^2}{90}}$$

$$(2.1,\ 7.5) \quad \Rightarrow \quad 2.1 < \mu_2 - \mu_1 < 7.5$$

We conclude, with 98% confidence, that the mean of urban scores is at least 2.1 units higher and can be as much as 7.5 units higher than the mean of rural scores.

**12.11.3 Independent Samples:** ***Normal populations, same unknown variance, small samples.*** When $n_1$ and $n_2$ are small and $\sigma_1^2$ and $\sigma_2^2$ are unknown, the formula for constructing a confidence interval that we have been discussing cannot be used. However, for independent samples from two normal populations having the same unknown variance $\sigma^2$, we can develop a confidence interval for $\mu_1 - \mu_2$ as follows :

If $\bar{X}_1$ and $\bar{X}_2$ are respectively, the means of two independent random samples taken from populations which are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, then $\bar{X}_1 \sim N(\mu_1, \sigma^2/n)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2/n)$ and that $\bar{X}_1$ is independent of $\bar{X}_2$.

Since any linear combination of independent normal random variables is also normally distributed, then $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

Thus the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a standard normal distribution. Thus, if $\hat{S}_1^2$ and $\hat{S}_2^2$ are the two sample variances ( both estimating the variance $\sigma^2$ common to both populations ), the pooled (weighted arithmetic mean ) estimator of $\sigma^2$, denoted by $S_p^2$, is

$$S_p^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2} = \frac{\Sigma(X_{i1} - \bar{X}_1)^2 + \Sigma(X_{i2} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{(\Sigma X_{i1}^2 - n_1 \bar{X}_1^2) + (\Sigma X_{i2}^2 - n_2 \bar{X}_2^2)}{n_1 + n_2 - 2}$$

Then the random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a $t$-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Then a two-sided $100(1 - \alpha)\%$ confidence interval for difference between means of the two populations $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\nu;1-\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If $\bar{x}_1$ and $\bar{x}_2$ are the means of the two observed random samples and $s_p$ is the pooled estimate of the common standard deviation of the two normal populations, then a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

It should be noted that this is used under the conditions when the sample sizes are small (*i. e.*, $n_1 \leq 30$ and $n_2 \leq 30$). When both $n_1$ and $n_2$ are greater than 30, it is legitimate to use

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

as a good approximation.

***Example* 12.18** *Suppose you want to estimate the difference in annual operation costs for automobiles with rotary engines and those with standard engines. You find 8 owners of cars with rotary engines and 12 owners with standard engines, who have purchased their cars within the last two years and are willing to participate in the experiment. Each of the 20 owners keeps accurate records of the amount spent on operating his or her car (including gasoline, oil, repairs, etc.) for a 12 month period. All costs are recorded on a per 1000 mile basis to adjust for differences in mileage driven during the 12 month period. The results are summarized below:*

| *Sample* | *Size* | *Mean* | *Standard deviation* |
|---|---|---|---|
| *Rotary*: I | $n_1 = 8$ | $\bar{x}_1 = 56.96$ | $\hat{s}_1 = 4.85$ |
| *Standard*: II | $n_2 = 12$ | $\bar{x}_2 = 52.73$ | $\hat{s}_2 = 6.35$ |

*Estimate the true difference* $(\mu_1 - \mu_2)$ *between the mean operating cost per* 1000 *miles of cars with rotary and standard engines. Use a* 90% *confidence level.*

**Solution.** The pooled estimate of population common standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(8 - 1)(4.85)^2 + (12 - 1)(6.35)^2}{8 + 12 - 2}} = 5.813$$

*Confidence coefficient*: $1 - \alpha = 0.90$

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$$

*Degrees of freedom*: $\nu = n_1 + n_2 - 2 = 8 + 12 - 2 = 18$

$t_{\nu;1-\alpha/2} = t_{18;95} = 1.734$ (From Table 12)

The two-sided 90% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(56.96 - 52.73) \pm 1.734\,(5.813) \sqrt{\frac{1}{8} + \frac{1}{12}}$$

$$(-0.37,\ 8.83) \quad \Rightarrow \quad -0.37 < \mu_1 - \mu_2 < 8.83$$

***Example* 12.19** *Given two random samples from two independent normal populations, with*

| Sample | Size | Mean | Sum of squares |
|---|---|---|---|
| I | $n_1 = 11$ | $\bar{x}_1 = 75$ | $\sum (x_{i1} - \bar{x}_1)^2 = 372.1$ |
| II | $n_2 = 14$ | $\bar{x}_2 = 60$ | $\sum (x_{i2} - \bar{x}_2)^2 = 365.17$ |

*Find a* 99% *confidence interval for* $(\mu_1 - \mu_2)$. *Assume that population variances are equal.*

**Solution.** The pooled estimate of population common standard deviation is

$$s_p = \sqrt{\frac{\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{372.1 + 365.17}{11 + 14 - 2}} = 5.66$$

*Confidence coefficient*: $1 - \alpha = 0.99$

$$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow 1 - \alpha/2 = 0.995$$

*Degrees of freedom*: $\nu = n_1 + n_2 - 2 = 11 + 14 - 2 = 23$

$t_{\nu;1-\alpha/2} = t_{23;0.995} = 2.807$ (From Table 12)

The two-sided 99% confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(75 - 60) \pm 2.807\,(5.66) \sqrt{\frac{1}{11} + \frac{1}{14}}$$

$$(8.6,\ 21.4) \quad \Rightarrow \quad 8.6 < \mu_1 - \mu_2 < 21.4$$

***Example* 12.20** *A course in mathematics is taught to 10 students by the conventional class room method. A second group of 12 students was given the same course by means of programmed materials. At the end of the semester, the same examination was given to each group. Their scores are given below:*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Group* I | 70 | 66 | 76 | 77 | 73 | 72 | 68 | 74 | 75 | 69 | | |
| *Group* II | 77 | 83 | 92 | 85 | 82 | 84 | 80 | 86 | 91 | 93 | 80 | 87 |

*Compute a 90% confidence interval for the difference between the average scores of the two populations. Assume the populations to be approximately normal with equal variance*

**Solution.** The sample means and pooled estimate of population common standard deviation are

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 70 | 66 | 76 | 77 | 73 | 72 | 68 | 74 | 75 | 69 | | | $\Sigma x_1 = 720$ |
| $x_2$ | 77 | 83 | 92 | 85 | 82 | 84 | 80 | 86 | 91 | 93 | 80 | 87 | $\Sigma x_2 = 1020$ |
| $x_1^2$ | 4900 | 4356 | 5776 | 5929 | 5329 | 5184 | 4624 | 5476 | 5625 | 4761 | | | $\Sigma x_1^2 = 51960$ |
| $x_2^2$ | 5929 | 6889 | 8464 | 7225 | 6724 | 7056 | 6400 | 7396 | 8281 | 8649 | 6400 | 7569 | $\Sigma x_2^2 = 86982$ |

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{720}{10} = 72$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{1020}{12} = 85$$

$$s_p = \sqrt{\frac{(\Sigma x_1^2 - n_1 \bar{x}_1^2) + (\Sigma x_2^2 - n_2 \bar{x}_2^2)}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{\{51960 - 10(72)^2 + \{86982 - 12(85)^2\}}{10 + 12 - 2}} = 4.48$$

*Confidence coefficient* : $1 - \alpha = 0.90$

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$$

*Degrees of freedom* : $\nu = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$

$$t_{\nu;1-\alpha/2} = t_{20;0.95} = 1.725 \quad \text{(From Table 12)}$$

The two-sided 90% confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{x}_2 - \bar{x}_1) \pm t_{\nu;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(85 - 72) \pm 1.725\,(4.48)\sqrt{\frac{1}{10} + \frac{1}{12}}$$

$$13 \pm 3.31 \quad \Rightarrow \quad 9.69 < \mu_2 - \mu_1 < 16.31$$

## Exercise 12.4

**1.** (*a*) A test in statistics was given to 50 girls and 75 boys. The girls made an average grade of 76 with a standard deviation of 6, while the boys made an average grade of 82 with a standard deviation of 8. Find a 96% confidence interval for the difference $\mu_1 - \mu_2$, where $\mu_1$ is the mean score of all boys and $\mu_2$ is the mean score of all girls who might take this test.

$(3.42 < \mu_1 - \mu_2 < 8.58)$

(*b*) A manufacturing company consists of two departments producing identical products. It is suspected that the hourly outputs in the two departments are different. Two random samples of production hours are respectively selected and the following data are obtained:

| | *Department 1* | *Department 2* |
|---|---|---|
| Sample size: | 64 | 49 |
| Sample mean: | 100 | 90 |

The variances of the hourly outputs for the two departments are known to be $\sigma_1^2 = 256$ and $\sigma_2^2 = 196$ respectively. What is the point estimate for the true difference between the mean outputs of the two departments ? Find the 95 percent confidence limits for the true difference.

$(\bar{x}_1 - \bar{x}_2 = 10;\ 4.456 < \mu_1 - \mu_2 < 15.544)$

**2.** (*a*) Two independent samples of 100 mechanists and 100 carpenters are taken to estimate the difference between the weekly wages of the two categories of workers. The relevant data are given below:

| | Sample mean wages | Population variance |
|---|---|---|
| Mechanists: | 345 | 196 |
| Carpenters: | 340 | 204 |

Determine the 95% and the 99% confidence limits for the true difference between the average wages for machinists and carpenters.

$(1.08 < \mu_1 - \mu_2 < 8.92;\ -0.152 < \mu_1 - \mu_2 < 10.152)$

(*b*) General Incorporated Mill's packing machinery is known to pour dry cereal into economy-size boxes with a standard deviation of 0.6 ounce. Two samples taken on two machines yields the following information:

| *Machine I* | *Machine II* |
|---|---|
| $n_1 = 15$ | $n_2 = 21$ |
| $\bar{x}_1 = 18.7$ ounces | $\bar{x}_2 = 21.9$ ounces |

Assuming machine I packages a content that is $N(\mu_1, 0.36)$ and machine II packages a content that is $N(\mu_2, 0.36)$, construct a 95% confidence interval estimate of $\mu_2 - \mu_1$.

$(2.8 < \mu_2 - \mu_1 < 3.6)$

**3.** (*a*) A sample of 150 brand *A* light bulbs showed a mean lifetime of 1400 hours with a standard deviation of 120 hours. A sample of 200 brand *B* light bulbs showed a mean lifetime of 1200 hours with a standard deviation of 80 hours. Find 95% and 99% confidence limits for the difference between the mean lifetime of the populations of brands *A* and *B*.

$(177.825 < \mu_1 - \mu_2 < 222.175;\ 170.856 < \mu_1 - \mu_1 < 229.144)$

(*b*) Let two independent random samples, each of size 100, from independent normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ yield $\bar{x}_1 = 4.8$, $\hat{s}_1^2 = 8.64$, $\bar{x}_2 = 5.6$, $\hat{s}_2^2 = 7$. Find a 95% confidence interval for $(\mu_2 - \mu_1)$.

$(0.025 < \mu_2 - \mu_1 < 1.575)$

**4.** (*a*) In order to ascertain the age distribution of operatives in a certain industry, random samples of 1720 males and 1230 females are drawn. The sample means and standard deviations were 33.93 years and 14.20 years for the males and 27.44 years and 10.79 years for the females. Calculate the 95 percent confidence interval for

(*i*) the mean age of all male operatives,

(*ii*) the mean age of all female operatives,

(*iii*) the difference between their mean ages.

$(33.259 < \mu_1 < 34.601;\ 26.837 < \mu_2 < 28.043;\ 5.588 < \mu_1 - \mu_2 < 7.392)$

(*b*) The means and variances of the weekly incomes in rupees of the workers employed in the different factories, from the samples are given below:

| *Sample* | *Size* | *Mean* | *Variance* |
|---|---|---|---|
| *Factory A* | 160 | 12.80 | 64 |
| *Factory B* | 220 | 11.25 | 49 |

(*i*) What is the maximum likelihood estimate of the difference in mean incomes?

(*ii*) Compute the 95 percent confidence interval estimate for the real differences in the incomes of the workers from the two factories.

{ (*i*) 1.55, (*ii*) $0.003 < \mu_1 - \mu_2 < 3.097$ }

**5.** (*a*) Let two independent random samples, each of size 100, from two independent normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ yield $\bar{x} = 4.8$, $\hat{s}_1^2 = 8.64$, $\bar{y} = 5.6$, $\hat{s}_2^2 = 7.88$. Find a 95% confidence interval for $(\mu_1 - \mu_2)$.

$(-1.6 < \mu_1 - \mu_2 < 0)$

(*b*) Given that

$$\bar{x}_1 = 75, \quad n_1 = 9, \quad \Sigma(x_{i1} - \bar{x}_1)^2 = 1482$$

$$\bar{x}_2 = 60, \quad n_2 = 16, \quad \Sigma(x_{i2} - \bar{x}_2)^2 = 1830$$

and assuming that the two samples were randomly selected from two normal populations in which $\sigma_1^2 = \sigma_2^2$ (but unknown), calculate an 80% confidence interval for the difference between the two population means.
$(8.4 < \mu_1 - \mu_2 < 21.6)$

**6.** (*a*) Two random samples of size $n_1 = 9$ and $n_2 = 16$ from two independent population having normal distributions provide the means and standard deviations; $\bar{x}_1 = 64$, $\bar{x}_2 = 59$, $\hat{s}_1 = 6$ and $\hat{s}_2 = 5$. Find a 95% confidence interval for $\mu_1 - \mu_2$ assuming $\sigma_1 = \sigma_2$.
$(0.37 < \mu_1 - \mu_2 < 9.63)$

(*b*) A course in mathematics is taught to 12 students by the conventional class- room method. A second group of 10 students was given the same course by means of programmed materials. At the end of the course, the same examination was given to each group. The 12 students meeting in the class room made an average grade of 85 with a standard deviation of 4, while the 10 students using programmed materials made an average of 81 with a standard deviation of 5. Find a 90% confidence interval for the difference between the population means, assuming the populations to be approximately normally distributed with equal variance.
$(0.693 < \mu_1 - \mu_2 < 7.307)$

◆

## 12.12 CONFIDENCE INTERVAL FOR DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS, $\pi_1 - \pi_2$

We now turn to statistical inferences concerning a comparison between the rates of incidence of a characteristic into populations. Comparing infant mortality in two groups, the unemployment rates in rural and urban populations, and the proportion of defective items produced by two competing manufacturing processes are the examples of this type. The unknown proportion of elements possessing the particular characteristic in population I and in population II are denoted by $\pi_1$ and $\pi_2$, respectively. Our aim is to construct confidence intervals for the parameter $\pi_1 - \pi_2$.

A random sample of size $n_1$ is taken from population I and the number of successes is denoted by $X_1$. An independent random sample of size $n_2$ is taken from population II and the number of successes is denoted by $X_2$. The sample proportions of successes are

$$P_1 = \frac{X_1}{n_1}, \qquad P_2 = \frac{X_2}{n_2}$$

An intuitively appealing estimator for $\pi_1 - \pi_2$ is the difference between the sample proportions $P_1 - P_2$. When constructing the confidence intervals for $\pi_1 - \pi_2$, we will use the sampling distribution of $P_1 - P_2$.

When both sample sizes $n_1$ and $n_2$ are large, the Central Limit Theorem assures that $P_1$ is approximately normal with mean $\pi_1$ and variance $\pi_1(1-\pi_1)/n_1$ and that $P_2$ is approximately normal with mean $\pi_2$ and variance $\pi_2(1-\pi_2)/n_2$ and that $P_1$ is independent of $P_2$

Since any linear combination of independent normal random variables is also normally distributed then for large sample sizes $n_1$ and $n_2$, the sampling distribution of the random variable $P_1 - P_2$ is approximately normal with mean

$$\mu_{P_1 - P_2} = \pi_1 - \pi_2$$

and standard deviation

$$\sigma_{P_1 - P_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

The first result shows that $P_1 - P_2$ is an unbiased estimator of $\pi_1 - \pi_2$. For large sample sizes $n_1$ and $n_2$, the random variable

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

is approximately standard normal. The estimate of the standard error of $P_1 - P_2$ can be obtained by replacing $\pi_1$ and $\pi_2$ by their sample estimates $P_1$ and $P_2$ as

$$\hat{\sigma}_{P_1 - P_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

The random variable $Z$ then becomes

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}}$$

Therefore, in the case of two large, independent random samples a $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ can be readily constructed from this approximation.

Then a two-sided $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is given by

$$(P_1 - P_2) \pm z_{1-\alpha/2}\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

If $p_1$ and $p_2$ are the proportions in the two large, independent observed random samples, then a $100(1-\alpha)\%$ interval for $\pi_1 - \pi_2$ is given by

$$(p_1 - p_2) \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

***Example* 12.21** *An antibiotic for pneumonia was injected into* 100 *patients with kidney malfunctions (called uremic patients ) and into* 80 *patients with no kidney malfunctions (called normal patients). Some allergic reaction developed in* 40 *of the uremic patients and in* 16 *of the normal patients. Construct a* 95% *confidence interval for the difference between the population proportions.*

**Solution.** The sizes, number of successes and proportions of successes in the two samples are

$$n_1 = 100, \qquad x_1 = 40, \qquad p_1 = \frac{x_1}{n_1} = \frac{40}{100} = 0.4$$

$$n_2 = 80, \qquad x_2 = 16, \qquad p_2 = \frac{x_2}{n_2} = \frac{16}{80} = 0.2$$

*Confidence coefficient*: $1 - \alpha = 0.95$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{1-\alpha/2} = z_{0.975} = 1.960 \qquad \{ \text{From Table 10 } (b) \}$$

The two-sided 95% confidence interval for difference in the population proportions $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$(0.4 - 0.2) \pm 1.960 \sqrt{\frac{0.4(1-0.4)}{100} + \frac{0.2(1-0.2)}{80}}$$

$$(0.07, 0.33) \qquad \Rightarrow \qquad 0.07 < \pi_1 - \pi_2 < 0.33$$

## Exercise 12.5

**1.** **(*a*)** In a poll of college students in a large state university, 300 out of 400 students living in dormitories approved a certain course of action, whereas 200 out of 300 students not living in dormitories approved it. Estimate the difference in the proportions favouring the course of action and compute 90% confidence interval for it.
$(p_1 - p_2 = 0.08; \quad 0.023 < \pi_1 - \pi_2 < 0.137)$

**(*b*)** In a random sample of 400 adults and 600 teenagers who watched a certain television programme. 100 adults and 300 teenagers indicated that they liked it. Construct 95% and 99% confidence limits for the difference in proportions of all adults and all teenagers who watched the programme and liked it.
$(0.19 < \pi_2 - \pi_1 < 0.31; \quad 0.17 < \pi_2 - \pi_1 < 0.33)$

**2.** **(*a*)** A poll is taken among the residents of a city and the surrounding country to determine the feasibility of a proposal to construct a civic centre. If 2400 of 5000 city residents favour the proposal and 1200 of 2000 country residents favour it, find a 95% confidence interval for the true difference in the proportions favouring the proposal

to construct the civic centre.
$(0.0945 < \pi_2 - \pi_1 < 0.1455)$

(*b*) The population of interest are the voting preferences of all registered voters in the Punjab and the Sind. Two independent random samples were taken from these populations and the values $n_1 = n_2 = 1000$, $p_1 = 0.54$ and $p_2 = 0.47$. Find a 95% confidence interval for $\pi_1 - \pi_2$.
$(0.026 < \pi_1 - \pi_2 < 0.114)$

3. (*a*) A market survey organization carried out a product taste study with consumers in two regions. In one region, a random sample of $n_1 = 400$ consumers was selected while in the other region an independent random sample of $n_2 = 300$ consumers was selected. Each person was asked to indicate which of two servings of product had a better taste. Unknown to the subject, one serving was a new high protein breakfast cereal and the other was an existing cereal. In the first region, proportion $p_1 = 0.55$ of the sample persons preferred the new cereal and in the second region the proportion was $p_2 = 0.65$. Construct a 90% confidence interval for $\pi_2 - \pi_1$.
$(0.039 < \pi_2 - \pi_1 < 0.161)$

(*b*) Independent random samples are selected from two populations with fractions of success $\pi_1$ and $\pi_2$. Construct a 95% confidence interval for $\pi_1 - \pi_2$ for each of the following cases.

(*i*) $n_1 = 100$ $p_1 = 0.72$ $n_2 = 100$ $p_2 = 0.61$

(*ii*) $n_1 = 130$ $p_1 = 0.16$ $n_2 = 210$ $p_2 = 0.25$

(*iii*) $n_1 = 70$ $p_1 = 0.53$ $n_2 = 60$ $p_2 = 0.48$

{ (*i*) $-0.02$ to $0.24$ (*ii*) $-0.176$ to $-0.004$ (*iii*) $-0.12$ to $0.22$ }

◆

## Exercise 12.6
### Objective Questions

1. Fill in the blanks.

(*i*) Statistical ——— is the conclusion made about the unknown value of population parameter by using the sample observations. (*inference*)

(*ii*) The statistical ——— is a procedure of making judgment about the unknown value of population parameters by using the sample observations. (*estimation*)

(*iii*) The object of ——— estimation is to obtain a single number from the sample that is intended for estimating the unknown true value of a population parameter. (*point*)

(*iv*) A point estimator is a ——— variable whereas an estimate is a constant. (*random*)

(*v*) An estimator is ——— if its expected value is equal to the population parameter to be estimated. (*unbiased*)

(*vi*) If $T$ is a biased estimator, then ———— is the difference of its expected value from the parameter $\theta$ to be estimated. (*bias*)

(*vii*) The sample mean $\overline{X}$ is an ———— estimator of population mean $\mu$. (*unbiased*)

(*viii*) The sample proportion $P$ is an ———— estimator of population proportion $\pi$. (*unbiased*)

(*ix*) The sample variance $\hat{S}^2 = \Sigma(X - \overline{X})^2/(n-1)$ is an ———— estimator of population variance $\sigma^2$. (*unbiased*)

(*x*) ———— estimation is a procedure of constructing an interval from a random sample, such that prior to sampling, it has a high specified probability of including the unknown true value of a population parameter. (*Interval*)

**2.** Fill in the blanks.

(*i*) The width of a confidence interval is ———— if the level of confidence ( $1 - \alpha$ ) is decreased. (*decreased*)

(*ii*) The width of a confidence interval is ———— related to confidence coefficient. (*directly*)

(*iii*) The precision of confidence interval is increased by ———— the level of confidence. (*decreasing*)

(*iv*) The width of a confidence interval ———— if the sample size is increased. (*decreases*)

(*v*) The confidence coefficient is also called ———— of confidence. (*level*)

(*vi*) $1 - \alpha$ is the ———— that the interval estimator includes the unknown true value of the population parameter. (*probability*)

(*vii*) A sample consisting of 30 or less observations is known as a ———— sample. (*small*)

(*viii*) A sample consisting of more than 30 observations is known as a ———— sample. (*large*)

**3.** Mark off the following statements as *true* or *false*

(*i*) The types of statistical inferences are estimation of parameters and testing of hypotheses. (*true*)l

(*ii*) The types of statistical estimation of parameters are point estimation and interval estimation. (*true*)

(*iii*) A point estimator is a sample statistic that is used to estimate the unknown true value of a population parameter. (*true*)

(*iv*) A point estimate is a specific value of an estimator computed from the sample data after the sample has been observed. (*true*)

(*v*) An estimate obtained from the sample observations is always a point estimate. (*false*)

(*vi*) Point estimators may be more useful than interval estimators because probability statements are attached to point estimates. (*false*)

(*vii*) The point estimation provides two values with a probability statement for estimating the unknown true value of a population parameter. (*false*)

(*viii*) A confidence interval is a type of statistical inference. (*true*)

(*ix*) A point estimate provides information about the precision of the estimate. (*false*)

**4.** Mark off the following statements as *true* or *false*

(*i*) We cannot control the precision of an interval estimate by the choice of sample size or level of confidence. (*false*)

(*ii*) The width of a confidence interval increases if the confidence coefficient is decreased. (*false*)

(*iii*) The width of a confidence interval decreases if the confidence coefficient is decreased. (*true*)

(*iv*) The width of a confidence interval can be decreased by decreasing the confidence coefficient. (*true*)

(*v*) The precision of an interval estimate can be increased by decreasing the sample size. (*false*)

(*vi*) The precision of an interval estimate can be increased either by increasing the sample size or by decreasing the confidence coefficient. (*true*)

(*vii*) $\alpha$ is the probability that the interval estimator includes the unknown true value of the population parameter. (*false*)

(*viii*) The statistic $T$ can be used in making confidence interval for $\mu$ when population is non-normal. (*false*)