# Chapter 13

*Nothing is good or bad by comparison.*

—Thomas Fuller

# Correlation Analysis

## LEARNING OBJECTIVES

After studying this chapter, you should be able to

- express quantitatively the degree and direction of the covariation or association between two variables.
- determine the validity and reliability of the covariation or association between two variables.
- provide a test of hypothesis to determine whether a linear relationship actually exists between the variables.

## 13.1 INTRODUCTION

The statistical methods, discussed so far, are used to analyse the data involving only one variable. Often an analysis of data concerning two or more quantitative variables is needed to look for any statistical relationship or association between them that can describe specific numerical features of the association. The knowledge of such a relationship is important to make inferences from the relationship between variables in a given situation. Few instances where the knowledge of an association or a relationship between two variables would be helpful to make decision are as follows:

- Family income and expenditure on luxury items.
- Yield of a crop and quantity of fertilizer used.
- Sales revenue and expenses incurred on advertising.
- Frequency of smoking and lung damage.
- Weight and height of individuals.

A statistical technique that is used to analyse the strength (magnitude) and direction of the relationship between two quantitative variables is called ***correlation analysis***. A few definitions of correlation analysis are as follows:

- An analysis of the relationship of two or more variables is usually called correlation.
  — A. M. Tuttle
- When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.
  —Croxton and Cowden

The ***coefficient of correlation*** is a number that indicates the *strength* and *direction* of statistical relationship between two variables.

- The ***strength*** of the relationship is determined by the closeness of the points to a straight line when a pair of values of two variables are plotted on a graph. A straight line is used as the frame of reference for evaluating the relationship.
- The ***direction*** is determined by whether one variable generally increases or decreases when the other variable increases.

The following questions determine the importance of examining the statistical relationship between two or more variables and accordingly requires the statistical methods to answer these questions:

(i) Is there an association between two or more variables? If yes, what is the form and degree of that relationship?
(ii) Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?
(iii) Can the relationship be used to predict the most likely value of a dependent variable for the given value of independent variable or variables?

The first two questions will be answered in this chapter, while the third question will be answered in next chapter.

For correlation analysis, the data on values of two variables must come from sampling in pairs, one for each of the two variables.

## 13.2 SIGNIFICANCE OF MEASURING CORRELATION

The objective of any scientific research is to establish relationships between two or more sets of observations or variables to arrive at some valid conclusion. Few advantages of measuring an association (or correlation) between two or more variables are as under:

1. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective.
   —W. A. Neiswanger
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality. — Tippett
3. In economic theory, an association (or correlation) between two or more variables, such as price, supply and quantity demanded; customers retention is related to convenience, amenities and service standards; yield of a crop is related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on is established.
4. In healthcare, an association (or correlation) between two or more variables such as validity and reliability of clinical measures; effect on health due to certain biological or environmental factors, blood pressure and age of a person; inter-observer reliability for two doctors who are assessing a patient's disease, and so on is established.

**Coefficient of Correlation:** A statistical measure of the degree of association between two variables.

## 13.3 CORRELATION AND CAUSATION

Correlation is one the three criteria for establishing a causal relationship between two or more variables. While correlation coefficient only measures the strength of a linear relationship but it does not necessarily imply a causal relationship. The following factors should be examined to interpret the nature and extent of relationship between two or more variables:

(i) ***Chance Coincidence:*** The inferences drawn from the value of correlation coefficient may not be of any statistical significance because variables might be entirely different and unrelated. Any association between them may be only by a chance. For example, (i) a positive correlation between growth in population and wheat production in the country has no statistical significance, and (ii) the correlation in sales revenue and expenditure on advertisements over a period of time should be statistically significant and not just due to biased sampling or sampling error.

(ii) ***Influence of Third Variable:*** Clinically, it has been proved that smoking causes lung damage. However, there are often multiple reasons such as stress, quality of food and air pollution, of health problems. Similarly, the yield of rice and tea is positively correlated because both the crops are influenced by the amount of rainfall. But the yield of any one is not influenced by other.

(iii) ***Mutual Influence:*** Although two variables might be highly correlated, still it is difficult to say as to which variable is influencing the other. For example, variables like price supply, and demand of a commodity are mutually correlated. As price of a commodity increases, its demand decreases, so price influences the demand level. But when demand of a commodity increases, its price also increases so demand influences the price.

## 13.4 TYPES OF CORRELATIONS

There are three broad types of correlations:

(i) Positive and negative
(ii) Linear and non-linear
(iii) Simple, partial and multiple

In this chapter, we will discuss simple linear positive or negative correlation analysis.

### 13.4.1 Positive and Negative Correlations

The ***positive (or direct) correlation*** refers to an association between two variables where their values change (i.e., increasing or decreasing) in the same direction. The ***negative (or inverse) correlation*** refers to an association between two variables where their values change (i.e., increasing or decreasing) in the opposite direction.

**Illustration**

*Positive Correlation*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Increasing | → $x$ | : | 5 | 8 | 10 | 15 | 17 |
| Increasing | → $y$ | : | 10 | 12 | 16 | 18 | 20 |
| Decreasing | → $x$ | : | 17 | 15 | 10 | 8 | 5 |
| Decreasing | → $y$ | : | 20 | 18 | 16 | 12 | 10 |

*Negative Correlation*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Increasing | → $x$ | : | 5 | 8 | 10 | 15 | 17 |
| Decreasing | → $y$ | : | 20 | 18 | 16 | 12 | 10 |
| Decreasing | → $x$ | : | 17 | 15 | 12 | 10 | 6 |
| Increasing | → $y$ | : | 2 | 7 | 9 | 13 | 14 |

**Remarks:** The change (increasing or decreasing) in values of both the variables may not be proportional or fixed.

### 13.4.2 Linear and Non-linear Correlations

A linear correlation refers to an association between two variables where variation in their values is either proportional or fixed. The following pattern of variation in the values of two variables $x$ and $y$ reveals linear correlation.

| | | | | | |
|---|---|---|---|---|---|
| $x$ : | 10 | 20 | 30 | 40 | 50 |
| $y$ : | 40 | 60 | 80 | 100 | 120 |

When these pairs of values of $x$ and $y$ are plotted on a graph paper, the line joining these points would be a straight line.

A non-linear (or curvy linear) correlation refers to an association between two variables where variation in their values is neither proportional nor fixed. The following pattern of variation in the values of two variables $x$ and $y$ reveals non-linear correlation.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ : | 8 | 9 | 9 | 10 | 10 | 28 | 29 | 30 |
| $y$ : | 80 | 130 | 170 | 150 | 230 | 560 | 460 | 600 |

When these pairs of values of $x$ and $y$ are plotted on a graph paper, the line joining these points would not be a straight line, rather it would be curvy linear.

### 13.4.3 Simple, Partial and Multiple Correlations

The distinction between simple, partial and multiple correlations is based upon the number of variables involved in the correlation analysis.

If only two variables are chosen to study correlation between them, then such a correlation is referred to as ***simple correlation***. A study on the yield of a crop with respect to only amount of fertilizer used, or sales revenue with respect to amount of money spent on advertisement, are a few examples of simple correlation.

In ***partial correlation***, two variables are chosen to study the correlation between them but the effect of other influencing variables is kept constant. For example (i) yield of a crop is influenced by the amount of fertilizer applied, whereas effect of other influencing variables such as rainfall, quality of seed, type of soil and pesticides is kept constant, and (ii) sales revenue from a product is influenced by the level of advertising expenditure, whereas effect of other influencing variables such as quality of the product, price, competitors, distribution and so on is kept constant.

In ***multiple correlation***, more than two variables are chosen to study the correlation among them. For example, (i) employer-employee relationship in any organization may be examined with reference to, training and development facilities; medical, housing and education to children facilities; salary structure; grievances handling system; and so on, and (ii) sales revenue from a product may be examined in relation with the level of advertising expenditure, quality of the product, price, competitors, distribution and so on.

## 13.5 METHODS OF CORRELATION ANALYSIS

The correlation between two ratio-scaled (numeric) variables is represented by the letter, $r$, which takes on values between –1 and +1 only and is referred to as '**Pearson product moment correction**' or **correlation coefficient**. The correlation coefficient is a relative (scale free) number and so its interpretation is independent of the units of measurement of two variables, say $x$ and $y$.

In this chapter, the following methods of calculating a correlation coefficient between two variables $x$ and $y$ are discussed:

- Scatter Diagram method
- Karl Pearson's Coefficient of Correlation method
- Spearman's Rank Correlation method
- Method of Least-squares

Figure 13.1 shows how the strength of the association between two variables is represented by the coefficient of correlation.

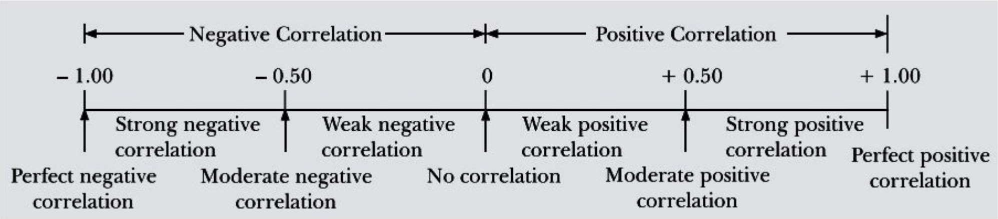

Negative Correlation
Positive Correlation
– 1.00
– 0.50
0
+ 0.50
+ 1.00
Strong negative correlation
Weak negative correlation
Weak positive correlation
Strong positive correlation
Perfect negative correlation
Moderate negative correlation
No correlation
Moderate positive correlation
Perfect positive correlation

**Figure 13.1** Interpretation of Correlation Coefficient

### 13.5.1 Scatter Diagram Method

The **scatter diagram** method is an at-a-glance method to understand an apparent relationship (if any) between two variables. A scatter diagram (or a graph) can be traced on a graph paper by plotting pairs of values of variables, $x$ and $y$, taking values of variable, $x$ on the $x$-axis and values of variable, $y$- on the $y$-axis. The horizontal and vertical axes are scaled in units corresponding to the variables $x$ and $y$, respectively. A straight line drawn through these pair of values describes different types of relationships between the two variables.

Figure 13.2 shows examples of different types of relationships based on pairs of values of $x$ and $y$ in a sample data. The patterns shown in Figs 13.2(a) and (b) represent linear relationships since the patterns are described by straight lines. The pattern in Fig. 13.2(a) shows a *positive* relationship since the value of $y$ tends to increase as the value of $x$ increases, whereas pattern in Fig. 13.2(b) shows a *negative* relationship since the value of $y$ tends to decrease as the value of $x$ increases.

The pattern shown in Fig. 13.2(c) illustrates very low or no relationship between the values of $x$ and $y$, whereas Fig. 13.2(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. The wider scattering indicates that there is a lower degree of association between the two variables $x$ and $y$ than there is in Fig. 13.2(a).

#### Interpretation of Correlation Coefficients

While interpreting correlation coefficient $r$, the following points should be taken into account:

- A low positive or negative value of correlation coefficient, $r$, indicates that the relationship is poorly described by a straight line. A non-linear relationship may also exist.
- A correlation is an observed association and does not indicate any *cause-and-effect* relationship.

**Scatter Diagram:** A graph of pairs of values of two variables that is plotted to indicate a visual display of the pattern of their relationship.

#### Types of Correlation Coefficients

Table 13.1 shows several types of correlation coefficients used in statistics along with the conditions of their use. All of them are appropriate for quantifying linear relationship between two variables $x$ and $y$.

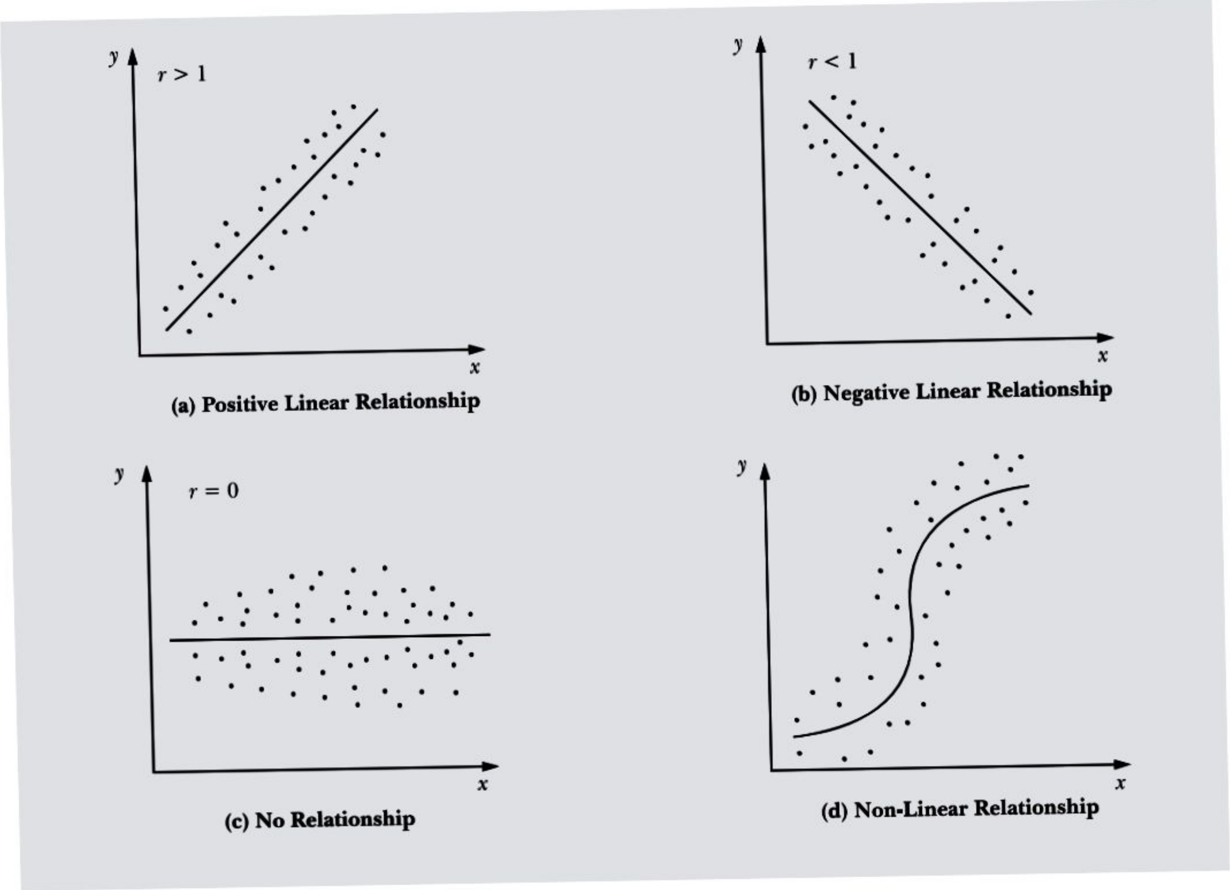
**Table 13.1:** Types of Correlation Coefficients

| *Coefficient* | *Conditions Applied for Use* |
|---|---|
| • $\phi$ (phi) | Both $x$ and $y$ variables are measured on a nominal scale |
| • $\rho$ (rho) | Both $x$ and $y$ variables are measured on, or changed to, ordinal scales (rank data) |
| • $r$ | Both $x$ and $y$ variables are measured on an interval or ratio scale scales (numeric data) |

The correlation coefficient, denoted by $\eta$(eta), is used for quantifying non-linear relationships (It is beyond the scope of this text). In this chapter, methods of calculating

Pearson's correlation coefficients, $r$ and Spearman's correlation coefficients, $R$, are discussed.

**Figure 13.2** Typical Examples of Correlation Coefficient



(a) Positive Linear Relationship

(b) Negative Linear Relationship

(c) No Relationship

(d) Non-Linear Relationship

### Features of the Correlation Coefficient

The following are the common features among all correlation coefficient:

(i) The value of correlation coefficient, $r$, depends on the slope of the line passing through the data points and the scattering of the pair of values of variables $x$ and $y$ about this line.

(ii) The sign of the correlation coefficient indicates the direction of the relationship. The positive correlation denoted by + (positive sign) indicates that the direction of increase (or decrease) in the value of two variables is same. While negative correlation denoted by – (minus sign) indicates that direction of increase (or decrease) in the value of two variables is opposite.

(iii) The values of the correlation coefficient range from + 1 to – 1 regardless of the units of measurements of $x$ and $y$. That is, correlation coefficient is a pure number independent of the unit of measurement.

(iv) The value of correlation coefficient $r = +1$ or $-1$ indicates perfect linear association (relationship) between two variables, $x$ and $y$. A perfect correlation implies that every observed pair of values of $x$ and $y$ falls on the straight line.

(v) The value of correlation coefficient indicates the strength of association (relationship) between two variables, i.e., a closeness of the observed pair of values of $x$ and $y$ to the straight line. The sign of the correlation coefficient indicates the strength of the linear relationship.

(vi) The value of correlation coefficient remains unchanged when a constant value is subtracted from every pair of values of variables $x$ and $y$ (also referred as change of origin), also when a pair of values of variables $x$ and $y$ are divided or multiplied by a constant (also referred to as change of scale).

(vii) The value of correlation coefficient, $r = 0$, indicates that the straight line through the data points is horizontal, and therefore no association (relationship) between two variables $x$ and $y$.

(viii) The square, $r^2$, of correlation coefficient, $r$, value is referred to as ***coefficient of determination***.

**Example 13.1:** Given the following data:

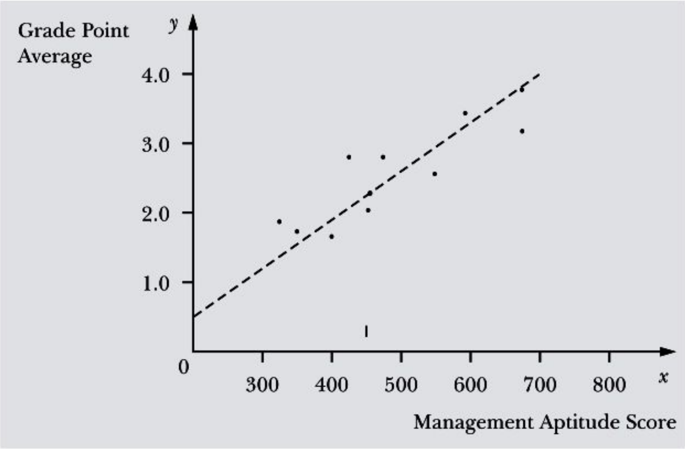
| Student | : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Management aptitude score | : | 400 | 675 | 475 | 350 | 425 | 600 | 550 | 325 | 675 | 450 |
| Grade point average | : | 1.8 | 3.8 | 2.8 | 1.7 | 2.8 | 3.1 | 2.6 | 1.9 | 3.2 | 2.3 |

(a) Draw this data on a graph paper.

(b) Is there any correlation between per capita national income and per capita consumer expenditure? If yes, what is your opinion.

**Solution:** By taking an appropriate scale on the $x$ and $y$ axes, the pairs of observations are plotted on a graph paper as shown in Fig. 13.3. The scatter diagram in Fig. 13.3 with straight line represents the relationship between $x$ and $y$ 'fitted' through it.

**Figure 13.3**
Scatter Diagram



***Interpretation:*** Since pairs of values of two variables are very close to a straight line passing through them, therefore it appears that there is a high degree of association between two variable values. The pattern of dotted points also indicates a high degree of linear positive correlation.

### 13.5.2 Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient quantitatively measures the degree of association (relationship) between two variables $x$ and $y$. For a set of $n$ pairs of values of $x$ and $y$, Pearson's correlation coefficient, $r$, is given by

$$r = \frac{\text{Covariance }(x, y)}{\sqrt{\text{var } x}\,\sqrt{\text{var } y}} = \frac{\text{Cov}\,(x, y)}{\sigma_x\,\sigma_y}$$

where $\quad \text{Cov}(x, y) = \frac{1}{n}\Sigma(x - \bar{x})\,(y - \bar{y})$

$$\sigma_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} \quad \leftarrow \text{ standard deviation of sample data on variable } x$$

$$\sigma_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}} \quad \leftarrow \text{ standard deviation of sample data on variable } y$$

Substituting values of Cov($x$, $y$), $\sigma_x$ and $\sigma_y$, we have

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}} \qquad (13\text{-}1)$$

### Step Deviation Method for Ungrouped Data

If actual mean values of variables $x$ and $y$ are in fraction, then calculation of Pearson's correlation coefficient can be simplified by taking deviations, $d_x = x - \text{A}$ and $d_y = y - \text{B}$, of $x$ and $y$ values from their assumed means A and B, respectively. The formula (13-1) becomes

$$r = \frac{n\,\Sigma d_x\, d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\,\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\,\Sigma d_y^2 - (\Sigma d_y)^2}} \qquad (13\text{-}2)$$

### Step Deviation Method for Grouped Data

If values of variables $x$ and $y$ values are classified into a frequency distribution, then formula (13-2) is modified as

$$r = \frac{n\,\Sigma fd_x\, d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\,\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{n\,\Sigma fd_y^2 - (\Sigma fd_y)^2}} \qquad (13\text{-}3)$$

### Assumptions for Using Pearson's Correlation Coefficient

1. Pearson's correlation coefficient is used only when both variables $x$ and $y$ are measured on an interval or a ratio scale.
2. Pearson's correlation coefficient is used only when two variables $x$ and $y$ are linearly related.

### Advantages and Disadvantages of Pearson's Correlation Coefficient

The numerical value of correlation coefficient between –1 and 1 indicates the strength as well as the direction (positive or negative) of association between two variables. Few limitations of Pearson's method are as follows:

1. Pearson's correlation coefficient is used only when two variables $x$ and $y$ are linearly related.
2. The value of the coefficient is unduly affected by the extreme values of two variable values.
3. Comparatively, the computational time required to calculate the value of Pearson's correlation coefficient, $r$, is lengthy.

## 13.5.3 Probable Error and Standard Error of Coefficient of Correlation

The probable error (PE) of Pearson's correlation coefficient, $r$, indicates the extent to which its value depends on the condition of random sampling. If $r$ is the value of correlation coefficient in a sample of $n$ pairs of observations, then its standard error $\text{SE}_r$ is given by

$$\text{SE}_r = \frac{1-r^2}{\sqrt{n}}$$

The probable error of the coefficient of correlation is calculated as follows:

$$\text{PE}_r = 0.6745\ \text{SE}_r = 0.6745\ \frac{1-r^2}{\sqrt{n}}$$

The amount of $\text{Pe}_r$ is helpful to determine the range, $\rho_r = r \pm \text{Pe}_r$, within which population coefficient of correlation is expected to fall where $\rho_r$(rho) represents population coefficient of correlation.

### Remarks

1. If $r < \text{PE}_r$ then the value of $r$ is not significant, i.e., there is no relationship between two variables of interest.
2. If $r > 6\text{PE}_r$ then value of $r$ is significant, i.e., there exists a relationship between two variables.

**Illustration:** If $r = 0.8$ and $n = 25$, then $\text{PE}_r$ becomes

$$\text{PE}_r = 0.6745\ \frac{1-(0.8)^2}{\sqrt{25}} = 0.6745\ \frac{0.36}{5} = 0.048$$

Thus, the range within which population correlation coefficient ($\rho_r$) should fall is

$$r \pm \text{PE}_r = 0.8 \pm 0.048 \text{ or } 0.752 \le \rho_r \le 0.848$$

## 13.5.4 Coefficient of Determination

The **coefficient of determination**, $r^2$, always has a value between 0 and 1. While squaring the value of correlation coefficient, $r$, the information about the strength of the relationship is retained but the information about the direction is lost. *The value of coefficient of determination represents the proportion (or percentage) of the total variability in the dependent variable, y, that is explained by the independent variable, x.* The proportion (or percentage) of variation in $y$ that $x$ can explain determines precisely the extent or strength of association between two variables $x$ and $y$ (see Chapter 14 for details).

According to Tuttle, *the coefficient of correlation, r has been grossly overrated and is used entirely too much. Its square, coefficient of determination $r^2$, is a much more useful measure of the linear covariance of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between two correlated variables.*

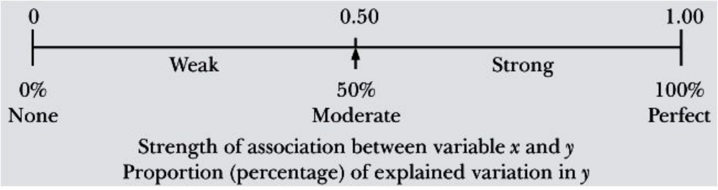
**Coefficient of determination:** A statistical measure of the proportion of the variation in the dependent variable that is explained by independent variable.

### Interpretation of Coefficient of Determination

The knowledge of coefficient of determination is helpful in interpreting the strength of association in terms of percentage between two variables. Figure 13.4 illustrates proportion (percentage) of explained variation in the value of dependent variable, $y$.

- If $r^2 = 0$, then no variation in $y$ due to *variation in values of x*. That is, there is *no association* between $x$ and $y$.
- If $r^2 = 1$, then entire variation in $y$ is due to *variation in values of x*. That is, there is *perfect association* between $x$ and $y$.
- If $0 \le r^2 \le 1$, then degree of variation in $y$ due to *variation in values of x* depends on the value of $r^2$. Value of $r^2$ close to zero shows low proportion of variation in $y$ due to variation in values of $x$. On the other hand, value of $r^2$ close to one shows that the entire variation in $y$ is due to *variation in values of x*.

**Figure 13.4**
Interpretation of Coefficient of Determination


0
0.50
1.00
Weak
Strong
0%
None
50%
Moderate
100%
Perfect
Strength of association between variable x and y
Proportion (percentage) of explained variation in y

Mathematically, the coefficient of determination is determined as

$$r^2 = 1 - \frac{\text{Explained variability in } y}{\text{Total variability in } y}$$

$$= 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{n\Sigma y^2 - a\Sigma y - b\Sigma xy}{n\Sigma y^2 - (\bar{y})^2}$$

where $\hat{y} = a + bx$ is the estimated value of $y$ for given values of $x$.

For example, let correlation between variable $x$ (height) and variable $y$ (weight) be $r = 0.70$. Then the coefficient of determination $r^2 = 0.49$ or 49 per cent implies that only 49 per cent of the variations (changes) in value of variable $y$(weight) is due to variable $x$(height). The remaining 51 per cent of the variations may be due to other factors, say tendency to eat fatty foods.

It is important to know that the 'variability' refers to the dispersion of values of variable, $y$, around its mean value. The greater the correlation coefficient, the greater the coefficient of determination, and the variability in dependent variable can be accounted for in terms of independent variable.

**Example 13.2:** The following table gives indices of industrial production and number of registered unemployed people (in lakh). Calculate the value of the correlation coefficient.

| Year | : | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|---|---|---|---|
| Index of Production | : | 100 | 102 | 104 | 107 | 105 | 112 | 103 | 99 |
| Number Unemployed | : | 15 | 12 | 13 | 11 | 12 | 12 | 19 | 26 |

**Solution:** Calculations of Karl Pearson's correlation coefficient are shown below:

| *Year* | *Production* $x$ | $dx = (x - \bar{x})$ | $d_x^2$ | *Unemployed* $y$ | $d_y = (y - \bar{y})$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|---|
| 1991 | 100 | −4 | 16 | 15 | 0 | 0 | 0 |
| 1992 | 102 | −2 | 4 | 12 | −3 | 9 | +6 |
| 1993 | 104 | 0 | 0 | 13 | −2 | 4 | 0 |
| 1994 | 107 | +3 | 9 | 11 | −4 | 16 | −12 |
| 1995 | 105 | +1 | 1 | 12 | −3 | 9 | −3 |
| 1996 | 112 | +8 | 64 | 12 | −3 | 9 | −24 |
| 1997 | 103 | −1 | 1 | 19 | +4 | 16 | −4 |
| 1998 | 99 | −5 | 25 | 26 | +11 | 121 | −55 |
| Total | 832 | 0 | 120 | 120 | 0 | 184 | −92 |

$$\bar{x} = \frac{\Sigma x}{n} = \frac{832}{8} = 104, \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{120}{8} = 15$$

Applying the formula, $$r = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n \Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n \Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{8 \times -92}{\sqrt{8 \times 120}\sqrt{8 \times 184}} = \frac{-92}{10.954 \times 13.564}$$

$$= \frac{-92}{148.580} = -0.619$$

***Interpretation:*** Since coefficient of correlation $r = -0.619$ is moderately negative, it indicates that there is a moderately large inverse correlation between the two variables. Hence, we conclude that as the production index increases, the number of unemployed decreases and vice versa.

**Example 13.3:** The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

| Size-group | : | 15–16 | 16–17 | 17–18 | 18–19 | 19–20 | 20–21 |
|---|---|---|---|---|---|---|---|
| No. of items | : | 200 | 270 | 340 | 360 | 400 | 300 |
| No. of defective items | : | 150 | 162 | 170 | 180 | 180 | 114 |

[*Delhi Univ., B.Com., 2007*]

**Solution:** Let group size be denoted by variable $x$ and number of defective items by variable $y$. Calculations for Karl Pearson's correlation coefficient are shown below:

| *Size–<br>Group* | *Mid-value*<br>*m* | $d_x = m - 17.5$ | $d_x^2$ | *Percent of*<br>*Defective Items* | $d_y = y - 50$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|---|
| 15–16 | 15.5 | −2 | 4 | 75 | +25 | 625 | −50 |
| 16–17 | 16.5 | −1 | 1 | 60 | +10 | 100 | −10 |
| 17–18 | 17.5 | 0 | 0 | 50 | 0 | 0 | 0 |
| 18–19 | 18.5 | +1 | 1 | 50 | 0 | 0 | 0 |
| 19–20 | 19.5 | +2 | 4 | 45 | −5 | 25 | −10 |
| 20–21 | 20.5 | +3 | 9 | 38 | −12 | 144 | −36 |
| | | 3 | 19 | | 18 | 894 | −106 |

Substituting values in the formula of Karl Pearson's correlation coefficient $r$, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\ \sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{6 \times -106 - 3 \times 18}{\sqrt{6 \times 19 - (3)^2}\ \sqrt{6 \times 894 - (18)^2}} = \frac{-636 - 54}{\sqrt{105}\ \sqrt{5040}}$$

$$= -\frac{690}{727.46} = -0.949$$

***Interpretation:*** Since value of $r$ is negative, and is close to –1, association (relationship) between $x$(size group) and $y$(percent of defective items) is moderate and negative. Hence, it may be concluded that when size of group increases, the number of defective items decreases and vice versa.

**Example 13.4:** The following data relate to age of employees and the number of days they reported sick in a month.

| Employees : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Age : | 30 | 32 | 35 | 40 | 48 | 50 | 52 | 55 | 57 | 61 |
| Sick days : | 1 | 0 | 2 | 5 | 2 | 4 | 6 | 5 | 7 | 8 |

Calculate Karl Pearson's coefficient of correlation and interpret it.

[*Kashmir Univ., B.Com., 2005*]

**Solution:** Let age and sick days be represented by variables $x$ and $y$, respectively. Calculations for value of correlation coefficient are shown below:

| *Age*<br>$x$ | $dx = x - \bar{x}$ | $d_x^2$ | *Sick days*<br>$y$ | $d_y = y - \bar{y}$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 30 | −16 | 256 | 1 | −3 | 9 | 48 |
| 32 | −14 | 196 | 0 | −4 | 16 | 56 |
| 35 | −11 | 121 | 2 | −2 | 4 | 22 |
| 40 | −6 | 36 | 5 | 1 | 1 | −6 |
| 48 | 2 | 4 | 2 | −2 | 4 | −4 |
| 50 | 4 | 16 | 4 | 0 | 0 | 0 |
| 52 | 6 | 36 | 6 | 2 | 4 | 12 |
| 55 | 9 | 81 | 5 | 1 | 1 | 9 |
| 57 | 11 | 121 | 7 | 3 | 9 | 33 |
| 61 | 15 | 225 | 8 | 4 | 16 | 60 |
| 460 | 0 | 1092 | 40 | 0 | 64 | 230 |

$$\bar{x} = \frac{\Sigma x}{n} = \frac{460}{10} = 46, \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{40}{10} = 4$$

Substituting values in the formula of Karl Pearson's correlation coefficient $r$, we have

$$r = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10 \times 230}{\sqrt{10 \times 1092}\sqrt{10 \times 64}}$$

$$= \frac{230}{264.363} = 0.870$$

***Interpretation:*** Since value of $r$ is positive, therefore age of employees and number of sick days are positively correlated to a high degree. Hence, we may conclude that as the age of an employee increases, he is likely to go on sick leave more often than others.

**Example 13.5:** The following table shows the frequency, according to the marks, obtained by 67 students in an intelligence test. Measure the degree of relationship between age and marks:

| | *Age in years* | | | | *Total* |
|---|---|---|---|---|---|
| *Test Marks* | *18* | *19* | *20* | *21* | |
| 200 – 250 | 4 | 4 | 2 | 1 | 11 |
| 250 – 300 | 3 | 5 | 4 | 2 | 14 |
| 300 – 350 | 2 | 6 | 8 | 5 | 21 |
| 350 – 400 | 1 | 4 | 6 | 10 | 21 |
| Total | 10 | 19 | 20 | 18 | 67 |

[*Allahabad Univ., B.Com., 2007*]

**Solution:** Let age of students and marks obtained by them be represented by variables $x$ and $y$, respectively. Calculations for correlation coefficient for this bivariate data are shown below:

| | | *Age in years* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $x$ | *18* | *19* | *20* | *21* | | | | |
| | $d_x$ | *−1* | *0* | *1* | *2* | *Total, f* | $fd_y$ | $fd_y^2$ | $fd_xd_y$ |
| $y$ | $d_y$ | | | | | | | | |
| 200–250 | −1 | 4<br>4 | 0<br>4 | −2<br>2 | −2<br>1 | 11 | −11 | 11 | 0 |
| 250–300 | 0 | 0<br>3 | 0<br>5 | 0<br>4 | 0<br>2 | 14 | 0 | 0 | 0 |
| 300–350 | 1 | −2<br>2 | 0<br>6 | 8<br>8 | 10<br>5 | 21 | 21 | 21 | 16 |
| 350–400 | 2 | −2<br>1 | 0<br>4 | 12<br>6 | 40<br>10 | 21 | 42 | 84 | 50 |
| | *Total, f* | 10 | 19 | 20 | 18 | $n$ = 67 | $\Sigma fd_y$ = 52 | $\Sigma fd_y^2$ = 116 | $\Sigma fd_xd_y$ = 66 |
| | $fd_x$ | −10 | 0 | 20 | 36 | $\Sigma fd_x$ = 46 | | | |
| | $fd_x^2$ | 10 | 0 | 20 | 72 | $\Sigma fd_x{}^2$ = 102 | | | |
| | $fd_xd_y$ | 0 | 0 | 18 | 48 | $\Sigma fd_xd_y$ = 66 | | | |

where $d_x = x - 19$, and $d_y = (m - 275)/50$

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\ \sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}} = \frac{67 \times 66 - 46 \times 52}{\sqrt{67 \times 102 - (46)^2}\ \sqrt{67 \times 116 - (52)^2}}$$

$$= \frac{4422 - 2392}{\sqrt{6834 - 2116}\ \sqrt{7772 - 2704}} = \frac{2030}{\sqrt{4718}\ \sqrt{5068}} = \frac{2030}{68.688 \times 71.19} = 0.415$$

***Interpretation*:** Since the value of $r$ is positive, therefore age of students and marks obtained in an intelligence test are positively correlated to the extent of 0.415. Hence, we may conclude that as the age of students increases, score of marks in intelligence test also increases.

**Example 13.6:** Calculate the coefficient of correlation from the following bivariate frequency distribution:

| *Sales Revenue (₹ in lakh)* | *Advertising Expenditure (₹ in '000)* | | | | *Total* |
|---|---|---|---|---|---|
| | *5–10* | *10–15* | *15–20* | *20–25* | |
| 75–125 | 4 | 1 | — | — | 5 |
| 125–175 | 7 | 6 | 2 | 1 | 16 |
| 175–225 | 1 | 3 | 4 | 2 | 10 |
| 225–275 | 1 | 1 | 3 | 4 | 9 |
| Total | 13 | 11 | 9 | 7 | 40 |

[*Delhi Univ., MBA,2005*]

**Solution:** Let advertising expenditure and sales revenue be represented by variables $x$ and $y$, respectively. The calculations for correlation coefficient are shown below:

| | | | *Advertising Expenditure* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *5–10* | *10–15* | *15–20* | *20–25* | | | | |
| | | | *7.5* | *12.5* | *17.5* | *22.5* | | | | |
| | | | *–1* | *0* | *1* | *2* | *Total, f* | $fd_y$ | $fd_y^2$ | $fd_xd_y$ |
| *Revenue* *y* | *Mid-value* *(m)* | $d_y$ | | | | | | | | |
| 75–125 | 100 | –2 | 8<br>4 | 0<br>1 | 0<br>— | 0<br>— | 5 | –10 | 20 | 8 |
| 125–175 | 150 | –1 | 7<br>7 | 0<br>6 | –2<br>2 | –2<br>1 | 16 | –16 | 16 | 3 |
| 175–225 | 200 | 0 | 0<br>1 | 0<br>3 | 0<br>4 | 0<br>2 | 10 | 0 | 0 | 0 |
| 225–275 | 250 | 1 | –1<br>1 | 0<br>1 | 3<br>3 | 8<br>4 | 9 | 9 | 9 | 10 |
| | | *Total, f* | 13 | 11 | 9 | 7 | $n$ = 40 | $\Sigma d_y$ = –17 | $\Sigma d_y^2$ = 45 | $\Sigma fd_xd_y$ = 21 |
| | | $fd_x$ | –13 | 0 | 9 | 14 | $\Sigma fd_x$ = 10 | | | |
| | | $fd_x^2$ | 13 | 0 | 9 | 28 | $\Sigma fd_x{}^2$ = 50 | | | |
| | | $fd_xd_y$ | 14 | 0 | 1 | 6 | $\Sigma fd_xd_y$ = 21 | | | |

where $d_x = (m - 12.5)/5$, and $d_y = (m - 200)/50$.

Substituting values in the formula of Karl Pearson's correlation coefficient, we have

$$r = \frac{n\,\Sigma\, fd_x d_y - (\Sigma\, fd_x)\,(\Sigma\, fd_y)}{\sqrt{n\,\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2}\;\sqrt{n\,\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2}} = \frac{40 \times 21 - 10 \times -17}{\sqrt{40 \times 50 - (10)^2}\;\sqrt{40 \times 45 - (-17)^2}}$$

$$= \frac{840 + 170}{\sqrt{1900}\;\sqrt{1511}} = \frac{1010}{1694.373} = 0.596$$

***Interpretation***: Since the value of $r$ is positive, advertising expenditure and sales revenue are positively correlated to the extent of 0.596. Hence, we may conclude that as expenditure on advertising increases, the sales revenue also increases.

**Example 13.7:** A computer while calculating the correlation coefficient between two variables $x$ and $y$ from 25 pairs of observations obtained the following results:

$$n = 25,\ \Sigma x = 125,\ \Sigma x^2 = 650 \text{ and } \Sigma y = 100,\ \Sigma y^2 = 460,\ \Sigma xy = 508$$

It was, however, discovered at the time of checking that he had copied down two pairs of observations as

| $x$ | $y$ |
|---|---|
| 6 | 14 |
| 8 | 6 |

instead of

| $x$ | $y$ |
|---|---|
| 8 | 12 |
| 6 | 8 |

Obtain the correct value of correlation coefficient between $x$ and $y$.

*[MD Univ., M.Com., 2006; Kumaon Univ., MBA, 2007]*

**Solution:** The corrected values of variables required for the formula of Pearson's correlation coefficient are determined as follows:

$$\begin{aligned}
\text{Correct } \Sigma x &= 125 - (6 + 8 - 8 - 6) = 125\\
\text{Correct } \Sigma y &= 100 - (14 + 6 - 12 - 8) = 100\\
\text{Correct } \Sigma x^2 &= 650 - \{(6)^2 + (8)^2 - (8)^2 - (6)2\}\\
&= 650 - \{36 + 64 - 64 - 36\} = 650\\
\text{Correct } \Sigma y^2 &= 460 - \{(14)^2 + (6)^2 - (12)^2 - (8)^2\}\\
&= 460 - \{196 + 36 - 144 - 64\} = 436\\
\text{Correct } \Sigma xy &= 508 - \{(6\times 14) + (8\times 6) - (8\times 12) - (6 \times 8)\}\\
&= 508 - \{84 - 48 - 96 - 48\} = 520
\end{aligned}$$

Applying the formula

$$r = \frac{n\,\Sigma xy - (\Sigma x)\,(\Sigma y)}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\;\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}} = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\;\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{13{,}000 - 12{,}500}{\sqrt{625}\;\sqrt{900}} = \frac{500}{25 \times 30} = 0.667$$

Thus, the correct value of correlation coefficient between $x$ and $y$ is 0.667.

## Self-practice Problems 13A

**13.1** Making use of the data summarized below, calculate the coefficient of correlation.

| *Case* | $x_1$ | $x_2$ | *Case* | $x_1$ | $x_2$ |
|---|---|---|---|---|---|
| A | 10 | 9 | E | 12 | 11 |
| B | 6 | 4 | F | 13 | 13 |
| C | 9 | 6 | G | 11 | 8 |
| D | 10 | 9 | H | 9 | 4 |

**13.2** Find the correlation coefficient by Karl Pearson's method between $x$ and $y$ and interpret its value.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ : | 57 | 42 | 40 | 33 | 42 | 45 | 42 | 44 | 40 | 56 | 44 | 43 |
| $y$ : | 10 | 60 | 30 | 41 | 29 | 27 | 27 | 19 | 18 | 19 | 31 | 29 |

**13.3** Calculate the coefficient of correlation from the following data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ : | 100 | 200 | 300 | 400 | 500 | 600 | 700 |
| $y$ : | 30 | 50 | 60 | 80 | 100 | 110 | 130 |

**13.4** Calculate the coefficient of correlation between $x$ and $y$ from the following data and calculate the probable errors. Assume 69 and 112 as the mean value for $x$ and $y$, respectively.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ : | 78 | 89 | 99 | 60 | 50 | 79 | 68 | 61 |
| $y$ : | 125 | 137 | 156 | 112 | 107 | 136 | 123 | 108 |

**13.5** Find the coefficient of correlation from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cost : | 39 | 65 | 62 | 90 | 82 | 75 | 25 | 98 | 36 | 78 |
| Sales : | 47 | 53 | 58 | 86 | 62 | 68 | 60 | 91 | 51 | 84 |

[*Madras Univ., B.Com., 2005*]

**13.6** Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below. Also calculate the probable error and comment on the value:

| | | | | | | |
|---|---|---|---|---|---|---|
| Age : | 20 | 21 | 22 | 23 | 24 | 25 |
| No. of students : | 500 | 400 | 300 | 240 | 200 | 160 |
| Regular players : | 400 | 300 | 180 | 96 | 60 | 24 |

[*HP Univ., MBA, 2005*]

**13.7** Find the coefficient of correlation between age and the sum assured (in 1000 ₹) from the following table:

| *Age Group (years)* | *Sum Assured (in ₹)* | | | | |
|---|---|---|---|---|---|
| | *10* | *20* | *30* | *40* | *50* |
| 20–30 | 4 | 6 | 3 | 7 | 1 |
| 30–40 | 2 | 8 | 15 | 7 | 1 |
| 40–50 | 3 | 9 | 12 | 6 | 2 |
| 50–60 | 8 | 4 | 2 | — | — |

[*Delhi Univ., MBA, 2007*]

**13.8** Family income and its percentage spent on food in the case of one hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and interpret its value.

| *Food* | *Monthly Family Income (₹)* | | | | |
|---|---|---|---|---|---|
| *Expenditure (in percent)* | *2000–3000* | *3000–4000* | *4000–5000* | *5000–6000* | *6000–7000* |
| 10–15 | — | — | — | 3 | 7 |
| 15–20 | — | 4 | 9 | 4 | 3 |
| 20–25 | 7 | 6 | 12 | 5 | — |
| 25–30 | 3 | 10 | 19 | 8 | — |

[*Delhi Univ., MBA, 2006*]

**13.9** With the following data in 6 cities, calculate Pearson's coefficient of correlation between the density of population and death rate:

| *City* | *Area in Kilometres* | *Population (in '000)* | *No. of Deaths* |
|---|---|---|---|
| A | 150 | 30 | 300 |
| B | 180 | 90 | 1440 |
| C | 100 | 40 | 560 |
| D | 60 | 42 | 840 |
| E | 120 | 72 | 1224 |
| F | 80 | 24 | 312 |

[*Sukhadia Univ., B.Com.*, 2006]

**13.10** The coefficient of correlation between two variables $x$ and $y$ is 0.3. The covariance is 9. The variance of $x$ is 16. Find the standard deviation of $y$ series.

## Hints and Answers

**13.1** $\bar{x}_1 = 80/8 = 10,\ \bar{x}_2 = 64/8 = 8;$

$$r = \frac{43}{\sqrt{32}\,\sqrt{72}} = 0.896$$

**13.2** $r = -0.554$ **13.3** $r = 0.997$

**13.4** $r = 0.014$ **13.5** $r = 0.780$

**13.6** $r = 0.005$ **13.7** $r = -0.256$

**13.8** $r = -0.438$ **13.9** $r = 0.988$

**13.10** Given $\sigma_x = \sqrt{16} = 4;\quad r = \dfrac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$

or $0.3 = \dfrac{9}{4\sigma_y}$ or $\sigma_y = 7.5.$

### 13.5.5 Spearman's Rank Correlation Coefficient

In 1904, a British psychologist Charles Edward Spearman developed a method to measure the statistical association (relationship) between two variables, say $x$ and $y$, when only *ordinal (or rank) data* are available. This implies that **Spearman's rank correlation coefficient** method is applied in a situation where quantitative measure of qualitative factors such as judgment, brands personalities, beauty, intelligence, honesty, efficiency, TV programmes, leadership, colour and taste cannot be fixed but individual observations can be arranged in a definite order (or rank). The ranking is done by using a set of ordinal rank numbers with 1 for the individual observation ranked first; 2 for the individual observation ranked second and so on either in terms of quantity or quality. Mathematically, Spearman's rank correlation coefficient is defined as

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad (13\text{-}4)$$

where R is rank correlation coefficient; $R_1$ is rank of observations with respect to first variable; $R_2$ is the rank of observations with respect to second variable; $d = R_1 - R_2$ is difference in a pair of ranks; and $n$ is the number of paired observations or individuals being ranked.

The number '6' in the formula as scaling device ensures that the possible range of R is from –1 to 1.

### Advantages and Disadvantages of Spearman's Correlation Coefficient Method

*Advantages*

(i) This method is easy to understand and its application is simpler than Pearson's method.

(ii) This method is useful for correlation analysis when variables are expressed in qualitative terms.

(iii) This method is developed to measure the statistical association (relationship) between two variables, say $x$ and $y$, when only *ordinal* (*or rank*) *data* are available.

*Disadvantages*

(i) Values of both the variables are assumed to be normally distributed and describing a linear relationship rather than non-linear relationship.

(ii) A large computational time is required when pairs of values of two variables exceed 30.

(iii) This method cannot be applied on grouped data to measure the association between two variables.

### Case I: When Ranks Are Given

If observations in a data set are already arranged in a particular order (rank), then take the differences in pairs of observations to determine the difference, $d$. Square these differences and obtain the total. Apply the formula to calculate Spearman's correlation coefficient.

**Example 13.8:** The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of the squares of the differences in ranks is given to be 48, then find the values of $n$.

***Solution:*** Apply the formula of Spearman's correlation coefficient:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Given R = 0.143, $\Sigma d^2 = 48$ and $n=7$. Substituting values in the formula, we get

$$0.143 = 1 - \frac{6 \times 48}{n(n^2 - 1)} = 1 - \frac{288}{n^3 - n}$$

$$0.143(n^3 - n) = (n^3 - n) - 288$$
$$n^3 - n - 336 = 0 \quad \text{or} \quad (n - 7)(n^2 + 7n + 48) = 0$$

This implies that either $n - 7 = 0$, that is, $n = 7$ or $n^2 + 7n + 48 = 0$. But $n^2 + 7n + 48 = 0$ on simplification gives undesirable value of $n$ because its discriminant $b^2 - 4ac$ is negative. Hence, $n = 7$.

**Example 13.9:** The ranks of 15 students in two subjects A and B are given below. The two numbers within brackets denote the ranks of a student in A and B subjects, respectively.

(1, 10), (2, 7), (3, 2), (4, 6), (5, 4), (6, 8), (7, 3), (8, 1),
(9, 11), (10, 15), (11, 9), (12, 5), (13, 14), (14, 12), (15, 13)

Find Spearman's rank correlation coefficient. [*Sukhadia Univ., MBA, 2006*]

**Solution:** Since ranks of students with respect to their performance in two subjects are given, calculations for rank correlation coefficient are shown below:

| *Rank in A* $R_1$ | *Rank in B* $R_2$ | *Difference* $d = R_1 - R_2$ | $d^2$ |
|---|---|---|---|
| 1 | 10 | −9 | 81 |
| 2 | 7 | −5 | 25 |
| 3 | 2 | 1 | 1 |
| 4 | 6 | −2 | 4 |
| 5 | 4 | 1 | 1 |
| 6 | 8 | −2 | 4 |
| 7 | 3 | 4 | 16 |
| 8 | 1 | 7 | 49 |
| 9 | 11 | −2 | 4 |
| 10 | 15 | −5 | 25 |
| 11 | 9 | 2 | 4 |
| 12 | 5 | 7 | 49 |
| 13 | 14 | −1 | 1 |
| 14 | 12 | 2 | 4 |
| 15 | 13 | 2 | 4 |
| | | | $\Sigma d^2 = 272$ |

Apply the formula, $$R = 1 - \Sigma\frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 272}{15\{(15)^2 - 1\}}$$

$$= 1 - \frac{1632}{3360} = 1 - 0.4857 = 0.5143$$

The result shows a moderate positive correlation between performances of students in two subjects.

**Example 13.10:** There are 12 clerks working in a office. The long-serving clerks feel that they should get seniority increment based on length of service built into their salary structure. Based on assessment of their efficiency by the HR department a ranking of efficiency was developed. The ranking of efficiency together with a ranking of their length of service is as follows:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ranking according to length of service : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Ranking according to efficiency : | 2 | 3 | 5 | 1 | 9 | 10 | 11 | 12 | 8 | 7 | 6 | 4 |

Do the data support the clerks' claim for seniority increment?

[*Sukhadia Univ., MBA, 2000*]

**Solution:** Since ranks are already given, calculations for rank correlation coefficient are shown below:

| *Rank According to Length of Service* $R_1$ | *Rank According to Efficiency* $R_2$ | *Difference* $d = R_1 - R_2$ | $d^2$ |
|---|---|---|---|
| 1 | 2 | –1 | 1 |
| 2 | 3 | –1 | 1 |
| 3 | 5 | –2 | 4 |
| 4 | 1 | 3 | 9 |
| 5 | 9 | –4 | 16 |
| 6 | 10 | –4 | 16 |
| 7 | 11 | –4 | 16 |
| 8 | 12 | –4 | 16 |
| 9 | 8 | 1 | 1 |
| 10 | 7 | 3 | 9 |
| 11 | 6 | 5 | 25 |
| 12 | 4 | 8 | 64 |
|  |  |  | $\Sigma d^2 = 178$ |

Applying the formula,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 178}{12(144 - 1)} = 1 - \frac{1068}{1716} = 0.378$$

The result shows a low degree positive correlation between length of service and efficiency, the claim of the clerks for a seniority increment based on length of service may not be justified.

**Example 13.11:** Ten competitors in a beauty contest are ranked by three judges in the following order:

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Judge 1: | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| Judge 2: | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| Judge 3: | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Use the rank correlation coefficient to determine which pair of judges has the nearest approach for judgment of beauty. [*MD Univ., MBA, 2004*]

**Solution:** The pair of judges who have the nearest approach for judgment of beauty can be obtained in ${}^3C_2 = 3$ ways as follows:

(i) Judge 1 and judge 2.

(ii) Judge 2 and judge 3.

(iii) Judge 3 and judge 1.

Calculations after comparing the ranking of judges are shown below:

| *Judge 1* $R_1$ | *Judge 2* $R_2$ | *Judge 3* $R_3$ | $d_1^2 = (R_1 - R_2)^2$ | $d_2^2 = (R_2 - R_3)^2$ | $d_3^2 = (R_3 - R_1)^2$ |
|---|---|---|---|---|---|
| 1 | 3 | 6 | 4 | 9 | 25 |
| 6 | 5 | 4 | 1 | 1 | 4 |
| 5 | 8 | 9 | 9 | 1 | 16 |
| 10 | 4 | 8 | 36 | 16 | 4 |
| 3 | 7 | 1 | 16 | 36 | 4 |
| 2 | 10 | 2 | 64 | 64 | 0 |
| 4 | 2 | 3 | 4 | 1 | 1 |
| 9 | 1 | 10 | 64 | 81 | 1 |
| 7 | 6 | 5 | 1 | 1 | 4 |
| 8 | 9 | 7 | 1 | 4 | 1 |
|  |  |  | 200 | 214 | 60 |

Applying the formula

$$R_{12} = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10(100-1)} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10(100-1)} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10(100-1)} = 1 - \frac{360}{990} = 0.636$$

Since the correlation coefficient $R_{13} = 0.636$ is highest, the judges 1 and 3 have nearest approach for judgment of beauty.

### Case 2: When Ranks Are Not Given

If observations in a data set are not arranged in a particular order (rank), then ranks are assigned by taking either the highest value or the lowest value as rank one and so on for values of both the variables.

**Example 13.12:** Quotations of index numbers of security prices of a certain joint stock company are given below:

| *Year* | *Debenture Price* | *Share Price* |
|---|---|---|
| 1 | 97.8 | 73.2 |
| 2 | 99.2 | 85.8 |
| 3 | 98.8 | 78.9 |
| 4 | 98.3 | 75.8 |
| 5 | 98.4 | 77.2 |
| 6 | 96.7 | 87.2 |
| 7 | 97.1 | 83.8 |

Use rank correlation method to determine the relationship between debenture prices and share prices. [*Calicut Univ., B.Com., 2005*]

**Solution:** Let us start ranking from the lowest value for both the variables as shown below:

| *Debenture Price* $(x)$ | *Rank* | *Share Price* $(y)$ | *Rank* | *Difference* $d = R_1 - R_2$ | $d^2 = (R_1 - R_2)^2$ |
|---|---|---|---|---|---|
| 97.8 | 3 | 73.2 | 1 | 2 | 4 |
| 99.2 | 7 | 85.8 | 6 | 1 | 1 |
| 98.8 | 6 | 78.9 | 4 | 2 | 4 |
| 98.3 | 4 | 75.8 | 2 | 2 | 4 |
| 98.4 | 5 | 77.2 | 3 | 2 | 4 |
| 96.7 | 1 | 87.2 | 7 | −6 | 36 |
| 97.1 | 2 | 83.8 | 5 | −3 | 9 |
| | | | | | $\Sigma d^2 = 62$ |

Applying the formula,

$$R = 1 - \frac{6\,\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 62}{(7)^3 - 7}$$

$$= 1 - \frac{372}{336} = 1 - 0.107 = -\,0.107$$

The result shows a low degree of negative correlation between the debenture prices and share prices of a certain joint stock company.

**Example 13.13** An economist wanted to find out whether there is any relationship between the unemployment rate in a country and its inflation rate. Data from 7 countries for the year 2009 are given below:

| *Country* | *Unemployment Rate (Per cent)* | *Inflation Rate (Per cent)* |
|---|---|---|
| A | 4.0 | 3.2 |
| B | 8.5 | 8.2 |
| C | 5.5 | 9.4 |
| D | 0.8 | 5.1 |
| E | 7.3 | 10.1 |
| F | 5.8 | 7.8 |
| G | 2.1 | 4.7 |

Find the degree of linear association between unemployment rate in a country and its inflation rate.

**Solution:** Ranking from the lowest value for both the variables as shown below:

| *Unemployment Rate (x)* | *Rank $R_1$* | *Inflation Rate (y)* | *Rank $R_2$* | *Difference $d = R_1 - R_2$* | $d^2 = (R_1 - R_2)^2$ |
|---|---|---|---|---|---|
| 4.0 | 3 | 3.2 | 1 | 2 | 4 |
| 8.5 | 7 | 8.2 | 5 | 2 | 4 |
| 5.5 | 4 | 9.4 | 6 | −2 | 4 |
| 0.8 | 1 | 5.1 | 3 | −2 | 4 |
| 7.3 | 6 | 10.1 | 7 | −1 | 1 |
| 5.8 | 5 | 7.8 | 4 | 1 | 1 |
| 2.1 | 2 | 4.7 | 2 | 0 | 0 |
| | | | | | $\Sigma d^2 = 18$ |

Applying the formula,

$$R = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \times 18}{(7)^3 - (7)} = 1 - \frac{108}{336} = 0.678$$

The result shows a moderately high degree of positive correlation between unemployment rate and inflation rate of seven countries.

### Case 3: When Ranks Are Equal

If more than one observations of equal size are found at the time of ranking observations in the data set by taking either the highest value or lowest value as rank one, then rank to be assigned to individual observations is an average of the ranks that these individual observations deserved. For example, if two observations are ranked equal at third place, then the average rank of (3 + 4)/2 = 3.5 is assigned to these two observations. Similarly, if three observations are ranked equal at third place, then the average rank of (3 + 4 + 5)/3 = 4 is assigned to these three observations.

The modified Spearman rank correlation coefficient formula for such a case is given below:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots\right\}}{n(n^2 - 1)}$$

where $m_i (i = 1, 2, 3, \ldots)$ stands for the number of times an observation is repeated in the data set for both variables.

**Example 13.14:** A financial analyst wanted to find out whether inventory turnover influences any company's earnings per share (in per cent). A random sample of 7 companies listed in a stock exchange was selected and the following data was recorded for each.

| *Company* | *Inventory Turnover (Number of Times)* | *Earnings per Share (Per cent)* |
|---|---|---|
| A | 4 | 11 |
| B | 5 | 9 |
| C | 7 | 13 |
| D | 8 | 7 |
| E | 6 | 13 |
| F | 3 | 8 |
| G | 5 | 8 |

Find the strength of association between inventory turnover and earnings per share. Interpret this finding.

**Solution:** Ranking from lowest value for both the variables. Since observations of equal size are found at the time of ranking in the data set, therefore rank to be assigned to repeat observations is an average of the ranks that these individual observations deserved as shown below.

| *Inventory Turnover (x)* | *Rank* $R_1$ | *Earnings Per Share (y)* | *Rank* $R_2$ | *Difference* $d = R_1 - R_2$ | $d^2 = (R_1 - R_2)^2$ |
|---|---|---|---|---|---|
| 4 | 2 | 11 | 5 | −3.0 | 9.00 |
| 5 | 3.5 | 9 | 4 | −0.5 | 0.25 |
| 7 | 6 | 13 | 6.5 | 0.5 | 0.25 |
| 8 | 7 | 7 | 1 | 6.0 | 36.00 |
| 6 | 5 | 13 | 6.5 | −1.5 | 2.25 |
| 3 | 1 | 8 | 2.5 | −1.5 | 2.25 |
| 5 | 3.5 | 8 | 2.5 | 1.0 | 1.00 |
| | | | | | $\Sigma d^2 = 51$ |

If may be noted that a value 5 of variable *x* is repeated twice ($m_1 = 2$) and values 8 and 13 of variable *y* is also repeated twice, so $m_2 = 2$ and $m_3 = 2$. Applying the formula:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{51 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right\}}{7(49 - 1)}$$

$$= 1 - \frac{6\{51 + 0.5 + 0.5 + 0.5\}}{336} = 1 - 0.9375 = 0.0625$$

The result shows a very week positive association between inventory turnover and earning per share.

**Example 13.15:** Obtain the rank correlation coefficient between the variables *x* and *y* from the following pairs of observed values.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *x* : | 50 | 55 | 65 | 50 | 55 | 60 | 50 | 65 | 70 | 75 |
| *y* : | 110 | 110 | 115 | 125 | 140 | 115 | 130 | 120 | 115 | 160 |

[*Mangalore Univ., B.Com., 2005*]

**Solution:** Ranking from lowest value for both the variables. Since observations of equal size are found at the time of ranking in the data set, therefore rank to be assigned to repeat observations is an average of the ranks that these individual observations deserved as shown below.

| *Variable* $x$ | *Rank* $R_1$ | *Variable* $y$ | *Rank* $R_2$ | *Difference* $d = R_1 - R_2$ | $d^2 = (R_1 - R_2)^2$ |
|---|---|---|---|---|---|
| 50 | 2 | 110 | 1.5 | 0.5 | 0.25 |
| 55 | 4.5 | 110 | 1.5 | 3.0 | 9.00 |
| 65 | 7.5 | 115 | 4 | 3.5 | 12.25 |
| 50 | 2 | 125 | 7 | –5.0 | 25.00 |
| 55 | 4.5 | 140 | 9 | –4.5 | 20.25 |
| 60 | 6 | 115 | 4 | 2.0 | 4.00 |
| 50 | 2 | 130 | 8 | –6.0 | 36.00 |
| 65 | 7.5 | 120 | 6 | 1.5 | 2.25 |
| 70 | 9 | 115 | 4 | 5.0 | 25.00 |
| 75 | 10 | 160 | 10 | 0.0 | 00.00 |
| | | | | | 134.00 |

It may be noted that for variable $x$, 50 is repeated thrice ($m_1 = 3$), 55 is repeated twice ($m_2 = 2$), and 65 is repeated twice ($m_3 = 2$). Also for variable $y$, 110 is repeated twice ($m_4 = 2$) and 115 thrice ($m_5 = 3$). Applying the formula:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \frac{1}{12}(m_4^3 - m_4) + \frac{1}{12}(m_5^3 - m_5)\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10(100 - 1)}$$

$$= 1 - \frac{6\,[134 + 2 + 0.5 + 0.5 + 0.5 + 2]}{990} = 1 - \frac{6 \times 139.5}{990} = 1 - \frac{837}{990}$$

$= 1 - 0.845 = 0.155.$

The result shows a weak positive association between variables $x$ and $y$.

### 13.5.6 Method of Least Squares

The method of least squares to calculate the correlation coefficient requires the values of regression coefficients $b_{xy}$ and $b_{yx}$, so that

$$r = \sqrt{b_{xy} \times b_{yx}}$$

In other words, correlation coefficient is the geometric mean of two regression coefficients (see Chapter 14 for details).

### 13.5.7 Auto Correlation Coefficient

The auto correlation coefficient describes mutual dependence between values of the same variable but at different time periods. Thus, it provides information on how a variable relates to itself for a specific time lag. The difference in the period before a cause-and-effect relationship is established is referred to as ***lead time or lag***. While computing the correlation, the time gap must be considered; otherwise misleading conclusions may be arrived at. For example, the decrease or increase in supply of a commodity may not immediately reflect on its price, it may take some lead time or time lag.

The formula for auto-correlation coefficient at time lag *k* is stated as:

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where *k* is length of time lag; *n* is the number of observations; and $\bar{x}$ is the mean of all observations.

**Example 13.16:** The monthly sales of a product, in thousands of units, in the last 6 months are given below:

| | | | | | | |
|---|---|---|---|---|---|---|
| Month : | 1 | 2 | 3 | 4 | 5 | 6 |
| Sales : | 1.8 | 2.5 | 3.1 | 3.0 | 4.2 | 3.4 |

Compute the auto-correlation coefficient up to lag 2. What conclusion can be derived from these values regarding the presence of a trend in the data?

**Solution:** The calculations for auto-correlation coefficient are shown below:

| *Time* | *Sales* $(x)$ | $x_1$ *= One Time Lag Variable Constructed From x* | $x_2$ *= Two Time Lags Variable Constructed From x* |
|---|---|---|---|
| 1 | 1.8 | 2.5 | 3.1 |
| 2 | 2.5 | 3.1 | 3.0 |
| 3 | 3.1 | 3.0 | 4.2 |
| 4 | 3.0 | 4.2 | 3.4 |
| 5 | 4.2 | 3.4 | — |
| 6 | 3.4 | — | — |

For $k = 1,\ \bar{x} = \frac{1}{6}(1.8 + 2.5 + \ldots + 3.4) = 3$

$$r_1 = \frac{\{(1.8-3)(2.5-3)+(2.5-3)(3.1-3)+(3.1-3)(3-3) + (3-3)(4.2-3)+(4.2-3)(3.4-3)\}}{(1.8-3)^2+(2.5-3)^2+(3.1-3)^2+(3-3)^2+(4.2-3)^2+(3.4-3)^2}$$

$$= \frac{(-1.2)(-0.5)+(-0.5)(0.1)+(0.1)(0)+(0)(1.2)+(1.2)(0.4)}{1.44+0.25+0.01+0+1.44+0.16}$$

$$= \frac{(0.6-0.5+0.48)}{3.3} = 0.312$$

For $k = 2$

$$r_2 = \frac{(1.8-3)(3.1-3)+(2.5-3)(3-3)+(3.1-3)(4.2-3)+(3-3)(3.4-3)}{(1.8-3)^2+(2.5-3)^2+\ldots+(3.4-3)^2}$$

$$= \frac{(-1.2\times 0.1)+(-0.5\times 0)+(0.1\times 1.2)+(0\times 0.4)}{3.3} = \frac{-0.12+0.12}{3.3} = 0$$

Since the value of *r*1 is positive, it implies that there is a seasonal pattern of 6 months duration and $r_2 = 0$ implies that there is no significant change in sales.

## Self-practice Problems 13B

**13.11** The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.2. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct coefficient of rank correlation.

[*Delhi Univ., B.Com., 2004*]

**13.12** The ranking of 10 students in accordance with their performance in two subjects A and B are as follows:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A: | 6 | 5 | 3 | 10 | 2 | 4 | 9 | 7 | 8 | 1 |
| B: | 3 | 8 | 4 | 9 | 1 | 6 | 10 | 7 | 5 | 2 |

Calculate the rank correlation coefficient and comment on its value.

**13.13** Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges $x$ and $y$ in a certain competitive test as shown below:

| *Student* | *Marks by Judge x* | *Marks by Judge y* |
|---|---|---|
| 1 | 52 | 65 |
| 2 | 53 | 68 |
| 3 | 42 | 43 |
| 4 | 60 | 38 |
| 5 | 45 | 77 |
| 6 | 41 | 48 |
| 7 | 37 | 35 |
| 8 | 38 | 30 |
| 9 | 25 | 25 |
| 10 | 27 | 50 |

**13.14** An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the accountancy and statistics papers, compute the rank correlation coefficient.

| *Applicant* | : | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|---|
| Marks in accountancy | : | 15 | 20 | 28 | 12 | 40 | 60 | 20 | 80 |
| Marks in statistics | : | 40 | 30 | 50 | 30 | 20 | 10 | 30 | 60 |

**13.15** Seven methods of imparting business education were ranked by the MBA students of two universities as follows:

| *Method of Teaching* | : | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Rank by students of Univ. A | : | 2 | 1 | 5 | 3 | 4 | 7 | 6 |
| Rank by students of Univ. B | : | 1 | 3 | 2 | 4 | 7 | 5 | 6 |

Calculate the rank correlation coefficient and comment on its value.

**13.16** An investigator collected the following data with respect to the socio-economic status and severity of respiratory illness.

| *Patient* | : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Socio-economic status (rank) | : | 6 | 7 | 2 | 3 | 5 | 4 | 1 | 8 |
| Severity of illness rank) | : | 5 | 8 | 4 | 3 | 7 | 1 | 2 | 6 |

Calculate the rank correlation coefficient and comment on its value.

**13.17** You are given the following data of marks obtained by 11 students in statistics in two tests, one before and other after special coaching:

| *First Test (Before coaching)* | *Second Test (After coaching)* |
|---|---|
| 23 | 24 |
| 20 | 19 |
| 19 | 22 |
| 21 | 18 |
| 18 | 20 |
| 20 | 22 |
| 18 | 20 |
| 20 | 22 |
| 18 | 20 |
| 17 | 20 |
| 23 | 23 |
| 16 | 20 |
| 19 | 17 |

Do the marks indicate that the special coaching has benefited the students? [*Delhi Univ., M.Com., 2000*]

**13.18** Two departmental managers ranked a few trainees according to their perceived abilities. The ranking are given below:

| *Trainee* | : | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Manager A | : | 1 | 9 | 6 | 2 | 5 | 8 | 7 | 3 | 10 | 4 |
| Manager B | : | 3 | 10 | 8 | 1 | 7 | 5 | 6 | 2 | 9 | 4 |

Calculate an appropriate correlation coefficient to measure the consistency in the ranking.

**13.19** In an office some keyboard operators, who were already ranked on their speed, were also ranked on accuracy by their supervisor. The results were as follows:

| *Operator* | : | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Speed | : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Accuracy | : | 7 | 9 | 3 | 4 | 1 | 6 | 8 | 2 | 10 | 5 |

Calculate the appropriate correlation coefficient between speed and accuracy.

**13.20** The personnel department is interested in comparing the ratings of job applicants when measured by a variety of standard tests. The ratings of 9 applicants on interviews and standard psychological test are shown below:

| *Applicant* | : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Interview | : | 5 | 2 | 9 | 4 | 3 | 6 | 1 | 8 | 7 |
| Standard test | : | 8 | 1 | 7 | 5 | 3 | 4 | 2 | 9 | 6 |

Calculate Spearman's rank correlation coefficient and comment on its value.

## Hints and Answers

**13.11** Given $R = 0.2, n = 10; R = 1 - \dfrac{6\,\Sigma\, d^2}{n\,(n^2-1)}$ or

$$0.2 = 1 - \frac{6\Sigma\ d^2}{10\,(100-1)} \text{ or } \Sigma\, d^2 = 100$$

Correct value of $R = 1 - \dfrac{6 \times 100}{10 \times 99} = 0.394$

**13.12** $R = 1 - \dfrac{6\Sigma d^2}{n\,(n^2-1)} = 1 - \dfrac{6 \times 36}{10(100-1)} = 0.782$

**13.13** $R = 1 - \dfrac{6\,\Sigma\, d^2}{n\,(n^2-1)} = 1 - \dfrac{6 \times 76}{10(100-1)} = 0.539$

**13.14** $R = 1 - \dfrac{6\left\{\Sigma\, d^2 + \dfrac{1}{12}\left(m_1^3 - m_1\right) + \dfrac{1}{12}\left(m_2^3 - m_2\right)\right\}}{n(n^2-1)}$

$$= 1 - \frac{6\left\{81.5 + \dfrac{1}{12}(2^3-2) + \dfrac{1}{12}(3^3-3)\right\}}{8(64-1)} = 0$$

**13.15** R = 0.50 **13.16** R = 0.477

**13.17** R = 0.71 **13.18** R = 0.842

**13.19** R = 0.006 **13.20** R = 0.817

## 13.6 HYPOTHESIS TESTING FOR CORRELATION COEFFICIENT

The sample correlation coefficient is often used as an estimator to test whether the possible strength of association between two random variables in the population exists. In other words, sample correlation coefficient, *r*; is used as an estimator for testing a null hypothesis about true *population correlation coefficient* ρ(Greek letter rho) with the assumption that two random variables, say *x* and *y*, are normally distributed.

### 13.6.1 Hypothesis Testing About Population Correlation Coefficient (Small Sample)

The null hypothesis that a linear relationship between two variables *x* and *y* exists requires the knowledge of sample correlation coefficient, *r*. The test of null hypothesis about linear relationship between *x* and *y* is the same as determining whether there is any significant correlation between these variables. To determine whether there exist any significant correlation between variables *x* and *y*, first take a hypothesis that the value of the population correlation coefficient, ρ, is equal to zero. The population correlation coefficient, ρ, measures the degree of association between two variables in a population of interest. The null and alternative hypotheses are expressed as

**Two-tailed Test**

$H_0: \rho = 0$ (No correlation between variables *x* and *y*)

$H_1: \rho \neq 0$ (Correlation exists between variables *x* and *y*)

**One-tailed Test**

$H_0: \rho = 0$, and $H_1: \rho > 0$ (or $\rho < 0$)

The *t*-test statistic for testing the null hypothesis is given by

$$t = \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

where $r$ is sample correlation coefficient; $s_r$ is standard error of correlation coefficient and $n$ is sample size.

The $t$-test statistic follows $t$-distribution with $n - 2$ degrees of freedom. If the sample size is large, then the standard error of correlation coefficient is given by $s_r = (1 - r_2)/\sqrt{n}$ .

### Decision Rule

The calculated value of $t$-test statistic is compared with its critical (or table) value at $n - 2$ degrees of freedom and level of significance $\alpha$ to arrive at a decision as follows:

| *One-tailed Test* | *Two-tailed Test* |
|---|---|
| • Reject $H_0$<br>if $t_{cal} > t_{\alpha,n-2}$ or $t_{cal} < -t_{\alpha.}$ | • Reject $H_0$ if $t_{cal} > t_{\alpha/2,\, n-2}$ |
| • Otherwise accept $H_0$ | • Otherwise accept $H_0$ |

**Example 13.17:** A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated? *[Delhi Univ., M.Com.,* 2005]

**Solution:** Take a null hypothesis that there is no significant difference in the sample and population correlation coefficients, that is,

$$H_0 : \rho = 0 \text{ and } H_1 : \rho \neq 0$$

Given $n = 27$, $df = n - 2 = 25$, $r = 0.42$. Applying $t$-test statistic as follows:

$$t = \frac{r - \rho}{s_r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.42}{\sqrt{\{1-(0.42)^2\}/(27-2)}}$$

$$= \frac{0.42}{0.908/5} = 2.312$$

Since the calculated value of $t_{cal} = 2.312$ is more than its critical value, $t_\alpha = 1.708$ at $\alpha = 0.05$ level of significance and $df = 25$, the null hypothesis is rejected. Hence, it may be concluded that there is significant difference in the sample and population correlation coefficients.

**Example 13.18:** Is a correlation coefficient of 0.5 significant which is obtained from a random sample of 11 pairs of values from a normal population? *[Madras Univ., B.Com., 2005]*

**Solution:** Let us take the null hypothesis that the given correlation coefficient is not sufficient. Applying t-test :

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.5}{\sqrt{\{1 - (0.5)^2\}/(11 - 2)}} = \frac{0.5}{0.866} \times 3 = 1.732$$

where $r = 0.5$, $n = 11$

The calculated value of $t_{cal} = 1.732$ at $\alpha = 0.05$ level of significance and $df, v = 9$ is less than the table value $t_\alpha = 2.26$ and hence the given correaltion coefficient is not significant.

**Example 13.19:** How many pairs of observations must be included in a sample so that an observed correlation coefficient of value 0.42 shall have a calculated value of $t$ greater than 2.72?

**Solution:** Given, $r = 0.42$, $t = 2.72$. Applying $t$-test statistic, we get

$$\frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = t \text{ or } r^2 \times \frac{n - 2}{1 - r^2} = t^2$$

$$(0.42)^2 \times \frac{(n - 2)}{1 - (0.42)^2} = (2.72)^2$$

$$n - 2 = \frac{(2.72)^2\,[1 - (0.42)^2]}{(0.42)^2} = \frac{7.3984(0.8236)}{0.1764}$$

$$= \frac{6.0933}{0.1764} = 34.542$$

$$n = 2 + 34.542 = 36.542 \cong 37$$

Hence, the sample size should be of 37 pairs of observations.

**Example 13.20:** To study the correlation between the stature of father and son, a sample of 1600 is taken from the universe of fathers and sons. The sample study gives the correlation between the two to be 0.80. Within what limits does it hold true for the universe?

**Solution:** Since the sample size is large, the standard error of the correlation coefficient is given by

$$\text{SE}_r = \frac{1 - r^2}{\sqrt{n}}$$

Given correlation coefficient, $r = 0.8$ and $n = 1600$. Thus,

$$\text{Standard error SE}_r = \frac{1-(0.8)^2}{\sqrt{1600}} = \frac{1-0.64}{40} = \frac{0.36}{40} = 0.009$$

The limits within which the correlation coefficient should hold true is given by

$$r \pm 3\text{SE}_r = 0.80 \pm 3(0.009) \text{ or } 0.773 \le r \le 0.827$$

### 13.6.2 Hypothesis Testing About Population Correlation Coefficient (Large Sample)

If distribution of sample correlation coefficient, $r$, is not normal and its probability curve is skewed in the neighborhood of population correlation coefficient, $\rho = \pm 1$, even for large sample size $n$, then use Fisher's $z$-transformation for transforming $r$ into $z$ as follows:

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

The value of $z$ for different values of $r$ can be seen from the standard table given in the Appendix.

Changing natural logarithm to the base $e$ to the base 10 by multiplying with the constant 2.3026 as follows:

$$\log e\, x = 2.3026 \log_{10} x$$

where $x$ is a positive integer. Thus the transformation formula becomes

$$z = \frac{1}{2}\,(2.3026) \log_{10}\frac{1+r}{1-r} = 1.1513 \log_{10}\frac{1+r}{1-r}$$

Fisher's $z$-transformation for transforming $r$ into $z$ with:

$$\text{Mean } z_\rho = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho} = 1.1513 \log_{10}\frac{1+\rho}{1-\rho}$$

and $$\text{Standard deviation } \sigma_z = \frac{1}{\sqrt{n-3}}$$

This approximation is useful for large sample sizes. However, it can also be used for small sample sizes of at least $n \ge 10$.

The $z$-test statistic to test the null hypothesis $H_0$: $\rho = 0$ and $H_1$: $\rho \ne 0$ is given by

$$Z = \frac{z - z_\rho}{\sigma_z} = \frac{z - z_\rho}{1/\sqrt{n-3}}$$

where $\sigma$ is the standard error of Z.

#### Decision Rule

- If $|Z \text{ cal}| <$ Table value of $Z_{\alpha/2}$, then accept null hypothesis $H_0$.
- Otherwise reject $H_0$.

### 13.6.3 Hypothesis Testing About the Difference Between Two Independent Correlation Coefficients

The $z$-test statistic for **testing a hypothesis about** correlation coefficient in the **single population** can be generalized to test the hypothesis of two correlation coefficients $r_1$ and $r_2$ derived from two independent samples as follows:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \frac{z_1 - z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$$

where $z_1 = \frac{1}{2}\log_e \frac{1+r_1}{1-r_1} = 1.1513 \log_{10}\frac{1+r_1}{1-r_1}$, and

$$z_2 = \frac{1}{2}\log_e \frac{1+r_2}{1-r_2} = 1.1513 \log_{10}\frac{1+r_2}{1-r_2}$$

are approximately normally distributed with zero mean and unit standard deviation.

### Decision Rules

- If absolute value $|Z \text{ cal}|$ is less than its table value, $Z_{\alpha/2}$, then accept the null hypothesis, $H_0$.
- Otherwise reject $H_0$.

**Example 13.21:** What is the probability that a correlation coefficient of 0.75 or less arises in a sample of 30 pairs of observations from a normal population in which the true correlation is 0.9?

**Solution:** Given, $r = 0.75, n = 30$, and $\rho = 0.9$. Applying Fisher's $z$-transformation, we get

$$z = 1.1513 \log_{10}\frac{1+r}{1-r} = 1.1513 \log_{10}\frac{1.75}{0.25}$$

$$= 1.1513[\log_{10}1.75 - \log_{10}0.25]$$

$$= 1.1513(0.24304 - \bar{1}.39794) = 0.973$$

The distribution of $z$ is normal around the true population correlation value $\rho = 0.9$. Thus,

$$\text{Mean}, z_\rho = 1.1513 \log_{10}\frac{1+\rho}{1-\rho} = 1.1513 \log_{10}\frac{1+0.90}{1-0.90} = 1.1513 \log_{10}\frac{1.90}{0.10}$$

$$= 1.1513 [\log_{10}1.90 - \log_{10}0.10] = 1.1513 (0.27875 + 1) = 1.47$$

The $Z$-test statistic is given by

$$Z = \frac{|z - z_\rho|}{\sigma_z} = \frac{|z - z_\rho|}{1/\sqrt{n-3}} = \frac{|0.973 - 1.47|}{1/\sqrt{30-3}} = 0.498 \times 5.196 = 2.59$$

Hence $\rho\,(r \le 0.75) = P[Z \le 2.59] = 1 - 0.9952 = 0.0048$.

**Example 13.22:** Test the significance of the correlation, $r = 0.5$ from a sample of size 18 against hypothesized population correlation, $\rho = 0.70$.

**Solution:** Take the null hypothesis that the difference is not significant, that is,

$$H_0 : \rho = 0.70 \text{ and } H_1 : \rho \ne 0.70$$

Given, $n = 18, r = 0.5$. Applying $z$-transformation, we have

$$z = 1.1513 \log_{10}\frac{1+r}{1-r} = 1.1513 \log_{10}\frac{1+0.5}{1-0.5}$$

$$= 1.1513 \log_{10}\frac{1.50}{0.5} = 1.1513 \log_{10}3$$

$$= 1.1513(0.4771) = 0.5492$$

and $\quad \text{Mean } z\rho = 1.1513 \log_{10}\frac{1+\rho}{1-\rho} = 1.1513 \log_{10}\frac{1+0.70}{1-0.70}$

$$= 1.1513 \log_{10}\frac{1.70}{0.30} = 1.1513 \log_{10}5.67$$

$$= 1.1513(0.7536) = 0.8676$$

Applying $Z$-test statistic, we get

$$Z = \frac{|z - z_\rho|}{\sigma_z} = \frac{|z - z_\rho|}{1/\sqrt{n-3}} = |z - z\rho|\,\sqrt{n-3}$$

$$= |\ 0.5492 - 0.8676\ | = 0.3184(3.872) = 1.233$$

Since calculated value of $Z_{\text{cal}} = 1.233$ is less than its table value $Z_{\alpha./2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence, it may be concluded that the difference (if any) is due to sampling error.

**Example 13.23:** Two independent samples of size 23 and 21 pairs of observations were analysed and their coefficient of correlation was found as 0.5 and 0.8, respectively. Does this value differ significantly?

**Solution:** Take the null hypothesis that two values do not differ significantly, that is, samples are drawn from the same population.

Given $n_1 = 23$, $r_1 = 0.5$; $n_2 = 28$, $r_2 = 0.8$. Applying $Z$-test statistic as follows:

$$z = \frac{|z_1 - z_2|}{\sigma_{z_1 - z_2}};$$

$$= \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{|0.55 - 1.10|}{\sqrt{\frac{1}{20} + \frac{1}{25}}}$$

$$= \frac{0.55}{0.30} = 1.833$$

$$z_1 = 1.1513 \log_{10}\frac{1 + r_1}{1 - r_1} = 1.1513 \log_{10}\frac{1 + 0.5}{1 - 0.5}$$

$$= 1.1513 \log_{10} 3 = 0.55$$

$$z_2 = 1.1513 \log_{10}\frac{1 + r_2}{1 - r_2} = 1.1513 \log_{10}\frac{1 + 0.8}{1 - 0.8}$$

$$= 1.1513 \log_{10} 9 = 1.10$$

Since calculated value of $z_{\text{cal}} = 1.833$ is less than its table value $z_{\alpha/2} = 1.96$ at 5 per cent significance level, the null hypothesis is accepted. Hence, the difference in correlation values is not significant.

## Conceptual Questions 13A

1. What is the meaning of the coefficient of correlation?
2. Explain the meaning and significance of the term correlation. [ *Delhi Univ., MBA,2003*]
3. What is meant by 'correlation'? Distinguish between positive, negative, and zero correlation. [*Ranchi Univ., MBA, 2004*]
4. What are the numerical limits of $r^2$ and $r$? What does it mean when $r$ equals one? zero? minus one?
5. What is correlation? Clearly explain its role with suitable illustration from simple business problems. [*Delhi Univ., MBA, 2005*]
6. What is the relationship between the coefficient of determination and the coefficient of correlation? How is the coefficient of determination interpreted?
7. Does correlation always signify a cause-and-effect relationship between the variables? [*Osmania Univ., MBA, 2000*]
8. What information is provided by the coefficient of correlation of a sample? Why is it necessary to perform a test of a hypothesis for correlation?
9. When the result of a test of correlation is significant, what conclusion is drawn if $r$ is positive? If $r$ is negative?
10. What is the $t$-statistic that is used in a test for correlation? What is meant by the number of degrees of freedom in a test for correlation and how is it used?
11. What is coefficient of rank correlation? Bring out its usefulness. How does this coefficient differ from the coefficient of correlation? [*Delhi Univ., MBA, 2006*]
12. What is Spearman's rank correlation coefficient? How does it differ from Karl Pearson's coefficient of correlation?
13. (a) What is a scatter diagram? How do you interpret a scatter diagram?
    (b) What is a scatter diagram? How does it help in studying the correlation between two variables, in respect of both its direction and degree? [*Delhi Univ., MBA, 2007*]
14. Define correlation coefficient '$r$' and give its limitations. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is (–)0.60 if only drivers with at least one accident are considered?

## Self-practice Problems 13C

**13.21** The correlation between the price of two commodities $x$ and $y$ in a sample of 60 is 0.68. Could the observed value have arisen
(a) from an uncorrelated population?
(b) from a population in which true correlation was 0.8?

**13.22** The following data give sample sizes and correlation coefficients. Test the significance of the difference between two values using Fisher's $z$-transformation.

| *Sample Size* | *Value of r* |
|---|---|
| 5 | 0.870 |
| 12 | 0.560 |

**13.23** A company wants to study the relationship between R&D expenditure (in ₹1000's) and annual profit (in ₹1000's). The following table presents the information for the last 8 years.

| *Year* | :1988 | 87 | 86 | 85 | 84 | 83 | 82 | 81 |
|---|---|---|---|---|---|---|---|---|
| R&D expenses : | 9 | 7 | 5 | 10 | 4 | 5 | 3 | 2 |
| Annual profit : | 45 | 42 | 41 | 60 | 30 | 34 | 25 | 20 |

(a) Estimate the sample correlation coefficient.
(b) Test the significance of correlation coefficient at a $\alpha = 5$ per cent level of significance.

**13.24** Find the least value of $r$ in a sample of 27 pairs from a bivariate normal population at $\alpha = 0.05$ level of significance, where $t_{\alpha = 0.05} = 2.06$ at $df = 25$.

**13.25** A small retail business has determined that the correlation coefficient between monthly expenses and profits for the past year, measured at the end of each month, is $r = 0.56$. Assuming that both expenses and profits are approximately normal, test at $\alpha = 0.05$ level of significance the null hypothesis that there is no correlation between them.

**13.26** The manager of a small shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in ₹1000's and analysed the results. Has the rise been significant?

| *Week* : | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Sales : | 2.69 | 2.62 | 2.80 | 2.70 | 2.75 | 2.81 |

Find the correlation coefficient between sales and week and test it for significance at $\alpha = 0.05$.

## Hints and Answers

**13.21** (a) $z = 1.1513 \log_{10} \dfrac{1+r}{1-r} = 1.1513 \log_{10} \dfrac{1+0.68}{1-0.68}$

$= 1.1513 \log_{10} \dfrac{1.68}{0.32} = 0.829$

Standard error, $\sigma_z = \dfrac{1}{\sqrt{n-3}} = \dfrac{1}{\sqrt{57}} = 0.13$

Test statistic $z = \dfrac{z - z_\rho}{\sigma_z} = \dfrac{0.829 - 0}{0.13} = 6.38$

Since deviation of $z$ from $z\rho$ is 6 times more than $\sigma_z$, the hypothesis is not correct, that is, population is correlated.

Mean $z_\rho = 1.1513 \log_{10} \dfrac{1+\rho}{1-\rho}$

$= 1.1513 \log_{10} \dfrac{1.8}{1.2} = 1.099$

$\text{Z} = \dfrac{|z - z_\rho|}{\sigma_z} = \dfrac{|0.829 - 1.099|}{0.13} = 2.08 > 2$ times standard error, $\rho$ is likely to be less than 0.8.

**13.22** Let $\text{H}_0$: samples are drawn from the same population.

$z_1 = 1.1513 \log_{10} \dfrac{1+r_1}{1-r_1} = 1.1513 \log \dfrac{1+0.87}{1-0.87}$

$= 1.333$

$z_2 = 1.1513 \log_{10} \dfrac{1+r_2}{1-r_2} = 1.1513 \log_{10} \dfrac{1+0.56}{1-0.56}$

$= 0.633$

$= \sqrt{\dfrac{1}{n_1-3} + \dfrac{1}{n_2-3}} = \sqrt{\dfrac{1}{5-3} + \dfrac{1}{12-3}} = 0.782$

$z = \dfrac{z_1 - z_2}{\sigma_{z_1 - z_2}} = \dfrac{0.7}{0.782} = 0.895$

Since the calculated value Z = 0.895 is less than its table value $\text{Z}_\alpha = 2.58$ at $\alpha = 0.01$ level of significance, $\text{H}_0$ is accepted.

**13.23** (a) $r = 0.95$ (*b*) Let $\text{H}_0$: $r = 0$ and $\text{H}_1$: $r \neq 0$

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.95}{\sqrt{\{1-(0.95)^2\}/(8-2)}}$$

$= 7.512$

Since $t_{\text{cal}} = 7.512 > t_{\alpha/2} = 2.447$ for $df = 6$, the $\text{H}_0$ is rejected.

**13.24** $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{r\sqrt{27-2}}{\sqrt{1-r}} = \dfrac{5r}{\sqrt{1-r}} > 2.06$

or $|\, r \,| = 0.381$

**13.25** $r = 0.560$ and $t_{\text{cal}} = 0.576$, $\text{H}_0$ is rejected.

**13.26** $r = 0.656$ and $t_{\text{cal}} = 0.729$, $\text{H}_0$ is rejected.

1. Karl Pearson's correlation coefficient

$$r = \frac{\text{Covariance between } x \text{ and } y}{\sigma_x \, \sigma_y}$$

- Deviation from actual mean

$$r = \frac{\Sigma (x - \overline{x})(y - \overline{y})}{\sqrt{\Sigma (x - \overline{x})^2} \sqrt{\Sigma (y - \overline{y})^2}}$$

- Deviation from assumed mean

$$r = \frac{n \Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{n \Sigma d_x^2 - (\Sigma d_x)^2} \sqrt{n \Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$d_x = x - \mathrm{A}, \; d_y = y - \mathrm{B}$$

$$\mathrm{A, B} = \text{constants}$$

- Bivariate frequency distribution

$$r = \frac{n \Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{n \Sigma f d_x^2 - (\Sigma f d_x)^2} \sqrt{n \Sigma f d_y^2 - (\Sigma f d_y)^2}}$$

- Using actual values of $x$ and $y$

$$r = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}$$

2. Standard error of correlation coefficient, $r$

$$\mathrm{SE}_r = \frac{1 - r^2}{\sqrt{n}}$$

- Probable error of correlation coefficient, $r$

$$\mathrm{PE}_r = 0.6745 \frac{1 - r^2}{\sqrt{n}}$$

3. Coefficient of determination

$$r_2 = \frac{\text{Explained variance}}{\text{Total variance}} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \overline{y})^2}$$

4. Spearman's rank correlation coefficient

- Ranks are not equal

$$\mathrm{R} = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)}$$

- Ranks are equal

$$\mathrm{R} = 1 - \frac{6 \left[ \Sigma d^2 + \frac{1}{12} (m_i^3 - m_i) \right]}{n(n^2 - 1)}$$

$$t = 1, 2, \ldots$$

5. Hypothesis testing

- Population correlation coefficient $r$ for a small sample

$$t = \frac{r - \rho}{\mathrm{SE}_r} = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Population correlation coefficient for a large sample

$$\mathrm{Z} = \frac{z - z_\rho}{\sigma_z} = \frac{z - z_\rho}{1/\sqrt{n - 3}}$$

# Chapter 8

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

—Cervantes

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

— H. L. Mencken

# Sampling and Sampling Distributions

## LEARNING OBJECTIVES

After studying this chapter, you should be able to

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

## 8.1 INTRODUCTION

In Chapters 2 to 5, certain statistical methods were introduced to understand characteristics of a population or a process. Also the concepts of probability and its distributions were elaborated in Chapters 6 and 7 to add to the knowledge on likelihood occurrence of random event(s) and its significance to understand features (characteristics) of a population or a process.

The process of selecting a sample or a portion of elements from a population or a process using a specific method is called sampling. For example, (i) a political analyst selects specific or random set of people for interviews to estimate the proportion of the votes that each candidate may get from the population of voters; (ii) an auditor selects a sample of vouchers and calculates the sample mean for estimating population average amount; and (iii) a doctor examines a few drops of blood to draw conclusions about the nature of disease or blood constitution of the whole body.

Samples drawn from population are analysed and sample results are called **sample statistic** (such as sample mean, sample proportion, sample standard deviation, etc.). **Statistical inferences** are drawn about the population characteristics. These *s*ample statistics are treated as estimator of population parameters as $\mu$, $\sigma$, $p$, etc.

## 8.2 REASONS OF SAMPLE SURVEY

A census involves a complete count of every individual member of the population of interest, such as people in state, eligible voters, buying habits of adults, households in a town, shops in a city, students in a college, and so on. Some of the reasons to prefer sampling method instead of census method are as follows:

1. **Movement of population element:** The population of fishes, birds, snakes and mosquitoes is large and constantly moving, being born and dying. So instead of attempting to count all members of such populations, it is desirable to make estimates about their population or certain behaviour by catching a suitable number (a sample) from predetermined places.
2. **Cost and/or time required for census:** Since census involves a complete count of every individual member of the population of interest, therefore apart from the cost and the large amount of resources (such as infrastructure and manpower) which are required, lot of time is also required to process the data and hence results are known after a big interval of time.
3. **Destructive nature of certain tests:** In certain cases where (i) size of population of interest is very large, and constantly changing in a state of movement and (ii) observation results required destruction, etc., the census becomes extremely difficult. For example, to test the strength of some manufactured items by applying a stress until the unit breaks, the amount of stress (value of the observation) that results in breakage is recorded. Such a procedure when applied to each member of the population, we may end up breaking each item. This type of testing procedure is called destructive and requires that a sample be used for requisite test.

## 8.3 TYPES OF BIAS DURING SAMPLE SURVEY

Results based on sampling method may be ***biased*** due to human or chance error. The following are the common types of bias that might occur in the sampling process:

(i) ***Under-coverage bias:*** This bias occurs when a sample does not represent the population of interest. For instance, the result of a survey based on sampling method to determine passengers attitude towards buying platform ticket may not represent attitude of all passengers at railway stations.

(ii) ***Non-response bias:*** This bias occurs when adequate number of respondents does not respond to a quarry by a researcher and only those respondents respond who are particularly concerned about the subject.

(iii) ***Wording bias:*** This bias occurs when questionnaire contains questions that tend to confuse the respondents. Consequently, respondents do not respond truly. For instance, questionnaire used for survey about the use of drug, payment of income tax and abusive behaviour must be worded and conducted carefully to minimize response bias.

### 8.3.1 Sampling and Non-sampling Errors

Any statistical inference based on sample statistic may not always be correct because such results may not truly estimate population features (or characteristics). This error is referred as **sampling error** because as compared to results obtained by one-to-one analysis of members of a population; the sample statistic may provide a different estimate of the population characteristic. In such a case, it is important to have an exact knowledge about the reliability of sample based estimates of population features. Any sampling error that exceeds any specified level must be mentioned in terms of probability of error, say 5 per cent and so on. In other words, if a decision maker wants to be 95 per cent or more confident that the range of sample statistic reflects the true characteristic of the population of interest, then this acceptable margin of error is important to develop a ***confidence level*** to arrive at certain conclusions about the characteristic of the population.

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

Non-sampling errors arise due to biases and mistakes such as (i) incomplete list of population members, (ii) non-random selection of samples, (iii) use of faulty questionnaire for data collection, or (iv) wrong editing, coding, and presenting of the responses received through the questionnaire.

The sampling errors can be minimized provided questionnaire contains precise and unambiguous questions, administered carefully, and responses are correctly processed.

### 8.3.2 Measurement of Sampling Error

**Standard error of estimate** is used as a measure of sampling error. In most cases, sampling error occurs due to degree of precision while drawing samples and also on size of the sample.

The standard error of estimate is inversely proportional to the square root of the sample size. Thus, as sample size increases, the amount sampling error is decreases.

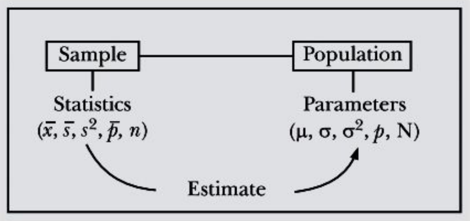
## 8.4 POPULATION PARAMETERS AND SAMPLE STATISTICS

**Parameters:** It is an exact but generally unknown value that describes population characteristics. For example, quantities such as mean μ, variance $\sigma^2$, standard deviation σ, median, mode and proportion $p$ computed from a data set (also called population) are referred as parameters. A parameter is usually denoted with letters of the lower case Greek alphabet.

**Sample Statistic:** It is a value obtained from the analyses of sample data. For example, quantities such as mean $\overline{x}$, standard deviation $s$, variance $\sigma^2$ and proportion $\overline{p}$ computed from a sample data are referred as sample statistic. A sample statistic is usually denoted by Roman letters.

The value of a sample statistic may vary from one sample to another whereas the value of a parameter remains constant. Since value of sample statistic depends on the sample drawn from a population, sample error associated with a statistical inference about a population also used on a sample. Figure 8.1 shows the estimation relationships between sample statistics and the population parameters.

**Figure 8.1**
Estimation Relationship Between Sample and Population Measures


Sample
Population
Statistics
($\overline{x}, \overline{s}, s^2, \overline{p}, n$)
Parameters
($\mu, \sigma, \sigma^2, p$, N)
Estimate

## 8.5 PRINCIPLES OF SAMPLING

The following two principles determine the possibility of arriving at a reliable statistical inference about the features of a population or process:

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

**Sample Statistic:** A sample measure, such as mean $\overline{x}$, standard deviation, $s$, pro-portion $\overline{p}$, and so on.

### 8.5.1 Principle of Statistical Regularity

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

1. **Size of sample:** As the size of sample increases, the inference about population features becomes easy and more accurate. But in actual practice, it is very expensive. Thus, a trade-off between the sample size, degree of accuracy desired and financial resources needs to be developed.
2. **Process of random sampling:** The process of random sampling can reduce the amount of efforts required to reach at a conclusion about the characteristic of any population. For example, instead of approaching each student to understand his/her book buying habit in a college, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

### 8.5.2 Principle of Inertia of Large Numbers

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

## 8.6 SAMPLING METHODS

The methods of drawing a sample or samples from the given population are divided into two categories as given in Table 8.1.

**Table 8.1** Types of Sampling Methods

| *Probability (Random)* | *Non-probability (Non-random)* |
|---|---|
| • Simple random sampling<br>• Stratified sampling<br>• Cluster sampling<br>• Systematic sampling<br>• Multi-stage sampling | • Convenience sampling<br>• Purposive sampling<br>• Quota sampling<br>• Judgement sampling |

### 8.6.1 Probabilistic Sampling Methods

The following are the probabilistic (or stochastic) sampling methods for selecting samples from a population or process:

#### Simple Random Sampling

In this method, every member of the population has an equal chance of being selected each time a sample is drawn from the population. For applying this method, an exhaustive list of members of the population of interest is prepared to identify each member by a distinct number. Such a list is called ***sampling frame for experiment.*** The frame for experiment helps to draw sample from the population using randomly generated numbers of the members to be included in the sample.

For example, to draw a random sample of 50 students from the population of 3500 students in a college, first assign each of 3500 students a unique identification number and then generate a set of 50 random numbers in the range of values from 1 and 3500 by computer or other means to draw a sample of 50 students. The procedure may be repeated any number of times.

One disadvantage with this method is that all members of the population have to be available for selection that may not be possible every time.

#### Stratified Sampling

This method is useful when the population consists of a number of heterogeneous sub-populations (age, industry type, gross sales, number of employees, etc.). These mutually exclusive sub-populations are called ***strata***. A desirable size of sample using simple random sampling method is drawn from each *strata* (or group). Individual stratum samples are combined to obtain an overall sample for analysis.

This sampling procedure is preferred over the simple random sampling procedure because, for the same sample size, more representative sample from each stratum is obtained. For example, if management of a company is concerned about low motivational level or high absentee rate among the employees, it makes sense to stratify the population of employees. Table 8.2 shows that 750 employees of an organization are classified according to their job levels. If a sample of 100 employees is to be taken for study, then their number to be selected from each stratum is shown in Table 8.2.

### 8.5.2 Principle of Inertia of Large Numbers

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

## 8.6 SAMPLING METHODS

The methods of drawing a sample or samples from the given population are divided into two categories as given in Table 8.1.

**Table 8.1** Types of Sampling Methods

| *Probability (Random)* | *Non-probability (Non-random)* |
|---|---|
| • Simple random sampling<br>• Stratified sampling<br>• Cluster sampling<br>• Systematic sampling<br>• Multi-stage sampling | • Convenience sampling<br>• Purposive sampling<br>• Quota sampling<br>• Judgement sampling |

### 8.6.1 Probabilistic Sampling Methods

The following are the probabilistic (or stochastic) sampling methods for selecting samples from a population or process:

#### Simple Random Sampling

In this method, every member of the population has an equal chance of being selected each time a sample is drawn from the population. For applying this method, an exhaustive list of members of the population of interest is prepared to identify each member by a distinct number. Such a list is called ***sampling frame for experiment.*** The frame for experiment helps to draw sample from the population using randomly generated numbers of the members to be included in the sample.

For example, to draw a random sample of 50 students from the population of 3500 students in a college, first assign each of 3500 students a unique identification number and then generate a set of 50 random numbers in the range of values from 1 and 3500 by computer or other means to draw a sample of 50 students. The procedure may be repeated any number of times.

One disadvantage with this method is that all members of the population have to be available for selection that may not be possible every time.

#### Stratified Sampling

This method is useful when the population consists of a number of heterogeneous sub-populations (age, industry type, gross sales, number of employees, etc.). These mutually exclusive sub-populations are called ***strata***. A desirable size of sample using simple random sampling method is drawn from each *strata* (or group). Individual stratum samples are combined to obtain an overall sample for analysis.

This sampling procedure is preferred over the simple random sampling procedure because, for the same sample size, more representative sample from each stratum is obtained. For example, if management of a company is concerned about low motivational level or high absentee rate among the employees, it makes sense to stratify the population of employees. Table 8.2 shows that 750 employees of an organization are classified according to their job levels. If a sample of 100 employees is to be taken for study, then their number to be selected from each stratum is shown in Table 8.2.

**Table 8.2** Proportionate and Disproportionate Stratified Random Samples

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

Disproportionate sampling decisions are made either when strata are too small/ large or more variability is seen within a particular stratum. This procedure is adopted when data collection is easy and less expensive.

The stratified sampling method will become more effective in terms of reliability, efficiency and precision, provided stratification brings

- maximum uniformity among members of each stratum.
- largest degree of variability among various strata.

### Cluster Sampling

This method, also known as area sampling method, helps to meet the problem of costs or inadequate sampling frames. For this method, the entire population is divided into smaller groups (called ***clusters***) and a sample is drawn using simple random sampling method. The members or elements of a cluster are called *elementary units*. A household where individuals live together is an example of a cluster.

If clusters with intra-heterogeneity and inter-homogeneity are found, then a random sample of the clusters can be drawn with information gathered from each member in the randomly chosen clusters. Cluster samples carry more heterogeneity within clusters and more homogeneity among clusters—the reverse of what is observed in stratified random sampling, where there is homogeneity within each strata and heterogeneity among strata.

For example, committees that are formed by choosing people from various departments in an organization to help making decisions on product development, budget allocations and marketing strategies represent clusters. Each of these clusters (or groups) contains a heterogeneous collection of members with different interests, orientations, values, philosophy and vested interests. However, based on individual and combined perceptions, it is possible to make decisions on certain issues.

Cluster sampling involves preparing only a list of clusters instead of a list of individual members. For example, airlines sometimes select randomly a set of flights to distribute questionnaire among passenger on those flights to measure customer satisfaction. In this case, each flight is a cluster. It is much easier for the airline to choose a random sample of flights than to identify and locate a random sample of individual passengers to distribute questionnaire.

### Multistage Sampling

This method of sampling is useful when the population is very widely spread and random sampling is not possible. The population is first stratified in different states (region) and further in urban and rural areas. A random sample of communities within these strata is chosen. These communities are then divided into different urban and rural areas as clusters and a few clusters are chosen randomly for study.

For example, to conduct a national pre-election opinion poll (i) *first stage* is to choose a specific state (region). The size of the sample (i.e. number of districts) from this region is determined by the relative population in each district, (ii) *second stage* is to choose a limited number of towns/cities in each of district and (iii) *third stage* is to choose from each selected towns/cities, a sample of respondents from the electoral roll.

The essence of this type of sampling is that a sub-sample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is chosen, it is important to classify all the urban areas in the region in such a way that the elements (towns/cities) of the population in those areas are given equal chances of being selected in the sample.

#### Systematic Sampling

This procedure is useful when elements of the population are already arranged in some order (e.g. alphabetic list of people with driving license, bank customers by account numbers, etc.). In such cases, one element of population is chosen at random from first $k$ elements and then every $k$th element (member) is included in the sample. The number, $k = N/n$, where N is the size of population and $n$ is the size of desired sample is called the *sampling interval*. For example, if a sample size of 50 is desired from a population consisting of 1000 accounts receivable, then sampling interval is $k = N/n = 1000/50 = 20$. Thus, a sample of 50 accounts is selected by moving through the population and identifying every 20th account after the first randomly selected account number.

### 8.6.2 Non-random Sampling Methods

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

#### Convenience Sampling

In this procedure, units to be included in the sample are selected at the convenience of the investigator. For example, (i) a student for the project on 'food habits among adults' may use his/her own friends in the college to constitute a sample because respondents are readily available and will participate for little or no cost, and (ii) public opinion surveys conducted by any TV channel near the railway station, bus stop, or in a market.

This method is easy for collecting data on a particular issue but samples may not truly represent the population and hence precautions should be taken in drawing inferences about a population characteristics based on convenient samples.

#### Purposive Sampling

Instead of collecting information from conveniently reachable respondents, the specific target respondents are approached to collect the desired information either because they are the only ones who can give the desired information or they satisfy some criteria set of the investigator.

#### Judgment Sampling

The judgment sampling is used when a specific number of respondents are in the best position to provide the desired information. The results of this method cannot be generalized because responses from a set of respondents who are conveniently available are considered. This method is useful only in those cases where desired information can only be obtained from a very specific section of respondents. However, the validity of the sample results depends on the judgment of the investigator in choosing the sample.

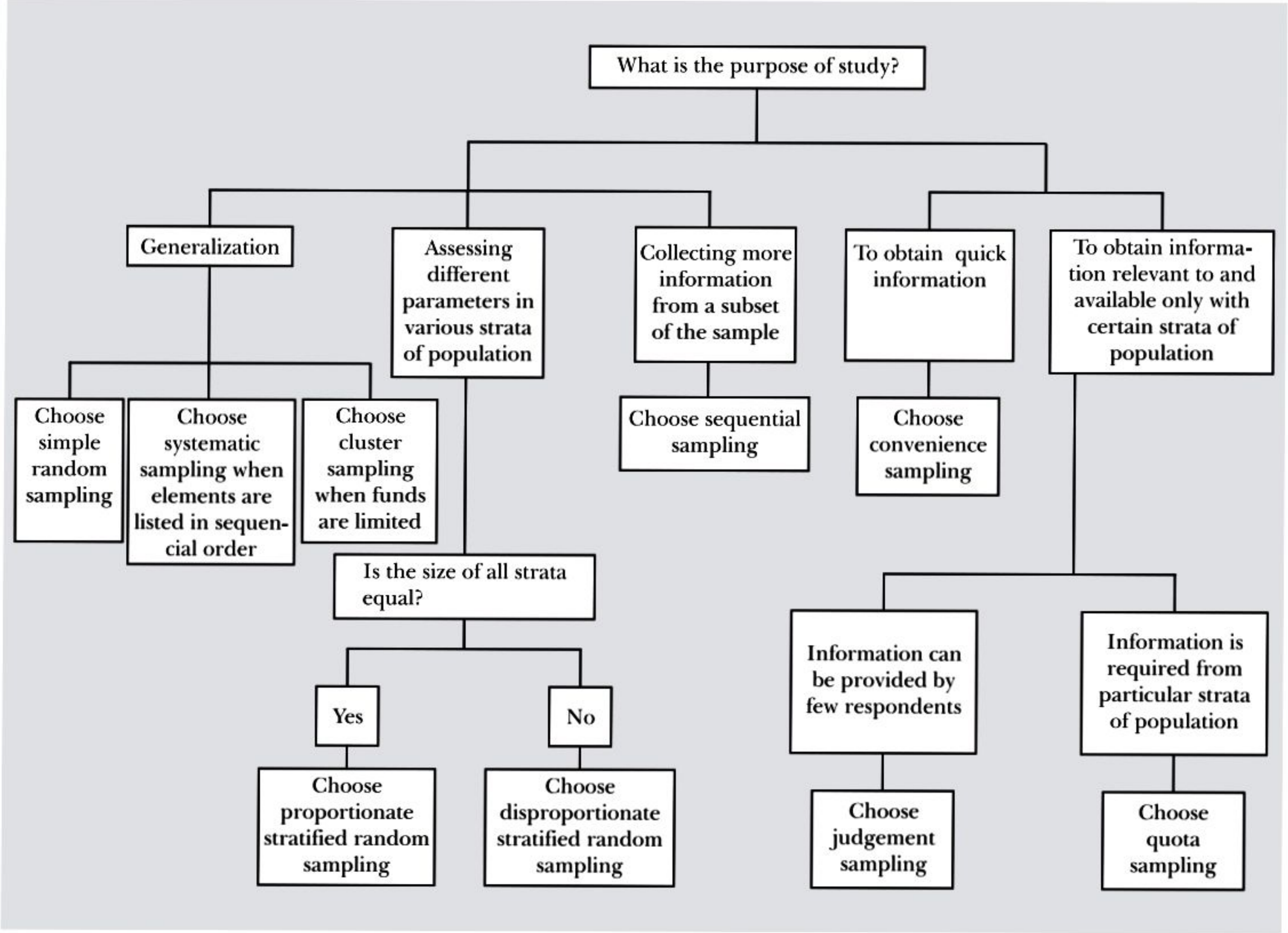
#### Quota Sampling

This method is a proportionate stratified sampling method in which a predetermined proportion of respondents is included in the sample from different groups in the population but on convenience basis that satisfies certain criteria for the study.

### 8.6.3 Choice of Sampling Methods

A particular sampling method is chosen based on certain factors such as nature of study, size of the population, size of the sample, availability of resources and degree of precision desired. A guideline for choosing a sampling plan is shown in Fig. 8.2.

**Figure 8.2**
Guidelines to Choose Sample


What is the purpose of study?
Generalization
Assessing different parameters in various strata of population
Collecting more information from a subset of the sample
To obtain quick information
To obtain information relevant to and available only with certain strata of population
Choose simple random sampling
Choose systematic sampling when elements are listed in sequencial order
Choose cluster sampling when funds are limited
Choose sequential sampling
Choose convenience sampling
Is the size of all strata equal?
Yes
No
Choose proportionate stratified random sampling
Choose disproportionate stratified random sampling
Information can be provided by few respondents
Information is required from particular strata of population
Choose judgement sampling
Choose quota sampling

#### Judging the Reliability of a Sample

The following tips are useful to ensure reliability of a sample to ensure dependable results:

- Take few samples from a population and compare their results. The variation among sample statistic of different samples should be within acceptable limit.
- Take a sub-sample from the main sample. If sample statistic of the sub-sample are similar to those of the main samples, then go ahead with the investigation, otherwise stop for correction in the sampling process.
- Compare theoretical properties of a population distribution with those of sampling distribution. If difference between them is not significant, then sample has given dependable results.

## 8.7 SAMPLING DISTRIBUTIONS

In Chapter 3, we have discussed several statistical methods to calculate parameters such as the mean and standard deviation of the population of interest. These values were used to describe the characteristics of the population. If a population is very large and the description of its characteristics is not possible by the census method, then to arrive at the

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

statistical inference, samples of a given size are drawn repeatedly from the population and a particular *'statistic'* is computed for each sample. The computed value of a particular statistic is likely to vary from sample to sample. Thus, it would be possible to construct a frequency table showing various values statistic and the frequency of their occurrence. This *distribution of values of a sample statistic is called a* **sampling distribution.** Since values of sample statistic are the outcome of several simple random samples, therefore sample statistic is random variables.

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

The concept of sampling distribution can be related to the various probability distributions. Probability distributions such as binomial, poisson and normal are theoretical distributions and helpful to determine probabilities of outcomes of random variables when any population or process generate these outcomes under certain specified conditions. For example, if mean values obtained from samples follow properties of normal distribution, then this distribution is useful to frame rules for making statistical inferences about a population on the basis of a single sample drawn from it, that is, without even repeating the sampling process. The sampling distribution of a sample statistic also helps in describing the extent of error associated with an estimate of the value of population parameters.

### 8.7.1 Standard Error of Statistic

Since sampling distribution describes how values of a sample statistic, say mean, is scattered around its own mean $\mu_{\bar{x}}$, therefore its standard deviation $\sigma_{\bar{x}}$ is called the *standard error* to distinguish it from the standard deviation $\sigma$ of a population. The population standard deviation describes the variation among values of the members of the population, whereas the standard deviation of sampling distribution measures the variation among values of the sample statistic (such as mean values and proportion values) due to sampling errors.

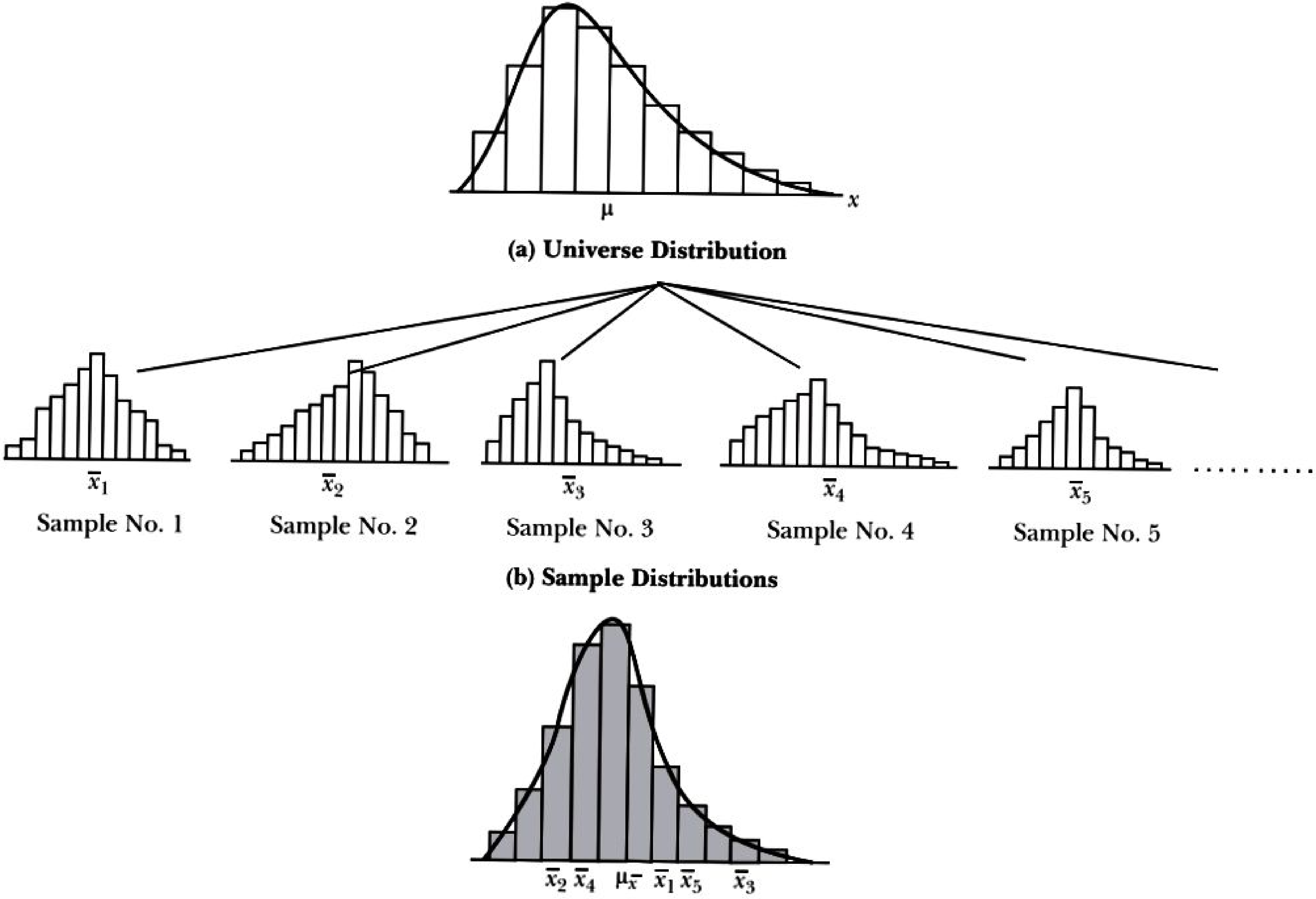
The sampling distribution of a sample statistic enables us to determine the probability of sampling error of the given magnitude. Consequently, standard deviation of sampling distribution of a sample statistic measures sampling error and is also known as ***standard error of statistic***.

### 8.7.2 Difference Between Population, Sample and Sampling Distributions

The *sampling distribution of statistic of interest* has its own mean and standard deviation (also called *standard error*). Such distributions describe the relative frequency of occurrence of values of a sample statistic and hence help to estimate values of population parameters, mean $\mu$, variance $\sigma^2$ and standard deviation $\sigma$.

Sample distribution is the distribution of measured values of *statistic* from random samples drawn from a given population. Each sample distribution is a discrete distribution [as shown in Fig 8.3(b)] because the value of sample mean would vary from sample to sample. This variability serves as the basis for the random sampling distribution. In Fig. 8.3(b), only five such samples are shown, however, there could be several such cases. In such distributions, the mean value represents the mean of all possible sample means denoted by $\bar{x}$; the standard deviation measures the variability among all possible values of the sample values, and is considered as a good approximation of the population's standard deviations $\sigma$. To estimate $\sigma$ of the population to greater accuracy the formula

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



**(a) Universe Distribution**

**(b) Sample Distributions**

**(c) Sampling Distribution**

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

## Conceptual Questions 8A

1. Briefly explain
   (a) The fundamental reason for sampling.
   (b) Some of the reasons why a sample is chosen instead of testing the entire population.
2. What is the relationship between the population mean, the mean of a sample and the mean of the distribution of the sample mean?
3. Is it possible to develop a sampling distribution for other statistics besides sample mean? Explain.
4. How does the standard error of mean measure sampling error? Is the amount of sampling error in the sample mean affected by the amount of variability in the universe? Explain.
5. If only one sample is selected in a sampling problem, how is it possible to have an entire distribution of the sample mean?
6. What is sampling? Explain the importance in solving business problems. Critically examine the well-known methods of probability sampling and non-probability sampling. [*Delhi Univ., MBA,2001*]
7. Point out the differences between a sample survey and a census survey. Under what conditions are these undertaken? Explain the law which forms the basis of sampling. [*Delhi Univ., MBA, 2004*]
8. Explain with the help of an example, the concept of sampling distribution of a sample statistic and point out its role in managerial decision making. [*Delhi Univ., MBA, 2006*]
9. Why does the sampling distribution of mean follow a normal distribution for a large sample size even though the population may not be normally distributed?
10. Explain the concept of standard error. Discuss the role of standard error in large sample theory.
11. What do you mean by sampling distribution of a statistic and its standard error? Give the expressions for the standard error of the sample mean.
12. Bring out the importance of sampling distribution and the concept of standard error in statistical application.
13. Explain the principles of 'Inertia of Large Numbers' and 'Statistical Regularity'.
14. Enumerate the various methods of sampling and describe two of them mentioning the situations where each one is to be used.
15. Distinguish between sampling and non-sampling errors. What are their sources? How can these errors be controlled?
16. (a) What is the distinction between a sampling distribution and a probability distribution?
    (b) What is the distinction between a standard deviation and a standard error?
17. Is the standard deviation of sampling distribution of mean the same as the standard deviation of the population? Explain.
18. Explain the terms 'population' and 'sample'. Explain, why is it sometimes necessary and often desirable to collect information about the population by conducting a sample survey instead of complete enumeration?
19. What are the main steps involved in a sample survey. Discuss different sources of error in such surveys and point out how these errors can be controlled.

## 8.8 SAMPLING DISTRIBUTION OF SAMPLE MEAN

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

### 8.8.1 Sampling Distribution of Mean When Population Has Non-normal Distribution

If population is not normally distributed, then **central limit theorem** is used to describe the random nature of the sample mean for large samples. The Central Limit Theorem states that

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

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

These results imply that the spread of the distribution of sample means is considerably less than the spread of the sampled population.

The values of sample statistic such as average' or 'proportion' are used to evaluate the probability of certain sample results using the normal distribution as follows:

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

The **central limit theorem** states that the approximation to normal distribution is valid as long as the sample size is large how much? The following guidelines are helpful in deciding an appropriate value of $n$:

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

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

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

**Central Limit Theorem:** A result that enables the use of normal probability distribution to approximate the sampling distribution of $\bar{x}$ and $\bar{p}$ .

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

### 8.8.2 Sampling Distribution of Mean When Population Has Normal Distribution

#### Population Standard Deviation σ Is Known

Regardless of population distribution, for a sample of size $n$ taken from a population with mean μ and standard deviation σ, the sampling distribution of a sample statistic such as mean and standard deviation are defined respectively by

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

If all possible samples of size $n$ are drawn *with replacement* from a population having normal distribution with mean μ and standard deviation σ, then the sampling distribution of mean $\bar{x}$ and standard error $\sigma_{\bar{x}}$ will also be normally distributed irrespective of the size of the sample. In particular, if the sampling distribution of mean, $\bar{x}$ is normal, then standard error of the mean, $\sigma_{\bar{x}}$ can be used to determine probability of various values of sample mean. For this purpose, sample mean $\bar{x}$, value is first converted into a value of normal variate, $z$ to know how any single mean value deviates from the mean, $\bar{x}$ of sample mean values by using the formula

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

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

- $\bar{x} \pm \sigma_{\bar{x}}$ covers about the middle 68 per cent of the total possible sample mean values.
- $\bar{x} \pm 2\,\sigma_{\bar{x}}$ covers about the middle 95 per cent of the total possible sample mean values.

**Procedure**

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

***Population standard deviation, σ value is known, and***

- population distribution is normal.
- population distribution is not normal but the sample size $n$ is large ($n \geq 30$).

In such a case sampling distribution of mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ are very close to the standard normal distribution given by

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

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

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

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

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

### Population Standard Deviation σ Is Not Known

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

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

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

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

the following formula, also called 'Student's $t$-distribution',

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

is used, where $s = \sqrt{\sum (x - \bar{x})^2/(n-1)}$ .

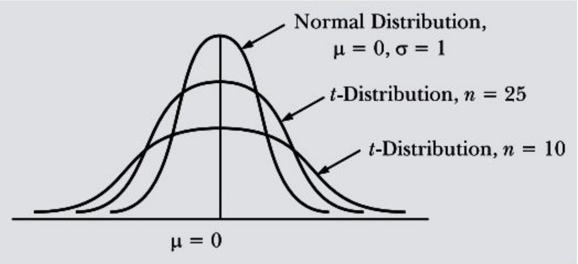


**Figure 8.4**
Comparison of *t*-Distributions with Standard Normal Distribution

In contrast to the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, the $t$-distribution is a family of symmetrical distributions with mean $\mu = 0$. The shape of the distribution depends on two statistics: $\bar{x}$ and $s$. However, the value of $s$ varies with sample size $n$. The higher the sample size $n$, $s$ will be a more accurate estimate of population standard deviation $\sigma$ and vice-versa. Figure 8.4 illustrates a comparison of $t$-distribution with that of the standard normal distribution.

### Degrees of Freedom

The devisor $(n - 1)$ in the formula for the sample variance, $s^2$ is called number of *degrees of freedom* (*df*) associated with $s^2$. The number of degrees of freedom refers to the number of values that are free to vary in a random sample (number of squared deviations in $s^2$ that are available for estimating population variance, $\sigma^2$). The shape of $t$-distribution varies with degrees of freedom. Obviously, more is the sample size $n$, higher is the degrees of freedom.

**Example 8.1:** A normal population has a mean 0.1 and a standard deviation 2.1. Find the probability that the mean of a sample of 900 elements of population will be negative.

**Solution:** Given that $\mu = 0.1$, $s = 2.1$, and $n = 900$. The standard error of mean ( $\bar{x}$ ) is given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.1}{\sqrt{900}} = \frac{2.1}{30} = 0.07 .$$

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

Let $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$, so that $z$ is a standard normal variate. Then

$$\bar{x} = \mu + \left(\frac{\sigma}{\sqrt{n}}\right) z = 0.1 + (0.07)\, z .$$

The mean $\bar{x}$ of sample will be negative provided $0.1+(0.07)\, z < 0$, i.e.,

$$0.07z < -0.1 \quad \text{or} \quad -z > \frac{0.1}{0.07} = 1.428.$$

Now $\qquad P(-z > 1.428) = P(z < 1.428) = 0.5 - P(0 < z < -1.428)$

$$= 0.5 - 0.4236 = 0.764.$$

**Example 8.2:** It is known that the means and standard deviations of a variable are respectively 100 and 10 in the universe. It is however considered sufficient to draw a sample of sufficient size to ensure that the mean of the sample would be in all probability within 0.10 per cent of the true value. How much would be the cost (exclusive of overhead charges) if the charges for drawing 100 members of a sample be one rupee? Find the extra cost necessary to double the precision.

**Solution:** Given that the sample mean should not differ from the true mean by 0.01 per cent = 0.01 as the true mean is given as 100. Also the standard error of the mean of the sample is, $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{n}$ .

If sample size is $n$ and $s$ is the standard deviation of the universe, then the range $\mu \pm 3s$ under normal curve contains almost all the values of the variable, where $\mu$ is the mean of the distribution. Thus,

$$|z| = \left|\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}\right| = 3.$$

$$\left|\frac{0.01}{\sigma / \sqrt{n}}\right| = 3 \quad \text{or} \quad \sqrt{n} = \frac{30}{0.01} = 300, \text{ i.e., } n = 90{,}000.$$

The sample size is 90,000, so the sampling charges are 90,000/100 = ₹900. To double the precision, we must have $\dfrac{0.005}{10/\sqrt{n}} = 3$ or $\sqrt{n} = \dfrac{30}{0.005} = 6{,}000$, i.e., $n = 3{,}60{,}00{,}000$.

Thus, the sample size is 3,60,00,000 and the sampling charges are 3,60,00,000/100 = ₹3,60,000.

The required extra cost = ₹(3,60,000 – 900) = ₹3,59,100.

**Example 8.3:** A research worker wishes to estimate the mean of a population using a sample sufficiently large that the probability will be 0.95, that the sample mean will not differ from the true mean by more than 0.25 per cent of the standard deviation. How large a sample should be taken?

**Solution:** Let the required size of the sample be $n$. Then

$$P\left[\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right|<1.96\right] = 0.95,$$

where $\bar{x}$ is the mean of the sample of size $n$; $\mu$ and $\sigma$ are the mean and standard deviations of the population from which the sample is drawn. Also, it is given that

$$|\,\bar{x}-\mu\,| \leq \frac{1}{4}\sigma \text{ or } \frac{1.96\sigma}{\sqrt{n}} \leq \frac{\sigma}{4} \text{ or } \sqrt{n} = 4(1.96)\text{, i.e. } n = 61.465.$$

Hence, $n = 62$ is the required minimum size of the sample.

**Example 8.4:** If the mean breaking strength of copper wire is 575 pounds with a standard deviation of 8.3 pounds, how large sample must be used in order that there be one chance in 100 that the mean breaking strength of the sample is less than 572 pounds?

**Solution:** Since $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \dfrac{572-575}{8.3/\sqrt{n}}$ or $|z| = \dfrac{3\sqrt{n}}{8.3}$.

But it is required that the value of random variable, $x$ should be such where area under normal curve to the right of normal variate $z$ be 0.01, so that area to the lift would be 1 – 0.01 = 0.99 (= 0.50 + 0.49)

From the area under the standard normal curve, the corresponding value of $z$ is 2.33. Hence,

$$|z| = \frac{3\sqrt{n}}{8.3} \text{ gives } 2.33 = \frac{3\sqrt{n}}{(8.3)}$$

or $$3\sqrt{n} = (2.33)\times(8.3) = 19.339$$

or $$n = (6.446)^2 = 41.55 \cong 42.$$

**Example 8.5:** The mean length of life of a certain cutting tool is 41.5 hours with a standard deviation of 2.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours?

*[Delhi Univ., MBA, 2003]*

**Solution:** From the data of the problem, we have $\mu = 41.5$ hours, $\sigma = 2.5$ hours and $n = 50$

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

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

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

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

The population distribution is unknown but sample size $n = 50$ is large enough to apply the central limit theorem. Hence, the normal distribution can be used to find the required probability as shown by the shaded area in Fig. 8.5.

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

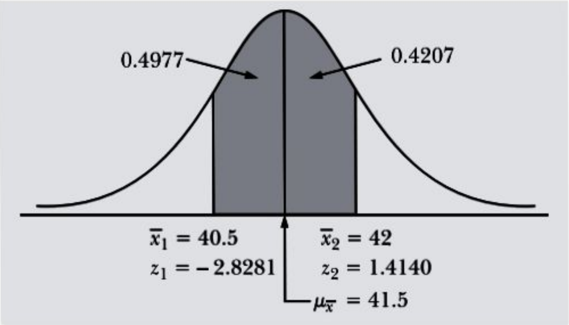
$$= P\left[\frac{40.5 - 41.5}{0.3536} \le z \le \frac{42 - 41.5}{0.3536}\right]$$

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

$$= P[z \le -2.8281] + P[z \le 1.4140]$$

$$= 0.4977 + 0.4207 = 0.9184$$

Figure 8.5
Normal Curve



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

**Example 8.6:** A continuous manufacturing process produces items whose weights are normally distributed with a mean weight of 800 gm and a standard deviation of 300 gm. A random sample of 16 items is to be drawn from the process.

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

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

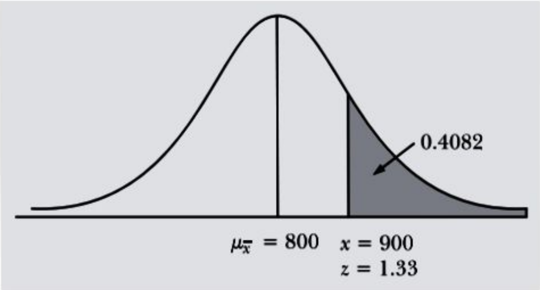
**Solution:** (a) From the data of the problem, we have, $\mu = 800$ g, $\sigma = 300$ g, and $n = 16$

Since population is normally distributed, the distribution of sample mean is normal with mean and standard deviation equal to $\mu_{\bar{x}} = \mu = 800$, and

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

The required probability $P(\bar{x} > 900)$ is represented by the shaded area in Fig. 8.6 of a normal curve. Hence,

Figure 8.6
Normal Curve



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

$$= P[z > 1.33] = 0.5000 - 0.4082 = 0.0918$$

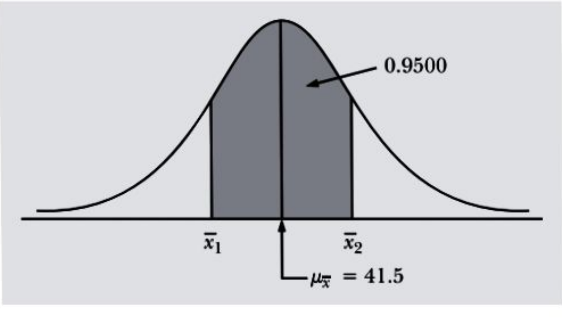
Hence, 9.18 per cent of all possible samples of size $n = 16$ will have a sample mean value greater than 900 g.

(b) Since $z = 1.96$ for the middle 95 per cent area under the normal curve as shown in Fig. 8.7, therefore using the formula for $z$ get the values of $\bar{x}$ in terms of the known values are as follows:

$$\bar{x}_1 = \mu_{\bar{x}} - z\sigma_{\bar{x}} = 800 - 1.96(75) = 653 \text{ g}$$

and $$\bar{x}_2 = \mu_{\bar{x}} + z\sigma_{\bar{x}} = 800 + 1.96(75) = 947 \text{ g}$$

Figure 8.7
Normal Curve



**Example 8.7:** An oil refinery has backup monitors to keep track of the refinery flows continuously and to prevent machine malfunctions from disrupting the process. One particular monitor has an average life of 4300 hours and a standard deviation of 730 hours. In addition to the primary monitor, the refinery has set up two standby units, which are duplicates of the primary one. In the case of malfunction of one of the monitors, another will automatically take over in its place. The operating life of each monitor is independent of the other.

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

(b) At most 12,630 hours?

**Solution:** From the data of the problem, we have, $\mu = 4300$ hours, $\sigma = 730$ hours, $n = 3$. The statistic of the sampling distribution are computed as

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

and

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

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

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

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

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

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

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

**Example 8.8:** Big Bazaar, a chain of 130 shopping malls has been bought out by another larger nationwide supermarket chain. Before the deal is finalized, the larger chain wants to have some assurance that Big Bazaar will be a consistent money maker. The larger chain has decided to look at the financial records of 25 of the Big Bazaar outlets. Big Bazaar claims that each outlet's profits have an approximately normal distribution with the same mean and a standard deviation of ₹40 million. If the Big Bazaar management is correct, then what is the probability that the sample mean for 25 outlets will fall within ₹30 million of the actual mean?

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

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

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

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

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

$$= P\,(-2.18 \leq z \leq 2.18) = 0.4854 + 0.4854 = 0.9708$$

**Example 8.9:** Chief Executive Officer (CEO) of a life insurance company wants to undertake a survey of the huge number of insurance policies that the company has underwritten. The company makes an yearly profit on each policy that is distributed with mean ₹8000 and standard deviation ₹300. It is desired that the survey must be large enough to reduce the standard error to no more than 1.5 per cent of the population mean. How large should sample be?

**Solution:** From the data of the problem, we have, $\mu$ = ₹8000, and $\sigma$ = ₹300. The aim is to find sample size $n$ to be large enough so that

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

or

$$\frac{300}{\sqrt{x}} \le 0.015 \times 8000 = 120$$

$$300 \le 120\sqrt{n} \text{ or } \sqrt{n} \ge 25, \text{ or } n \ge 625$$

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

**Example 8.10:** Safal, a tea manufacturing company, is interested in determining the consumption rate of tea per household in Delhi. The management believes that yearly consumption per household is normally distributed with an unknown mean $\mu$ and standard deviation of 1.50 kg.

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

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

**Solution:** From the data of the problem, we have $\mu$ = 500 gm, $n$ = 25 and $s = \sigma/\sqrt{n} = 1.5/\sqrt{25} = 0.25$ kg.

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

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

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

$$= 0.4772 + 0.4772 = 0.9544$$

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

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

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

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

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

Hence, the management of the company should sample at least 50 households.

**Example 8.11:** A motorcycle manufacturing company claims that its particular brand of motorcycle gave an average highway km/litre rating of 90. An independent agency tested it to verify the claim. Under controlled conditions, the motorcycle was driven for a distance of 100 km on each of 25 different occasions. The actual km/litres achieved during the trip were recorded on each occasion. Over the 25 trials, the average and standard deviation turned out to be 87 and 5 km/litre, respectively. It is believed that the distribution of the actual highway km/litre for this motorcycle is close to a normal distribution.

If the rating of 90 km/litre of the agency is correct, find the probability that the average km/litre over a random sample of 25 trials would be 87 or less.

**Solution:** Since the population standard deviation $\sigma$ is unknown, $t$-Student's test will be applicable to calculate the desired probability $P(\bar{x} \le 87)$ as follows:

$$P(\bar{x} \le 87) = P\left[t \le \frac{\bar{x}-\mu}{s/\sqrt{n}}\right] = P\left[t \le \frac{87-90}{5/\sqrt{25}}\right]$$

$$= P[t \le -3]$$

with degrees of freedom $(n - 1) = (25 - 1) = 24$.

The desired probability of $t \le -3.00$ with $df = 24$ from $t$-distribution table is 0.0031. Hence, the probability that the average km/litre is less than or equal to 87 is very small.

### 8.8.3 Sampling Distribution of Difference Between Two Sample Means

The concept of sampling distribution of sample mean can also be used to compare a population of size $N_1$ having mean $\mu_1$ and standard deviation $\sigma_1$ with another similar type of population of size $N_2$ having mean $\mu_2$ and standard deviation $\sigma_2$.

Let $\bar{x}_1$ and $\bar{x}_2$ be the mean of sampling distribution of mean values of two populations, respectively. Then the difference between their mean values $\mu_1$ and $\mu_2$ of the two populations can be estimated by generalizing the formula of standard normal variable as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{\bar{x}_1} - \mu_{\bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

where $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$ (mean of sampling distribution of difference of two means)

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$ (Standard error of sampling distribution of two means)

$n_1, n_2$ = independent random samples drawn from first and second population, respectively.

Since random samples are drawn independently from two populations with replacement, therefore the sampling distribution of the difference of two means, $\bar{x}_1 - \bar{x}_2$ will also be normal provided sample size is sufficiently large.

**Example 8.12:** Car stereos of manufacturer A have a mean lifetime of 1400 hours with a standard deviation of 200 hours while those of manufacturer B have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If a random sample of 125 stereos of each manufacturer are tested, what is the probability that manufacturer A's stereos will have a mean lifetime which is at least (a) 160 hours more than manufacturer B's stereos and (b) 250 hours more than the manufacturer B's stereos? *[Delhi Univ., MBA, 2006]*

**Solution:** From the data of the problem, we have

Manufacturer A: $\mu_1 = 1400$ hours, $\sigma_1 = 200$ hours, $n_1 = 125$

Manufacturer B: $\mu_2 = 1200$ hours, $\sigma_2 = 100$ hours, $n_2 = 125$

Thus,

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$

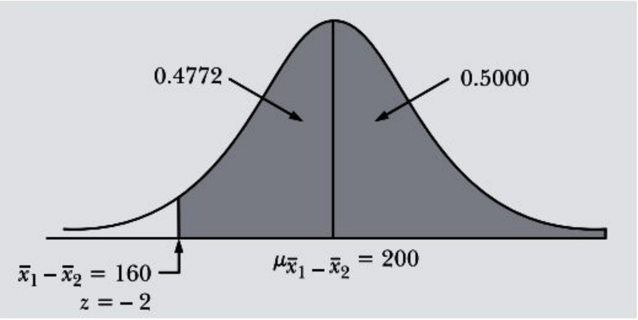
$$= \mu_1 - \mu_2 = 1400 - 1200 = 200$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}}$$

$$= \sqrt{80 + 320} = \sqrt{400} = 20$$

(a) Let $\bar{x}_1 - \bar{x}_2$ be the difference in mean lifetime of stereo manufactured by the two manufacturers. Then, it is required to find the probability that this difference is more than or equal to 160 hours as shown in Fig. 8.8.

**Figure 8.8**
Normal Curve



$$P[(\bar{x}_1 - \bar{x}_2) \geq 160] = P\left[z \geq \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$$

$$= P\left[z \geq \frac{160-200}{20}\right]$$

$$= P[z \geq -2]$$

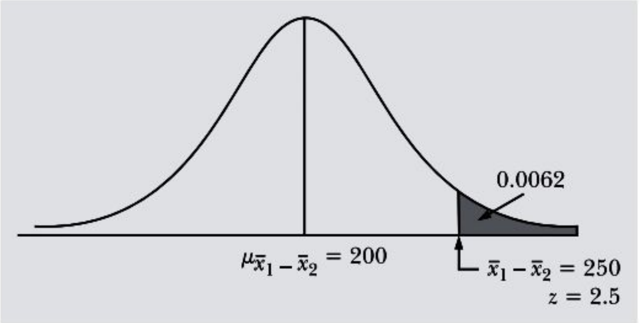
$$= 0.5000 + 0.4772$$

$= 0.9772$ (Area under normal curve)

Hence, the probability is very high that the mean lifetime of the stereos of A is 160 hours more than that of B.

(b) Proceeding in the same manner as in part (a) as follows:

**Figure 8.9**
Normal Curve



$$P[(\bar{x}_1 - \bar{x}_2) \geq 250] = P\left[z \geq \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$$

$$= P\left[z \geq \frac{250-200}{20}\right]$$

$$= P[z \geq 2.5]$$

$$= 0.500 - 0.4938$$

$= 0.0062$ (Area under normal curve)

Hence, the probability is very less that the mean lifetime of the stereos of A is 250 hours more than that of B as shown in Fig. 8.9.

**Example 8.13:** The particular brand of ball bearings weighs 0.5 kg with a standard deviation of 0.02 kg. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2 gm.

**Solution:** From the data of the problem, we have

Lot 1: $\mu_{\bar{x}_1} = \mu_1 = 0.50\text{kg}; \sigma_1 = 0.02$ kg and $n_1 = 100$

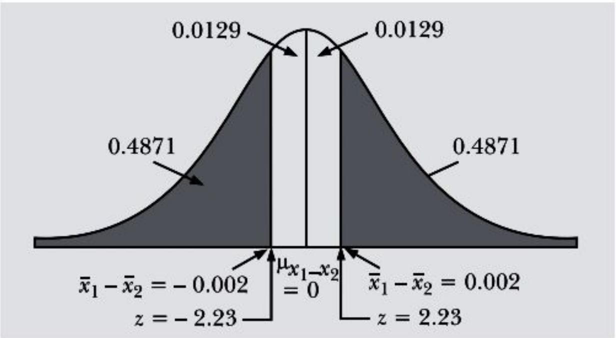
Lot 2: $\mu_{\bar{x}_2} = \mu_2 = 0.50\text{kg}; \sigma_2 = 0.02$ kg and $n_1 = 100$

Thus $\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 0$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000}+\frac{(0.02)^2}{1000}} = 0.000895$$

A difference of 2 gm in two lots is equivalent to a difference of 2/100 = 0.002 kg in mean weights. If $\bar{x}_1 - \bar{x}_2 \geq 0.002$ or $\bar{x}_1 - \bar{x}_2 \geq -0.002$, then the required probability that each ball bearing will differ by more than 2 gm is calculated as follows and shown in Fig. 8.10.

**Figure 8.10**
Normal Curve



$$P[-0.002 \leq \bar{x}_1 - \bar{x}_2 \leq 0.002]$$

$$= P\left[\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}} \leq z \leq \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$$

$$= P\left[\frac{-0.002}{0.000895} \leq z \leq \frac{0.002}{0.000895}\right]$$

$$= P[-2.33 \leq z \leq 2.33]$$

$$= 2\,[0.5000 - 0.4871] = 0.0258$$

## Self-practice Problems 8A

**8.1** A diameter of a component produced on a semi-automatic machine is known to be distributed normally with a mean of 10 mm and a standard deviation of 0.1 mm. If a random sample of size 5 is picked up, what is the probability that the sample mean will be between 9.95 mm and 10.05 mm? *[Delhi Univ., MBA, 2003]*

**8.2** The time between two arrivals at a queuing system is normally distributed with a mean of 2 minutes and standard deviation 0.25 minute. If a random sample of 36 is drawn, what is the probability that the sample mean will be greater than 2.1 minutes?

**8.3** The strength of the wire produced by company A has a mean of 4,500 kg and a standard deviation of 200 kg. Company B has a mean of 4,000 kg and a standard deviation of 300 kg. If 50 wires of company A and 100 wires of company B are selected at random and tested for strength, what is the probability that the sample mean strength of A will be atleast 600 kg more than that of B? *[Delhi Univ., MBA, 2004]*

**8.4** For a certain aptitude test, it is known from past experience that the average score is 1000 and the standard deviation is 125. If the test is administered to 100 randomly selected individuals, what is the probability that the value of the average score for this sample will lie in the interval 970 and 1030? Assume that the population distribution is normal.

**8.5** A manufacturing process produces ball bearings with mean 5 cm and standard deviation 0.005 cm. A random sample of 9 bearings is selected to measure their average diameter and find it to be 5.004 cm. What is the probability that the average diameter of 9 randomly selected bearings would be at least 5.004 cm?

**8.6** A population of items has an unknown distribution but a known mean and standard deviation of 50 and 100, respectively. Based upon a randomly drawn sample of 81 items drawn from the population, what is the probability that the sample arithmetic mean does not exceed 40?

**8.7** A marketing research team has determined the standard error of sampling distribution of mean for a proposed market research sample size of 100 consumers. However, this standard error is twice the level that the management of the organization considers acceptable. What can be done to achieve an acceptable standard error for mean?

**8.8** Assume that the height of 300 soldiers in an army battalion is normally distributed with mean 68 inches and standard deviation 3 inches. If 80 samples consisting of 25 soldiers each are taken, what would be the expected mean and standard deviation of the resulting sampling distribution of means if the sampling is done (a) with replacement and (b) without replacement?

**8.9** How well have equity mutual funds performed in the past compared with BSE Stock Index? A random sample of 36 funds averages a 16.9 per cent annual investment return for 2001–2 with a standard deviation of 3.6 per cent annual return. The BSE Stock Index grew at an annual average rate of 16.3 per cent over the same period. Do these data show that, on the average, the mutual funds out-performed the BSE Stock Index during this period?

**8.10** The average annual starting salary for an MBA is ₹3,42,000. Assume that for the population of MBA (Marketing majors), the average annual starting salary is $\mu$ = 3,40,000 and the standard deviation is $\sigma$ = 20,000. What is the probability that a simple random sample of MBA (Marketing majors) will have a sample mean within ± ₹2,500 of the population mean for each sample sizes: 50,100 and 200? What is your conclusion? *[Delhi Univ., MBA, 2005]*

## Hints and Answers

**8.1** Given $\mu_{\bar{x}} = \mu = 10$, $\sigma = 0.1$ and $n = 10$. Thus $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.1/\sqrt{5} = 0.047$

$$P[9.95 \le \bar{x} \le 10.05]$$

$$= P\left[\frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}_1}} \le z \le \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}_2}}\right]$$

$$= P\left[\frac{9.95 - 10}{0.047} \le z \le \frac{10.05 - 10}{0.047}\right]$$

$$= P[-1.12 \le z \le 1.12] = P[z \le -1.12] + P[z \le 1.12]$$

$$= 0.3686 + 0.3686 = 0.7372$$

**8.2** Given $\mu_{\bar{x}} = \mu = 2$, $\sigma = 0.25$ and $n = 36$. Thus,

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}\ \frac{3,37,000 - 3,40,000}{20,000/\sqrt{50}} = 0.25/\sqrt{36} = 0.042$$

$$P[\bar{x} \ge 2.1] = P\left[z \ge \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = P\left[z \ge \frac{2.1 - 2}{0.042}\right]$$

$$= P[z \ge 2.38] = 0.5000 - 0.4913 = 0.0087$$

**8.3** Given $\mu_1 = 4500$, $\sigma_1 = 200$ and $n_1 = 50$ ; $\mu_2 = 4000$, $\sigma_2 = 300$ and $n_2 = 100$. Then

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 4500 - 4000 = 500$$

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

$$\text{P}[(\bar{x}_1 - \bar{x}_2) \geq 600] = \text{P}\left[z \geq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}\right]$$

$$= \text{P}\left[z \geq \frac{600 - 500}{41.23}\right] = \text{P}(z \geq 2.43)$$

$$= 0.5000 - 0.4925 = 0.0075$$

**8.4** Given $\mu_{\bar{x}} = \mu = 1000$, $\sigma = 125$ and $n = 100$. Thus,

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 125/\sqrt{100} = 12.5$$

$$\text{P}(970 \leq \bar{x} \leq 1030)$$

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

$$= \text{P}\left[\frac{970 - 1000}{12.5} \leq z \leq \frac{1030 - 1000}{12.5}\right]$$

$$= \text{P}(-2.4 \leq z \leq 2.4) = \text{P}(z \leq 2.4) + \text{P}(z \leq -2.4)$$

$$= 0.4918 + 0.4918 = 0.9836$$

**8.5** Given $= \mu = 5$, $\sigma = 0.005$ and $n = 9$. Thus, $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.005/\sqrt{9} = 0.0017$

$$\text{P}(\bar{x} \geq 5.004) = \text{P}\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = \text{P}\left[z \geq \frac{5.004 - 5.000}{0.0017}\right]$$

$$= \text{P}(z \geq 2.4) = 1 - \text{P}(z \geq 2.4)$$

$$= 1 - 0.9918 = 0.0082$$

**8.6** Given $\mu_{\bar{x}} = \mu = 50$, $\sigma = 100$ and $n = 81$. Thus, $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 100/\sqrt{81} = 11.1$

$$\text{P}(\bar{x} \leq 40) = \text{P}\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = \text{P}\left[z \leq \frac{40 - 50}{11.1}\right]$$

$$= \text{P}(z \leq -0.90) = 0.5000 - 0.3159$$

$$= 0.1841$$

**8.7** Since standard error is inversely proportional to the square root of the sample size, therefore to reduce the standard error determined by the market research team, the sample size should be increased to $n = 400$ (four times of $n = 100$).

**8.8** The number of possible samples of size 25 each from a group of 3000 soldiers with and without replacement are $(3000)^{25}$ and ${}^{300}\text{C}_{25}$, respectively. These numbers are much larger than 80—actually drawn samples. Thus we will get only an experimental sampling distribution of means rather than true sampling distribution. Hence mean and standard deviation would be close to those of the theoretical distribution. That is,

(a) $\mu_{\bar{x}} = \mu = 68$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/\sqrt{25} = 0.60$

(b) $\mu_{\bar{x}} = \mu = 68$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{\text{N} - n}{\text{N} - 1}}$

$$= \frac{3}{\sqrt{25}}\sqrt{\frac{3000 - 25}{3000 - 1}} = 1.19$$

**8.9** Given $\mu_{\bar{x}} = \mu = 16.9$, $\sigma = 3.6$ and $n = 36$. Thus $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3.6/\sqrt{36} = 0.60$

$$\text{P}(\bar{x} \geq 16.3) = \text{P}\left[z \geq \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}\right] = \text{P}\left[z \geq \frac{16.3 - 16.9}{0.60}\right]$$

$$= \text{P}[z \geq -1] = 0.5000 + 0.1587$$

$$= 0.6587$$

**8.10** Given $\mu = 3{,}40{,}000$; $\sigma = 20{,}000$, $n_1 = 50$, $n_2 = 100$, and $n_3 = 200$

For $n_1 = 50$:

$$z_1 = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{3{,}42{,}000 - 3{,}40{,}000}{20{,}000/\sqrt{50}} = 0.88$$

$$z_2 = \frac{3{,}37{,}000 - 3{,}40{,}000}{20{,}000/\sqrt{50}} = -0.88$$

$\text{P}(-0.88 \leq z \leq 0.88) = 0.3106 \times 2 = 0.6212$

Similar calculations for $n_2 = 100$ and $n_2 = 200$ give

$\text{P}(-1.25 \leq z \leq 1.25) = 0.3944 \times 2 = 0.7888$

$\text{P}(-1.76 \leq z \leq 1.76) = 0.4616 \times 2 = 0.9282$

## 8.9 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

There are many situations in which each element of the population can be classified into two mutually exclusive categories such as success or failure, accept or reject, head or tail of a coin, and so on. These situations provide practical examples of binomial experiments under required conditions of binomial probability distribution. If a random sample of $n$ elements is selected from the binomial population and $x$ of these possess the specified characteristic, then the sample proportion is the best statistic to use for statistical inferences about the population proportion parameter $p$. The sample proportion can be defined as

$$\bar{p} = \frac{\text{Elements of sample having characteristic, } x}{\text{Sample size, } n}$$

With the same logic of sampling distribution of mean, the sampling distribution of sample proportions with mean $\mu_{\bar{p}}$ and standard deviation (also called *standard error*) $\sigma_{\bar{p}}$ is given by

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

If the sample size $n$ is large ($n \geq 30$), the sampling distribution of $\bar{p}$ can be approximated by a normal distribution. The approximation will be adequate if $np \geq 5$ and $n(1-p) \geq 5$. It may be noted that the sampling distribution of the proportion would actually follow binomial distribution because population is binomially distributed.

The mean and standard deviation of the sampling distribution of proportion are valid for a finite population in which sampling is with replacement. However, for finite population in which sampling is done without replacement, we have

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

For a large sample size $n$ ($\geq 30$), the sampling distribution of proportion is closely approximated by a normal distribution with mean and standard deviation as stated above. Hence, to standardize sample proportion, $\bar{p}$, the standard normal variable

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately the standard normal distribution.

### 8.9.1 Sampling Distribution of the Difference of Two Proportions

For two populations of size $N_1$ and $N_2$ draw samples of size $n_1$ from first population and compute sample proportion $\bar{p}_1$ and standard deviation $\sigma_{\bar{p}_1}$. Similarly, draw samples of size $n_2$ from second population and compute sample proportion $\bar{p}_2$ and standard deviation $\sigma_{\bar{p}_2}$.

For all combinations of samples drawn from two populations, obtain a sampling distribution of the difference $\bar{p}_1 - \bar{p}_2$ of samples proportions. Such a distribution is called *sampling distribution of difference of two proportions*. The mean and standard deviation of such distribution are given by

$$\mu_{\bar{p}_1-\bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

and

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\sigma^2_{\bar{p}_1} + \sigma^2_{\bar{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If sample size $n_1$ and $n_2$ are large ($n_1 \geq 30$ and $n_2 \geq 30$), then the sampling distribution of difference of proportions is closely approximated by a normal distribution.

**Example 8.14:** A manufacturer of watches has determined from experience that 3 per cent of the watches he produces are defective. If a random sample of 300 watches is examined, what is the probability that the proportion defective is between 0.02 and 0.035?

[*Delhi Univ., MBA, 2005*]

**Solution:** From the data of the problem, we have $\mu_{\bar{p}} = p = 0.03$, $\bar{p}_1 = 0.02$, $\bar{p}_2 = 0.035$ and $n = 300$. The standard error of proportion is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.03 \times 0.97}{300}}$$

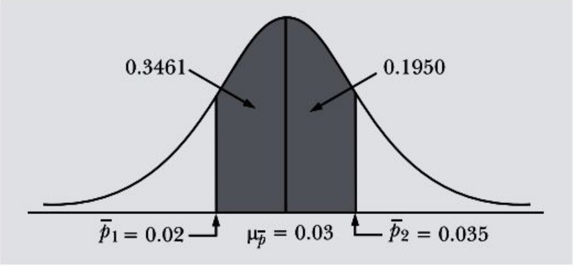
$$= \sqrt{0.000097} = 0.0098$$

For calculating the desired probability, apply the following formula

$$P[0.02 \leq \bar{p} \leq 0.035] = P\left[\frac{\bar{p}_1 - p}{\sigma_{\bar{p}}} \leq z \leq \frac{\bar{p}_2 - p}{\sigma_{\bar{p}}}\right]$$

$$= P\left[\frac{0.02 - 0.03}{0.0098} \leq z \leq \frac{0.035 - 0.03}{0.0098}\right]$$

**Figure 8.11**
Normal Curve



$$= P[-1.02 \le z \le 0.51]$$
$$= P(z \le -1.02) + P(z \le 0.51)$$
$$= 0.3461 + 0.1950 = 0.5411$$

Hence, the probability that the proportion of defectives will lie between 0.02 and 0.035 is 0.5411.

**Example 8.15:** Few years back, a policy was introduced to give loan to unemployed engineers to start their own business. Out of 1,00,000 unemployed engineers, 60,000 accept the policy and got the loan. A sample of 100 unemployed engineers is taken at the time of allotment of loan. What is the probability that sample proportion would have exceeded 50 per cent acceptance?

**Solution:** From the data of the problem, we have $\mu_{\bar{p}} = p = 0.60$, N = 1,00,000 and $n = 100$

The standard error of proportion in a finite population of size 1,00,000 is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.60 \times 0.40}{100}}\sqrt{\frac{1{,}00{,}000-100}{1{,}00{,}000-1}}$$
$$= \sqrt{0.0024}\,\sqrt{0.9990} = 0.0489 \times 0.9995 = 0.0488$$

The probability that sample proportion would have exceeded 50 per cent acceptance is given by

$$P(x \ge 0.50) = P\left[z \ge \frac{\bar{p}-p}{\sigma_{\bar{p}}}\right] = P\left[z \ge \frac{0.50-0.60}{0.0489}\right]$$
$$= P[z \ge -2.04] = 0.5000 + 0.4793 = 0.9793$$

**Example 8.16:** Ten per cent of machines produced by company A are defective and five per cent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B. What is the probability that the difference in sample proportion is less than or equal to 0.02?

[*South Gujarat Univ., MBA, 2000; Delhi Univ., MBA, 2002* ]

**Solution:** From the data of the problem, we have $\mu_{\bar{p}_1-\bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2 = 0.10 - 0.05 = 0.05$; $n_1 = 250$ and $n_2 = 300$.

The standard error of the difference in a sample proportion is given by

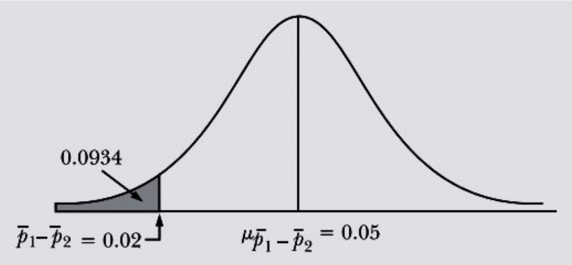
$$\mu_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.10 \times 0.90}{250} + \frac{0.05 \times 0.95}{300}}$$
$$= \sqrt{\frac{0.90}{250} + \frac{0.0475}{300}} = \sqrt{0.00052} = 0.0228$$

The desired probability of difference in sample proportions is given by

$$P[(\bar{p}_1 - \bar{p}_2) \le 0.02] = P\left[z \le \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}}\right]$$
$$= P\left[z \le \frac{0.02 - 0.05}{0.0228}\right]$$
$$= P[z \le -1.32]$$
$$= 0.5000 - 0.4066 = 0.0934$$

Hence, the desired probability for the difference in sample proportions is 0.0934.

Figure 8.12
Normal Curve



## Self-practice Problems 8B

**8.11** Assume that 2 per cent of the items produced in an assembly line operation are defective, but that the firm's production manager is not aware of this situation. What is the probability that in a lot of 400 such items, 3 per cent or more will be defective?

**8.12** If a coin is tossed 20 times and the coin falls on head after any toss, it is a success. Suppose the probability of success is 0.5, what is the probability that the number of successes is less than or equal to 12?

**8.13** The quality control department of a paints manufacturing company, at the time of dispatch of decorative paints, discovered that 30 per cent of the containers are defective. If a random sample of 500 containers is drawn with replacement from the population, what is the probability that the sample proportion will be less than or equal to 25 per cent defective?

**8.14** A manufacturer of screws has found that on an average 0.04 of the screws produced are defective. A random sample of 400 screws is examined for the proportion of defective screws. Find the probability that the proportion of defective screws in the sample is between 0.02 and 0.05.

**8.15** A manager in the billing section of a mobile phone company checks on the proportion of customers who are paying their bills late. Company policy dictates that this proportion should not exceed 20 per cent. Suppose that the proportion of all invoices that were paid late is 20 per cent. In a random sample of 140 invoices, determine the probability that more than 28 per cent invoices were paid late.

## Hints and Answers

**8.11** $\mu_{\bar{p}} = np = 400 \times 0.02 = 8;$

$\sigma_p = \sqrt{npq} = \sqrt{400 \times 0.02 \times 0.98} = 2.8$

and 3 per cent of 400 = 12 defective items. Thus,

$$P(\bar{p} \geq 12) = P\left[z \geq \frac{\bar{p} - np}{\sigma_p}\right] = P\left[z \geq \frac{12-8}{2.8}\right]$$

$$= P(z \geq 1.42) = 0.5000 - 0.4222 = 0.0778$$

**8.12** Given $\mu_{\bar{p}} = np = 20 \times 0.50 = 10;$

$\sigma_{\bar{p}} = \sqrt{npq} = \sqrt{20 \times 0.50 \times 0.50} = 2.24$

$$P(\bar{p} \leq 12) = P\left[z \leq \frac{\bar{p} - np}{\sigma_{\bar{p}}}\right] = P\left[z \leq \frac{12-10}{2.24}\right]$$

$$= P(z \leq 0.89) = 0.8133$$

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

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

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

$$= P[z \leq -2.43] = 0.5000 - 0.4927$$

$$= 0.0083$$

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

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

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

$$= P\left[\frac{0.02 - 0.04}{0.009} \leq z \leq \frac{0.05 - 0.04}{0.009}\right]$$

$$= P[-2.22 \leq z \leq 2.22]$$

$$= P[z \leq -2.22] + P[z \leq 2.22]$$

$$= 0.4861 + 0.4861 = 0.9722$$

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

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

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

$$= P\left[z \geq \frac{0.28 - 0.20}{0.033}\right] = P[z \geq 2.42]$$

$$= 0.5000 - 0.4918 = 0.0082$$

## 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 $n < 0.5\mathrm{N}$ ; $n$, 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

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

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

## Chapter Concepts Quiz

### True or False

1. [T] [F] The sampling distribution provides the basis for statistical inference when sample results are analysed.
2. [T] [F] The sampling distribution of mean is the probability density function that describes the distribution of the possible values of a sample mean.
3. [T] [F] The expected value (mean) is equal to the population mean from which the sample is chosen.
4. [T] [F] As the sample size is increased, the sampling distribution of the mean approaches the normal distribution regardless of the population distribution.
5. [T] [F] A sample size of $n \geq 30$ is considered large enough to apply the central limit theorem.
6. [T] [F] Standard error of the mean is the standard deviation of the sampling distribution of the mean.
7. [T] [F] The finite correction factor may be omitted of $n < 0.5$ N.
8. [T] [F] The principles of the 'inertia of large number' and 'statistical regularity' govern random sampling.
9. [T] [F] Every member of the population is tested in a sample survey.
10. [T] [F] Simple random sampling is non-probability sampling method.
11. [T] [F] Expected value (mean) of samples drawn randomly from a population are always same.
12. [T] [F] The standard error becomes stable with an increase in sample size.
13. [T] [F] The principle of 'inertia of large number' is a corollary of the principle of 'statistical regularity'.
14. [T] [F] Cluster sampling is a non-random sampling method.
15. [T] [F] Quota sampling method is used when the population is widely scattered.

### Multiple Choice Questions

16. Which of the following is the principle on which theory of sampling is based?
(a) Statistical regularity
(b) Inertia of large numbers

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

# Chapter 14

*The cause is hidden, but the result is known.*

—Ovid

*I never think of the future, it comes soon enough.*

—Albert Einstein

# Regression Analysis

## LEARNING OBJECTIVES

After studying this chapter, you should be able to

- use simple linear regression for building models to business data.
- understand how the method of least squares is used to predict values of a dependent (or response) variable based on the values of an independent (or explanatory) variable.
- measure the variability (residual) of the dependent variable about a straight line (also called regression line) and examine whether regression model fits to the data.

## 14.1 INTRODUCTION

In Chapter 13, we introduced the concept of statistical relationship between two variables such as level of sales and amount of advertising; yield of a crop and the amount of fertilizer used; price of a product and its supply, and so on. Such statistical relationship indicates the degree (strength) and direction of association between two variables but fails to ascertain whether there is any functional (or algebraic) relationship between two variables? If yes, can it be used to estimate the most likely value of one variable, given the value of other variable?

The statistical technique that expresses a functional (or algebraic) relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called *regression analysis*. The variable whose value is to be estimated is called *dependent* (or *response*) *variable* and the variable whose value is used to estimate this value is called *independent* (regressor or *predictor*) *variable*. The linear algebraic equations that express a dependent variable in terms of an independent variable are called *linear regression equation*.

Sir Francis Galton in 1877, while studying the relationship between the height of father and sons found that though 'tall father has tall sons', the average height of sons of tall father is $x$ above the general height and the average height of sons is $2x/3$ above the general height. He described such a fall in the average height as 'regression to mediocrity'. The term regression in the literary sense is also referred as ***'moving backward'***.

### Difference Between Correlation and Regression Analysis

1. Developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable is referred to as regression analysis.

2. Measuring the strength (or degree) and direction of the relationship between two variables is referred as correlation analysis. The direction (direct or inverse) of the relationship is indicated by the correlation coefficient, and the absolute value of correlation coefficient indicates the extent (strength or degree) of the relationship.
3. Correlation analysis determines the strength (or degree) of association between two variables $x$ and $y$ but does not establish a cause-and-effect relationship. Regression analysis establishes the cause-and-effect relationship between $x$ and $y$, that is, a change in the value of independent variable $x$ *causes* a change (*effect*) in the value of dependent variable, $y$ assuming that all other factors that may affect $y$ remain unchanged.
4. In linear regression analysis one variable is considered as dependent variable and other as independent variable, while in correlation analysis both variables are considered to be independent.
5. *The coefficient of determination $r^2$ indicates the proportion of total variance in the dependent variable that is explained or accounted for due to variation in the independent variable.* Since value of $r^2$ is determined from a sample, its value is subject to sampling error.

## 14.2 ADVANTAGES OF REGRESSION ANALYSIS

The following are few advantages of regression analysis:

1. Regression analysis helps in developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable around the regression line. Closer the pair of values ($x$, $y$) fall around the regression line, better the line fits the data and hence smaller the variance and error of estimate. Thus, a good estimate can be made of the value of variable $y$ when all the points fall on the line, i.e. standard error of estimate equals zero.
3. If the sample size is large ($n \geq 30$), then interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either $x$ or $y$. The magnitude of $r^2$ remains the same regardless of the values of the two variables.

## 14.3 TYPES OF REGRESSION MODELS

A regression model is an algebraic equation between two variables based on the given data and estimating the value of a dependent variable based on the known values of one or more independent variables. A particular form of regression model depends upon the nature of the problem under study and the type of data available.

### 14.3.1 Simple and Multiple Regression Models

If a regression model represents the relationship between a dependent, $y$, and only one independent variable, $x$, then such a regression model is called a *simple regression model*. But if more than one independent variable is associated with a dependent variable, then such a regression model is called a *multiple regression model*. For example, sales turnover of a product (a dependent variable) is associated with more than one independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors and so on. Thus, estimation of possible sales turnover with respect to only one of these independent variables is an example of a simple regression model, otherwise multiple regression model.

### 14.3.2 Linear and Non-linear Regression Models

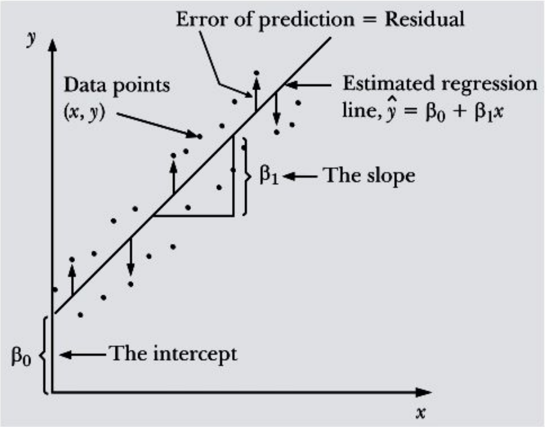
If the change (increase or decrease) in the values of a dependent (response) variable $y$ in a regression model is directly proportional to a unit change (increase or decrease) in the values of independent (predictor) variable $x$, then such a model is called a **linear regression model**. Thus, the relationship between these two variables can be represented by a straight-line relationship in terms of population parameters $\beta_0$ and $\beta_1$ as follows:

$$E(y) = \beta_0 + \beta_1 x \tag{14-1}$$

where $\beta_0$ = $y$-intercept that represents mean (or average) value of the dependent variable $y$ when $x = 0$.

$\beta_1$ = slope of the regression line that represents the expected change (positive or negative) in the value of dependent variable, $y$ for a unit change in the value of independent variable, $x$.

**Figure 14.1**
Straight Line Relationship


Error of prediction = Residual
$y$
Data points
$(x, y)$
Estimated regression line, $\hat{y} = \beta_0 + \beta_1 x$
$\beta_1$ ← The slope
$\beta_0$
← The intercept
$x$

The intercept $\beta_0$ and the slope $\beta_1$ are *unknown regression coefficients*. The value of both $\beta_0$ and $\beta_1$ is to be calculated to predict average value of $y$ for a given value of $x$ by substituting these values in equation (14-1).

Figure 14.1 presents a scatter diagram of each pair of values $(x_i, y_i)$ around the regression line. Although, mean (or average) value of dependent variable, $y$, is a linear function of independent variable, $x$, but not all values of $y$ fall exactly on the straight line. Since few points do not fall on the regression line, therefore values of $y$ are not exactly equal to the values obtained by equation (14-1). Thus, such a straight line is also called ***line of mean deviations*** of observed $y$ value from the regression line. This situation arises due to *random error* (also called *residual variation or residual error*) in the prediction of the value of dependent variable $y$ for given value of independent variable, $x$. This implies that the variable, $x$, is not alone responsible for all variability in the value of variable, $y$. For example, sales volume is related to the level of expenditure on advertisement, but if other factors related to sales such as price of the product, quality of the product, competitors, etc., are ignored, then a regression equation to predict the sales volume ($y$) based on budget of advertising ($x$) only may cause an error. Thus for a fixed value of independent variable, $x$, the actual value of dependent variable, $y$, is determined by the ***mean value function plus a random error term,*** $e$, as follows:

$$\begin{aligned} y &= \text{Mean value function} + \text{Deviation} \\ &= \beta_0 + \beta_1 x + e \end{aligned} \tag{14-2}$$

The equation (14-2) is referred to as **simple probabilistic linear regression model**. The error term, $e$, in equation (14-2) is called *random error* because its value associated with each value of variable, $y$, is assumed to vary unpredictably. The extent of random error associated with each value of variable, $y$, for a given value of $x$ is measured by the error variance. Lower the value of, $e$, better the regression model fit to a sample data.

The random errors corresponding to different observations $(x_i, y_i)$ for all $i$ are assumed to follow a normal distribution with mean zero and (unknown) constant standard deviation.

If the line passing through the pair of values of variables $x$ and $y$ is not linear, then the relationship between variables $x$ and $y$ is *non-linear*. A non-linear relationship implies that expected change (positive or negative) in the value of dependent variable, $y$, is not directly proportional to a unit change in the value of independent variable, $x$. A non-linear relationship is not very useful for predictions.

In this chapter, we will discuss methods of simple linear regression analysis involving single independent variable, whereas those involving two or more independent variables will be discussed in Chapter 15.

## 14.4 ESTIMATION: THE METHOD OF LEAST SQUARES

A sample of $n$ pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ is drawn from the population under study to estimate the values of regression coefficients $\beta_0$ and $\beta_1$. The method that provides the best linear unbiased estimate of the values of $\beta_0$ and $\beta_1$ is called the **method of least squares**. The estimated values of $\beta_0$ and $\beta_1$ should result in a straight line where most pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ fall very close (*best fit*) to it. Such a straight line is referred to as *'best fitted' (least squares or estimated) regression line because the sum of the squares of the vertical deviations (difference between the actual values of y and the estimated values predicted from the fitted line) is as small as possible.*

Rewriting equation (14-2) as follows:

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ or } e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for all } i$$

Mathematically, we intend to minimize

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let $b_0$ and $b_1$ be the least-squares estimators of $\beta_0$ and $\beta_1$, respectively. The least-squares estimators $b_0$ and $b_1$ must satisfy following equations:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{b_0 b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{b_0 b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)\, x_i = 0$$

After simplifying these two equations, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \qquad (14\text{-}3)$$

$$\sum_{i=1}^{n} x_i\, y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Equation (14-3) is called the *least-squares normal equation*. The values of least squares estimators $b_0$ and $b_1$ can be obtained by solving equation (14-3). Hence, the *fitted* or *estimated regression line* is given by

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ (called $y$ hat) is the *predicted value of y* falling on the fitted regression line for a given value of $x$ and $e_i = y_i - \hat{y}_i$ is called the *residual* that describes the amount of error in fitting of the regression line to the values of $y_i$.

**Remark:** $\sum e_i = 0$, i.e. sum of the residuals is zero for any least-squares regression line.

## 14.5 ASSUMPTIONS FOR A SIMPLE LINEAR REGRESSION MODEL

To form a basis for application of simple linear regression models, certain assumptions about the population from which a sample of observations is drawn and the way in which observations are made to draw statistical inference are illustrated in Figure 14.2.

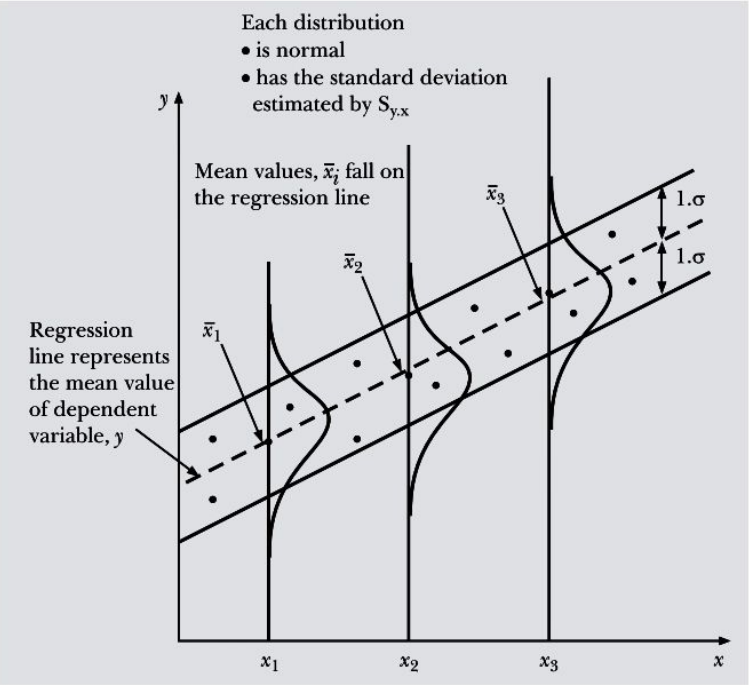

Each distribution
• is normal
• has the standard deviation estimated by $S_{y.x}$
Mean values, $\bar{x}_i$ fall on the regression line
Regression line represents the mean value of dependent variable, y
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
1.σ
1.σ
y
$x_1$
$x_2$
$x_3$
x

**Figure 14.2** Assumptions in Regression Analysis

### Assumptions

1. There is a linear relationship between the dependent variable $y$ and independent variable $x$. This relationship can be described by a linear regression equation $y = a + bx + e$, where $e$ represents the deviation in the value of dependent variable, $y$, from its expected value for a given value of independent variable, $x$.
2. The set of expected (or mean) values of the dependent variable, $y$, for given values of independent variable, $x$, are normally distributed. The mean of these normally distributed values falls on the line of regression.
3. The dependent variable $y$ is a continuous random variable, whereas values of the independent variable $x$ are fixed and not random.
4. The sampling error associated with the expected value of the dependent variable $y$ is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The amount of deviation (error) in the value of dependent variable, $y$, may be different in successive observations.
5. The standard deviation and variance of expected values of the dependent variable about the regression line are constant for all values of the independent variable $x$ for the set of observations in a sample.
6. The expected value of the dependent variable cannot be obtained for a value of an independent variable falling outside the range of values in the sample.

## 14.6 PARAMETERS OF SIMPLE LINEAR REGRESSION MODEL

The objective regression analysis is to ascertain that a regression equation (line) should provide the best fit of sample data to the population data so that the error of variance is as small as possible. J. R. Stockton stated that *the device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression*.

The two variables $x$ and $y$ which are correlated can be expressed in terms of each other in the form of straight line equations called *regression equations* as follows:

- The regression equation of $y$ on $x$

$$y = a + bx$$

is used for estimating the value of $y$ for given values of $x$.

- Regression equation of $x$ on $y$

$$x = c + dy$$

is used for estimating the value of $x$ for given values of $y$.

### Remarks

1. Regression lines coincide (overlap) when variables $x$ and $y$ are perfectly correlated (either positive or negative).
2. Higher the degree of correlation, the two regression lines come closer to each other. Lesser the degree of correlation, the two regression lines are away from each other. Also, when variables $x$ and $y$ are not correlated, i.e. $r = 0$, the two regression lines are at right angle to each other.
3. The point of interaction of two linear regression lines represents the average value of variables $x$ and $y$.

## 14.6.1 Regression Coefficients

To estimate values of population parameter $\beta_0$ and $\beta_1$, the fitted (or estimated) simple linear regression model (equation or line) is written as

$$\hat{y} = a + bx$$

where $\hat{y}$ is estimated average (mean) value of dependent variable $y$ for a given value of independent variable $x$; $a$ or $b_0$ is $y$-intercept that represents average value of $\overline{y}$; and $b$ is the slope of regression line that represents the expected change in the value of $y$ for unit change in the value of $x$.

To determine the value of for a given value of $x$, this equation requires the determination of two unknown constants $a$ (intercept) and $b$ (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable $y$ for a given value of independent variable $x$.

The particular values of $a$ and $b$ define a specific linear relationship between $x$ and $y$ based on sample data. The coefficient '$a$' represents the *level of fitted line* (i.e., the distance of the line above or below the origin) when $x$ equals zero, whereas coefficient '$b$' represents the *slope of the line* (a measure of the change in the estimated value of $y$ for a one-unit change in $x$).

The regression coefficient '$b$' is also denoted as

- $b_{yx}$ (*regression coefficient of y on x*) in the regression line, $y = a + bx$
- $b_{xy}$ (*regression coefficient of x on y*) in the regression line, $x = c + dy$

### Properties of Regression Coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, i.e., $r = \sqrt{b_{yx} \times b_{xy}}$ .
2. If one regression coefficient is greater than one, then other regression coefficient must be less than one because the value of correlation coefficient $r$ cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the case of opposite sign of two regression coefficients.
4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \times 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients $b_{xy}$ and $b_{yx}$ is more than or equal to the correlation coefficient, $r$, i.e., $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and

$b_{xy}$ = – 0.234, then the arithmetic mean of these two values is (– 0.664 – 0.234)/2 = – 0.449, and this value is more than the value of $r$ = – 0.394.

6. Regression coefficients are independent of origin but not of scale.

## 14.7 METHODS TO DETERMINE REGRESSION COEFFICIENTS

The following are the methods to determine the parameters of a fitted regression equation.

### 14.7.1 Least Squares Normal Equations

Let $\hat{y} = a + bx$ be the least squares line of $y$ on $x$, where $\hat{y}$ is the estimated average value of dependent variable $y$. Since best-fitted least squares line minimizes the sum of squares of the deviations of the observed values of $y$ from $\hat{y}$, therefore sum of residuals for any least-square line is minimum, i.e.,

$$L = \Sigma\,(y - \hat{y})^2 = \Sigma\{y - (a + bx)\}^2 = \text{minimum, where } a, b = \text{constants}$$

Differentiating $L$ with respect to $a$ and $b$ and equating to zero, we have

$$\frac{\partial \text{L}}{\partial a} = -2\Sigma\{y - (a + bx)\} = 0$$

$$\frac{\partial \text{S}}{\partial b} = -2\Sigma\{y - (a + bx)\}x = 0$$

Solving these two equations, we get the same set of equations as equations (14-3)

$$\begin{aligned} \Sigma\, y &= n\,a + b\Sigma\, x \qquad (14\text{-}4)\\ \Sigma\, xy &= a\,\Sigma x + b\Sigma\, x^2 \end{aligned}$$

where $n$ is the total number of pairs of values of $x$ and $y$ in a sample data. The equation (14-4) is called *normal equations* with respect to the regression line of $y$ on $x$. After solving these equations for $a$ and $b$, the values of $a$ and $b$ are substituted in the regression equation, $y = a + bx$.

Similarly if a least squares line is $x = c + dy$ of $x$ on $y$, where $x$ is the estimated average value of dependent variable $x$, then the normal equations will be

$$\begin{aligned} \Sigma\, x &= nc + d\,\Sigma\, y\\ \Sigma\, xy &= n\,\Sigma\, y + d\,\Sigma\, y^2 \end{aligned}$$

These equations are solved for constants $c$ and $d$. The values of these constants are substituted to the regression equation $x = c + dy$.

#### Alternative Method to Calculate Value of Constants

Instead of using the algebraic method to calculate values of constants $a$ and $b$ or $c$ and $d$, we may directly use the results of the solutions of these normal equations.

The gradient '$b$' (regression coefficient of $y$ on $x$) and '$d$' (regression coefficient of $x$ on $y$) are calculated as

$$b = \frac{\text{S}_{xy}}{\text{S}_{xx}}, \quad \text{where} \quad \text{S}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$\text{S}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

and $$d = \frac{\text{S}_{yx}}{\text{S}_{yy}}, \quad \text{where} \quad \text{S}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y\right)^2$$

Since the regression line passes through the point ($\bar{x}, \bar{y}$), the mean values of $x$ and $y$ and the regression equations can be used to find the value of constants $a$ and $c$ as follows:

$$a = \bar{y} - b\bar{x} \text{ for regression equation of } y \text{ on } x$$

$$c = \bar{x} - d\,\bar{y} \text{ for regression equation of } x \text{ on } y$$

The calculated values of $a$, $b$ and $c$, $d$ are substituted in the regression line $y = a + bx$ and $x = c + dy$, respectively, to determine the exact relationship.

**Example 14.1:** Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 per cent in advertising expenditure.

| *Firm* | *Annual Percentage Increase in Advertising Expenditure* | *Annual Percentage Increase in Sales Revenue* |
|---|---|---|
| A | 1 | 1 |
| B | 3 | 2 |
| C | 4 | 2 |
| D | 6 | 4 |
| E | 8 | 6 |
| F | 9 | 8 |
| G | 11 | 8 |
| H | 14 | 9 |

**Solution:** Assume sales revenue ($y$) is dependent on advertising expenditure ($x$). Calculations for regression line using following normal equations are shown in Table 14.1

$$\Sigma y = na + b\Sigma x \text{ and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

**Table 14.1** Calculation for Normal Equations

| *Sales Revenue* $y$ | *Advertising Expenditure*, $x$ | $x^2$ | $xy$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 3 | 9 | 6 |
| 2 | 4 | 16 | 8 |
| 4 | 6 | 36 | 24 |
| 6 | 8 | 64 | 48 |
| 8 | 9 | 81 | 72 |
| 8 | 11 | 121 | 88 |
| 9 | 14 | 196 | 126 |
| 40 | 56 | 524 | 373 |

*Normal Equations Approach:*

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 40 = 8a + 56b$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{or} \quad 373 = 56a + 524b$$

Solving these equations, we get $a = 0.072$ and $b = 0.704$

Substituting these values in the regression equation

$$y = a + bx = 0.072 + 0.704x$$

For $x = 7.5$ per cent or 0.075 an increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704\,(0.075) = 0.1248 \text{ or } 12.48\%$$

*Short-cut Method*

$$b = \frac{S_{xy}}{S_{xx}} = \frac{93}{132} = 0.704,$$

where
$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 373 - \frac{40 \times 56}{8} = 93$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 524 - \frac{(56)^2}{8} = 132$$

The intercept '*a*' on the *y*-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{40}{8} - 0.704 \times \frac{56}{8} = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of $a = 0.072$ and $b = 0.704$ in the regression equation, we get

$$y = a + bx = 0.072 + 0.704\,x$$

For $x = 0.075$, we have $y = 0.072 + 0.704\ (0.075) = 0.1248$ or 12.48 per cent.

**Example 14.2:** The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in ₹1000's and analysed the results

| Week : | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Sales : | 2.69 | 2.62 | 2.80 | 2.70 | 2.75 | 2.81 |

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

**Solution:** Assume sales (*y*) are dependent on weeks (*x*). Then the normal equations for regression equation: $y = a + bx$ are written as

$$\Sigma y = n\,a + b\,\Sigma x \text{ and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 14.2.

**Table 14.2** Calculations of Normal Equations

| *Week* (*x*) | *Sales* (*y*) | $x^2$ | $xy$ |
|---|---|---|---|
| 1 | 2.69 | 1 | 2.69 |
| 2 | 2.62 | 4 | 5.24 |
| 3 | 2.80 | 9 | 8.40 |
| 4 | 2.70 | 16 | 10.80 |
| 5 | 2.75 | 25 | 13.75 |
| 6 | 2.81 | 36 | 16.86 |
| 21 | 16.37 | 91 | 57.74 |

The gradient '*b*' is calculated as

$$b = \frac{S_{xy}}{S_{xx}} = \frac{0.445}{17.5} = 0.025; \quad S_{xy} = \Sigma xy - \frac{\Sigma x\,\Sigma y}{n} = 57.74 - \frac{21 \times 16.37}{6} = 0.445$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 91 - \frac{(21)^2}{6} = 17.5$$

The intercept '*a*' on the *y*-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 \times \frac{21}{6}$$

$$= 2.728 - 0.025 \times 3.5 = 2.64$$

Substituting the values $a = 2.64$ and $b = 0.025$ in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For $x = 7$, we have $y = 2.64 + 0.025(7) = 2.815$

Hence, the expected sales during the 7th week are likely to be ₹2.815 (in ₹1000's).

### 14.7.2 Deviations Method

Computation time while using least squares normal equations method becomes lengthy when values of *x* and *y* are in more than two digits. The computational time may be reduced by using following two methods:

**(a) Deviations Taken from Actual Mean Values of *x* and *y*** If deviations of actual values of variables *x* and *y* are taken from their mean values, then regression equations can be written as

- Regression equation of *y* on *x*

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where $b_{yx}$ = regression coefficient of *y* on *x*

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

- Regression equation of *x* on *y*

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where $b_{xy}$ = regression coefficient of *x* on *y*

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

**(b) Deviations Taken from Assumed Mean Values for *x* and *y*** If mean value of either *x* or *y* or both are not integer, then prefer to take deviations of actual values of variables *x* and *y* from their assumed means.

- Regression equation of *y* on *x*

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{where } b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2}$$

$n$ = number of observations
$d_x = x - A$; A is assumed mean of *x*
$d_y = y - B$; B is assumed mean of *y*

- Regression equation of *x* on *y*

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{where } b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2}$$

$n$ = number of observations
$dx = x - A$; A is assumed mean of *x*
$dy = y - B$; B is assumed mean of *y*

**(c) Regression Coefficients in Terms of Correlation Coefficient** If deviations are taken from actual mean values, then the values of regression coefficients can be calculated as follows:

$$byx = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\text{Covariance}(x, y)}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$bxy = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\text{Covariance}(x, y)}{\sigma_y^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

**Example 14.3:** The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (₹ in 1000's)

| Salesmen | : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Test scores | : | 50 | 60 | 50 | 60 | 80 | 50 | 80 | 40 | 70 |
| Weekly sales | : | 30 | 60 | 40 | 50 | 60 | 30 | 70 | 50 | 60 |

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales. *[HP Univ., M.Com.,2006 ]*

**Solution:** Assume weekly sales (*y*) as dependent variable and test scores (*x*) as independent variable. Calculations for the following regression equation are shown in Table 14.3.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

**Table 14.3** Calculation for Regression Equation

| *Weekly Sales, x* | $dx = x - 60$ | $d_x^2$ | *Test Score, y* | $dy = y - 50$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 50 | –10 | 100 | 30 | –20 | 400 | 200 |
| 60 | 0 | 0 | 60 | 10 | 100 | 0 |
| 50 | –10 | 100 | 40 | –10 | 100 | 100 |
| 60 | 0 | 0 | 50 | 0 | 0 | 0 |
| 80 | 20 | 400 | 60 | 10 | 100 | 200 |
| 50 | –10 | 100 | 30 | –20 | 400 | 200 |
| 80 | 20 | 400 | 70 | 20 | 400 | 400 |
| 40 | –20 | 400 | 50 | 0 | 0 | 0 |
| 70 | 10 | 100 | 60 | 10 | 100 | 100 |
| 540 | 0 | 1600 | 450 | 0 | 1600 | 1200 |

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{540}{9} = 60;\quad \bar{y} = \frac{\Sigma y}{n} = \frac{450}{9} = 50$

$$b_{yx} = \frac{\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1200}{1600} = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75\,(x - 60) \text{ or } y = 5 + 0.75x$$

For test score $x = 65$ of salesman, we have

$$y = 5 + 0.75\,(65) = 53.75$$

Hence, we conclude that the weekly sales are expected to be ₹53.75 (₹ in 1000's) for a test score of 65.

**Example 14.4:** A company is introducing a job evaluation scheme in which all jobs are graded by points for skill, responsibility, and so on. Monthly pay scales (₹ in 1000's) are then drawn up according to the number of points allocated and other factors such as experience and local conditions. To date the company has applied this scheme to 9 jobs:

| Job | : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Points | : | 5 | 25 | 7 | 19 | 10 | 12 | 15 | 28 | 16 |
| Pay (₹) | : | 3.0 | 5.0 | 3.25 | 6.5 | 5.5 | 5.6 | 6.0 | 7.2 | 6.1 |

(a) Find the least-squares regression line for linking pay scales to points.

(b) Estimate the monthly pay for a job graded by 20 points.

**Solution:** Assume monthly pay ($y$) as the dependent variable and job grade points ($x$) as the independent variable. Calculations for the following regression equation are shown in Table 14.4.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

**Table 14.4** Calculations for Regression Equation

| *Grade Points, x* | $d_x = x - 15$ | $d_x^2$ | *Pay Scale, y* | $d_y = y - 5$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 5 | –10 | 100 | 3.0 | –2.0 | 4 | 20 |
| 25 | 10 | 100 | 5.0 ← B | 0 | 0 | 0 |
| 7 | –8 | 64 | 3.25 | –1.75 | 3.06 | 14 |
| 19 | 4 | 16 | 6.5 | 1.50 | 2.25 | 6 |
| 10 | –5 | 25 | 5.5 | 0.50 | 0.25 | –2.5 |
| 12 | –3 | 9 | 5.6 | 0.60 | 0.36 | –1.8 |
| 15 ← A | 0 | 0 | 6.0 | 1.00 | 1.00 | 0 |
| 28 | 13 | 169 | 7.2 | 2.2 | 4.84 | 28.6 |
| 16 | 1 | 1 | 6.1 | 1.1 | 1.21 | 1.1 |
| 137 | 2 | 484 | 48.15 | 3.15 | 16.97 | 65.40 |

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{137}{9} = 15.22;\ \bar{y} = \frac{\Sigma y}{n} = \frac{48.15}{9} = 5.35$

Since mean values $\bar{x}$ and $\bar{y}$ are non-integer value, therefore deviations are taken from assumed mean as shown in Table 14.4.

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9 \times 65.40 - 2 \times 3.15}{9 \times 484 - (2)^2} = \frac{582.3}{4352} = 0.133$$

Substituting values in the regression equation, we have

$$y - \bar{y} = b_{yx}(x - \bar{x}) \text{ or } y - 5.35 = 0.133(x - 15.22) = 3.326 + 0.133x$$

(b) For job grade point $x = 20$, the estimated average pay scale is given by

$$y = 3.326 + 0.133x = 3.326 + 0.133\,(20) = 5.986$$

Hence, likely monthly pay for a job with grade points 20 is ₹5986.

**Example 14.5:** The following data, based on 450 students, are given for marks is Statistics and Economics at a certain examination:

| | | |
|---|---|---|
| Mean marks in Statistics | : | 40 |
| Mean marks in Economics | : | 48 |
| S.D. of marks in Statistics | : | 12 |
| The variance of marks in Economics | : | 256 |
| Sum of the product of deviation of marks from their respective mean | : | 42075 |

Obtain equations of the two lines of regression and estimate the average marks in Economics of candidates who obtained 50 marks in Statistics. [*Nagpur Univ., M.Com., 1996*]

**Solution:** Let the marks in Statistics be denoted by $x$ and marks in Economics by $y$. Then given that

$$\bar{x} = 40,\ \bar{y} = 48,\ \sigma_x = 12,\ \sigma_y = \sqrt{256} = 16.$$

Regression equation of $x$ on $y$ : $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

$$x - 40 = 0.487\frac{12}{16}(y - 48),$$
$$= 0.365\,y - 17.52 \text{ or } x = 22.48 + 0.365y.$$

where, $r = \frac{\Sigma d_x d_y}{n\,\sigma_x \sigma_y} = \frac{42075}{450 \times 12 \times 16} = 0.487$

Regression equation of $y$ on $x$ : $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$y - 48 = 0.487\frac{16}{12}(x - 40)$$
$$= 0.649x - 25.96 \text{ or } y = 22.04 + 0.649x.$$

The estimated marks in Economics for a candidate who has obtained $x = 50$ marks in Statistics will be

$$y = 22.04 + 0.649\,(50) = 54.49.$$

**Example 14.6:** The following data give the ages and blood pressure of 10 women.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | : | 56 | 42 | 36 | 47 | 49 | 42 | 60 | 72 | 63 | 55 |
| Blood pressure | : | 147 | 125 | 118 | 128 | 145 | 140 | 155 | 160 | 149 | 150 |

(a) Find the correlation coefficient between age and blood pressure.
(b) Determine the least-squares regression equation of blood pressure on age.
(c) Estimate the blood pressure of a woman whose age is 45 years.

[*Ranchi Univ. MBA*, 2003; *South Gujarat Univ., MBA, 2007*]

**Solution:** Assume blood pressure ($y$) as the dependent variable and age ($x$) as the independent variable. Calculations for regression equation of blood pressure on age are shown in Table 14.5.

**Table 14.5** Calculations for Regression Equation

| *Age, x* | $d_x = x - 49$ | $d_x^2$ | *Blood, y* | $d_y = y - 145$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 56 | 7 | 49 | 147 | 2 | 4 | 14 |
| 42 | −7 | 49 | 125 | −20 | 400 | 140 |
| 36 | −13 | 169 | 118 | −27 | 729 | 351 |
| 47 | −2 | 4 | 128 | −17 | 289 | 34 |
| $\textcircled{49} \leftarrow$ A | 0 | 0 | $\textcircled{145} \leftarrow$ B | 0 | 0 | 0 |
| 42 | −7 | 49 | 140 | −5 | 25 | 35 |
| 60 | 11 | 121 | 155 | 10 | 100 | 110 |
| 72 | 23 | 529 | 160 | 15 | 225 | 345 |
| 63 | 14 | 196 | 149 | 4 | 16 | 56 |
| 55 | 6 | 36 | 150 | 5 | 25 | 30 |
| 522 | 32 | 1202 | 1417 | −33 | 1813 | 1115 |

(a) Coefficient of correlation between age and blood pressure is given by

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

$$= \frac{11150 + 1056}{\sqrt{12020 - 1024}\sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892$$

We may conclude that there is a high degree of positive correlation between age and blood pressure.

(b) The regression equation of blood pressure on age is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{522}{10} = 52.2; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{1417}{10} = 141.7$$

and
$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$$

Substituting these values in the above equation, we have

$$y - 141.7 = 1.11\,(x - 52.2) \text{ or } y = 83.758 + 1.11x$$

This is the required regression equation of $y$ on $x$.

(*c*) For a women whose age is 45, the estimated average blood pressure will be

$$y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the likely blood pressure of a woman of 45 years is 134.

**Example 14.7:** The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of readymade men's wear—is toying with the idea of increasing his sales to ₹80,000. On checking the records of sales during the last 10 years, it was found that the annual sale proceeds and advertisement expenditure were highly correlated to the extent of 0.8. It was further noted that the annual average sale has been ₹45,000 and annual average advertisement expenditure ₹30,000, with a variance of ₹1600 and ₹625 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement would you suggest the General Sales Manager of the enterprise to incur to meet his target of sales?

[*Kurukshetra Univ., MBA, 2008*]

**Solution:** Assume advertisement expenditure ($y$) as the dependent variable and sales ($x$) as the independent variable. Then the regression equation advertisement expenditure on sales is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Given $r = 0.8$, $\sigma_x = 40$, $\sigma_y = 25$, $\bar{x} = 45{,}000$, $\bar{y} = 30{,}000$. Substituting these values in the above equation, we have

$$(y - 30{,}000) = 0.8\,\frac{25}{40}\,(x - 45{,}000) = 0.5(x - 45{,}000)$$

$$y = 30{,}000 + 0.5x - 22{,}500 = 7500 + 0.5x$$

When a sales target is fixed at $x = 80{,}000$, the estimated amount likely to the spent on advertisement would be

$$y = 7500 + 0.5\times 80{,}000 = 7500 + 40{,}000 = ₹47{,}500$$

**Example 14.8:** Find the most likely production corresponding to a rainfall of 40″ from the following data:

| | *Rainfall (x)* | *Production (y)* |
|---|---|---|
| Average | 30″ | 500 kg |
| Standard deviation | 5″ | 100 kg |

Coefficient of correlation, $r = 0.8$ *[GGSIP Univ., MBA, 2002]*

**Solution:** Given that, $r = 0.8$, $\bar{x} = 30$, $\bar{y} = 500$, $\sigma_y = 100$, and $\sigma_x = 5$.

The regression equation of production ($y$) on rainfall ($x$) is:

$$\begin{aligned} y - \bar{y} &= r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \\ y - 500 &= 0.8\frac{100}{5}(x - 30) \\ &= 16x - 480 \text{ or } y = 20 + 16x. \end{aligned}$$

When rainfall, $x = 40''$, the likely production will be $y = 16(40) + 20 = 660$ kg.

**Example 14.9:** You are given the following information about advertising expenditure and sales:

| | *Advertisement (x)*<br>*(Rs in lakh)* | *Sales (y)*<br>*(Rs in lakh)* |
|---|---|---|
| Arithmetic mean, $\bar{x}$ | 10 | 90 |
| Standard deviation, $\sigma$ | 3 | 12 |

Correlation coefficient = 0.8

(a) Obtain the two regression equations.

(b) Find the likely sales when advertisement budget is ₹15 lakh.

(c) What should be the advertisement budget if the company wants to attain sales target of ₹120 lakh? *[Kumaon Univ., MBA, 2000, MBA, Delhi Univ., 2002]*

**Solution:** (*a*) Regression equation of $x$ on $y$ is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Given $\bar{x} = 10$, $r = 0.8$, $\sigma_x = 3$, $\sigma_y = 12$, $\bar{y} = 90$. Substituting these values in the above regression equation, we have

$$x - 10 = 0.8\,\frac{3}{12}\,(y - 90) \text{ or } x = -8 + 0.2y$$

Regression equation of $y$ on $x$ is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 90 = 0.8\ \frac{12}{3} \quad \text{or } y = 58 + 3.2x$$

(b) Substituting $x = 15$ in regression equation of $y$ on $x$. The likely average sales volume would be

$$y = 58 + 3.2(15) = 58 + 48 = 106$$

Thus the likely sales for advertisement budget of ₹15 lakh is ₹106 lakh.

(c) Substituting $y = 120$ in the regression equation of $x$ on $y$. The likely advertisement budget to attain desired sales target of ₹120 lakh would be

$$x = -8 + 0.2\ y = -8 + 0.2(120) = 16$$

Hence, the likely advertisement budget of ₹16 lakh should be sufficient to attain the sales target of ₹120 lakh.

**Example 14.10:** In a partially destroyed laboratory record of an analysis of regression data, the following results only are legible:

Variance of $x = 9$

Regression equations: $8x - 10y + 66 = 0$ and $40x - 18y = 214$

Find on the basis of the above information:

(a) The mean values of $x$ and $y$.

(b) Coefficient of correlation between $x$ and $y$.

(c) Standard deviation of $y$. [*Pune Univ., MBA, 2006*]

**Solution:** (a) Since two regression lines always intersect at a point $(\bar{x}, \bar{y})$ representing mean values of the variables involved, solving given regression equations to get the mean values $\bar{x}$ and $\bar{y}$ as shown below:

$$8x - 10y = -66$$
$$40x - 18y = 214$$

Multiplying the first equation by 5 and subtracting from the second, we have

$$32y = 544 \text{ or } y = 17, \text{ i.e. } \bar{y} = 17$$

Substituting the value of $y$ in the first equation, we get

$$8x - 10(17) = -66 \text{ or } x = 13, \text{ i.e. } \bar{x} = 13$$

(b) To find correlation coefficient $r$ between $x$ and $y$, we need to determine the regression coefficients $b_{xy}$ and $b_{yx}$.

Rewriting the given regression equations in such a way that the coefficient of dependent variable is less than one at least in one equation.

$$8x - 10y = -66 \text{ or } 10y = 66 + 8x \text{ or } y = \frac{66}{10} + \frac{8}{10}x$$

That is, $b_{yx} = 8/10 = 0.80$

$$40x - 18y = 214 \text{ or } 40x = 214 + 18y \text{ or } x = \frac{214}{40} + \frac{18}{40}y$$

That is, $b_{xy} = 18/40 = 0.45$

Hence coefficient of correlation $r$ between $x$ and $y$ is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.80} = 0.60$$

(c) To determine the standard deviation of $y$, consider the formula:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \text{ or } \sigma_y = \frac{b_{yx}\,\sigma_x}{r} = \frac{0.80 \times 3}{0.6} = 4$$

**Example 14.11:** There are two series of index numbers, P for price index and S for stock of a commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4, respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off values of S for various values of P?

**Solution:** The regression equation to read off values of P for various values S is given by

$$P = a + bS \text{ or } (P - \overline{P}) = r\frac{\sigma_p}{\sigma_s}(S - \overline{S})$$

Given $\overline{P} = 100$, $\overline{S} = 103$, $\sigma_p = 8$, $\sigma_s = 4$, $r = 0.4$. Substituting these values in the above equation, we have

$$P - 100 = 0.4\frac{8}{4} \text{ or } P = 17.6 + 0.8\,S$$

This equation cannot be used to read off values of S for various values of P. Thus to read off values of S for various values of P, we use another regression equation of the form:

$$S = c + dP \text{ or } S - \overline{S} = \frac{\sigma_s}{\sigma_p}(P - \overline{P})$$

Substituting given values in this equation, we have

$$S - 103 = 0.4\,\frac{4}{8}\,(P - 100) \text{ or } S = 83 + 0.2P$$

**Example 14.12:** For certain $x$ and $y$ series which are correlated, the two lines of regression are:

$$5x - 6y + 90 = 0 \text{ and } 15x - 8y - 130 = 0.$$

Find the means of the two series and the correlation. *[MD Univ., M.Com., 2001]*

**Solution:** Solving two simultaneous regression equations to find mean value, we get $\overline{x} = 30$ and $\overline{y} = 40$.

Rewriting first regression equation as follows to find correlation coefficient, $r$:

$$6y = 5x + 90, \text{ i.e., } y = x\frac{5}{6} + 15 \text{ or } b_{yx} = \frac{5}{6}.$$

Also, $$15x = 8y + 130, \text{ i.e., } x = \frac{8}{15}y + \frac{130}{15} \text{ or } b_{xy} = \frac{8}{15}.$$

Since both regression coefficients $b_{xy}$ and $b_{yx}$ are less than one, applying following formula to get correlation coefficient:

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = 0.667.$$

**Example 14.13:** From 10 observations of price ($x$) and supply ($y$) of a commodity, the following summary figures were obtained (in appropriate units):

$$\Sigma x = 130;\ \Sigma y = 220;\ \Sigma x^2 = 2288;\ \Sigma y^2 = 5506;\ \text{and } \Sigma xy = 3467$$

Compute a regression line of $y$ on $x$ and estimate the supply when the price is 16.

**Solution:** Given that $\overline{y} = \frac{1}{n}\Sigma y = \frac{220}{10} = 22$ and $\overline{x} = \frac{1}{n}\Sigma x = \frac{130}{10} = 13.$

Regression line of $y$ on $x$ is given by

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x}), \text{ where } r\frac{\sigma_y}{\sigma_x} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{607}{598} = 1.015.$$

$$\begin{aligned} y - 22 &= 1.015\,(x - 13) \\ &= 1.015\,x - 13.195 \\ n &= 1.015\,x + 8.805. \end{aligned}$$

When price $x = 16$, the corresponding supply $y$ becomes

$$y = 1.015\,(16) + 8.805 = 25.045.$$

Thus, the estimated supply is of 25.45 units when price is 16 units.

**Example 14.14:** For a given set of bivariate data, the following results were obtained:

$$\bar{x} = 53.2,\ \bar{y} = 27.9, \text{ Regression coefficient of } y \text{ on } x = -1.5$$

Regression coefficient of $x$ on $y = -0.2$. Find the most probable value of $y$ when $x$ is 60.

[*Mysore Univ., B.Com., 2002*]

**Solution:** For finding the most probable value of $y$ when $x = 60$, a regression equation of $y$ on $x$ is written as:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}),$$

$$y - 27.9 = -1.5(x - 53.2)$$

$$= -1.5\,x + 79.8 \quad \text{or} \quad y = 107.7 - 1.5\,x.$$

For $x = 60$, $y = 107.7 - 1.5\,(60) = 17.7$ and $r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{-0.2 \times -1.5} = -0.548$.

**Example 14.15:** The two regression lines obtained in a correlation analysis of 60 observations are

$$5x = 6x + 24 \text{ and } 1000y = 768x - 3708$$

What is the correlation coefficient and what is its probable error? Show that the ratio of the coefficient of variability of $x$ to that of $y$ is 5/24. What is the ratio of variances of $x$ and $y$?

**Solution:** Rewriting the regression equations

$$5x = 6y + 24 \text{ or } x = \frac{6}{5}y + \frac{24}{5}$$

That is, $b_{xy} = 6/5$. Also

$$1000y = 768x - 3708 \text{ or } y = \frac{768}{1000}x - \frac{3708}{1000}$$

That is, $b_{yx} = 768/1000$. Since

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ and } b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{768}{1000},$$

therefore $\quad b_{xy}b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} = 0.9216 \quad \text{or} \quad r = \sqrt{0.9216} = 0.96.$

$$\text{Probable error of } r = 0.6745\frac{1 - r^2}{\sqrt{n}} = 0.6745\ \frac{1 - (0.96)^2}{\sqrt{60}}$$

$$= \frac{0.0528}{7.7459} = 0.0068$$

Solving the given regression equations for $x$ and $y$, we get $\bar{x} = 6$ and $\bar{y} = 1$. Also

$$r\,\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } 0.96\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } \frac{\sigma_x}{\sigma_y} = \frac{6}{5 \times 0.96} = \frac{5}{4}$$

$$\text{Ratio of coefficient of variability} = \frac{\sigma_x/\bar{x}}{\sigma_y/\bar{y}} = \frac{\bar{y}}{\bar{x}}\cdot\frac{\sigma_x}{\sigma_y} = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}.$$

### 14.7.3 Regression Coefficients for Grouped Sample Data

If data set is grouped or classified into frequency distribution of either variable $x$ or $y$ or both, then values of regression coefficients $b_{xy}$ and $b_{yx}$ are calculated by using the formulae:

$$b_{xy} = \frac{n\Sigma\, d_x d_y - \Sigma\, fd_x\, \Sigma\, fd_y}{n\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2} \times \frac{h}{k}$$

$$b_{yx} = \frac{n\Sigma\, fd_x d_y - \Sigma\, fd_x\, \Sigma\, fd_y}{n\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2} \times \frac{k}{h}$$

where $h$ is width of the class interval of sample data on $x$ variable and $k$ is the width of the class interval of sample data on $y$ variable.

**Example 14.16:** The following bivariate frequency distribution relates to sales turnover (₹ in lakh) and money spent on advertising (₹ in 1000's). Obtain the two regression equations

| *Sales Turnover (₹ in lakh)* | *Advertising Budget (₹ in 1000's)* | | | |
|---|---|---|---|---|
| | *50–60* | *60–70* | *70–80* | *80–90* |
| 20– 50 | 2 | 1 | 2 | 5 |
| 50– 80 | 3 | 4 | 7 | 6 |
| 80–110 | 1 | 5 | 8 | 6 |
| 110–140 | 2 | 7 | 9 | 2 |

Estimate (a) the sales turnover corresponding to advertising budget of ₹1,50,000, and (b) the advertising budget to achieve a sales turnover of ₹200 lakh.

**Solution:** Let $x$ and $y$ represent sales turnover and advertising budget, respectively. Then the regression equation for estimating the sales turnover ($x$) on advertising budget ($y$) is expressed as

$$x - \bar{x} = b_{xy}\,(y - \bar{y})$$

where $b_{xy} = \dfrac{n\Sigma\, fd_x d_y - \Sigma\, fd_x\, \Sigma\, fd_y}{n\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2}$

Similarly, the regression equation for estimating the advertising budget ($y$) on sales turnover of ₹200 lakh is written as

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where $b_{yx} = \dfrac{n\Sigma\, fd_x d_y - (\Sigma\, fd_x)\,(\Sigma\, fd_y)}{n\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2}$

The calculations for regression coefficients $b_{xy}$ and $b_{yx}$ are shown in Table 14.6.

$$\bar{x} = \text{A} + \frac{\Sigma\, fd_x}{n} \times h = 65 + \frac{50}{70} \times 30 = 65 + 21.428 = 86.428$$

$$\bar{y} = \text{B} + \frac{\Sigma\, fd_y}{n} \times k = 75 - \frac{14}{70} \times 10 = 75 - 2 = 73$$

**Table 14.6:** Calculations for Regression Coefficients

| | | | *Advertising Budget* | | | | $f$ | $fd_x$ | $fd_x^2$ | $fd_x\,d_y$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $y$ | *50–60* | *60–70* | *70–80* | *80–90* | | | | |
| | | $m.v$ | *55* | *65* | *75* | *85* | | | | |
| *Sales* | | $d_y$ | *−2* | *−1* | *0* | *1* | | | | |
| $x$ | *m.v.* | $d_x$ | | | | | | | | |
| 20–50 | 35 | 1 | 2<br>$\textcircled{4}$ | 1<br>$\textcircled{1}$ | 2<br>— | 5<br>$\textcircled{-5}$ | 10 | −10 | 10 | 0 |
| 50–80 | 65 | 0 | 3<br>— | 4<br>— | 7<br>— | 6<br>— | 20 | 0 | 0 | 0 |
| 80–110 | 95 | 1 | 1<br>$\textcircled{-2}$ | 5<br>$\textcircled{-5}$ | 8<br>— | 6<br>$\textcircled{6}$ | 20 | 20 | 20 | −1 |
| 110–140 | 125 | 2 | 2<br>$\textcircled{-8}$ | 7<br>$\textcircled{-14}$ | 9<br>— | 2<br>$\textcircled{4}$ | 20 | 40 | 80 | −18 |
| | | $f$ | 8 | 17 | 26 | 19 | $n$ = 70 | 50 = $\Sigma fd_x$ | 110 = $\Sigma fd_x^2$ | −19 $\Sigma fd_x d_y$ |
| | | $fd_y$ | −16 | −17 | 0 | 19 | −14 = $\Sigma fd_y$ | | | |
| | | $fd_y^2$ | 32 | 17 | 0 | 19 | 68 = $\Sigma fd_y^2$ | | | |
| | | $fd_xd_y$ | −6 | −18 | 0 | 5 | −19 = $\Sigma fd_xd_y$ | | | |

$$b_{xy} = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_y^2 - (\Sigma fd_y)^2} \times \frac{h}{k} = \frac{70 \times -19 - (50)(-14)}{70 \times 68 - (-14)^2} \times \frac{30}{10}$$

$$= \frac{-1330 + 700}{4760 - 196} \times \frac{30}{10} = \frac{-18{,}900}{45{,}640} = -\,0.414$$

$$b_{yx} = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{k}{h} = \frac{70 \times -19 - (50)(-14)}{70 \times 110 - (50)^2} \times \frac{10}{30}$$

$$= \frac{-1330 + 700}{7700 - 2500} \times \frac{10}{30} = \frac{-6300}{1{,}56{,}000} = -\,0.040$$

Substituting these values in the two regression equations, we get

(a) Regression equation of sales turnover ($x$) to advertising budget ($y$) is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 86.428 = -\,0.414\,(y - 73), \text{ or } x = 116.65 - 0.414y$$

For $y = 150$, we have $x = 116.65 - 0.414 \times 150 =$ ₹54.55 lakh

(b) Regression equation of advertising budget ($y$) on sales turnover ($x$) is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 73 = -\,0.040\,(x - 86.428) \text{ or } y = 76.457 - 0.04x$$

For $x = 200$, we have $y = 76.457 - 0.04\,(200) =$ ₹68.457 thousand.

## Self-practice Problems 14A

**14.1** The following calculations have been made for prices of twelve stocks ($x$) at the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares ($y$). From these calculations find the regression equation of price of stocks on the volume of sales of shares.

$$\Sigma x = 580, \quad \Sigma y = 370, \quad \Sigma xy = 11494,$$
$$\Sigma x^2 = 41658, \; \Sigma y^2 = 17206.$$

*[Rajasthan Univ., M.Com., 2005]*

**14.2** A survey was conducted to study the relationship between expenditure (in ₹) on accommodation ($x$) and expenditure on food and entertainment ($y$) and the following results were obtained:

| | *Mean* | *Standard Deviation* |
|---|---|---|
| • Expenditure on accommodation | 173 | 63.15 |
| • Expenditure on food and entertainment | 47.8 | 22.98 |

Coefficient of correlation $r = 0.57$

Write down the regression equation and estimate the expenditure on food and entertainment if the expenditure on accommodation is ₹200.

*[Bangalore Univ., B.Com.,2008]*

**14.3** The following data give the experience of machine operators and their performance ratings given by the number of good parts turned out per 100 pieces:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Operator : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| experience ($x$) : | 16 | 12 | 18 | 4 | 3 | 10 | 5 | 12 |
| Performance ratings ($y$) : | 87 | 88 | 89 | 68 | 78 | 80 | 75 | 83 |

Calculate the regression lines of performance ratings on experience and estimate the probable performance if an operator has 7 years experience.

*[Jammu Univ., M.Com.; Lucknow Univ., MBA, 2006]*

**14.4** A study of prices of a certain commodity at Delhi and Mumbai yield the following data:

| | *Delhi* | *Mumbai* |
|---|---|---|
| • Average price per kilo (Rs) | 2.463 | 2.797 |
| • Standard deviation | 0.326 | 0.207 |
| • Correlation coefficient between prices at Delhi and Mumbai $r = 0.774$ | | |

Estimate from the above data the most likely price (a) at Delhi corresponding to the price of ₹2.334 per kilo at Mumbai (b) at Mumbai corresponding to the price of 3.052 per kilo at Delhi.

**14.5** The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Aptitude scores ($x$) : | 60 | 62 | 65 | 70 | 72 | 48 | 53 | 73 | 65 | 82 |
| Productivity index ($y$) : | 68 | 60 | 62 | 80 | 85 | 40 | 52 | 62 | 60 | 81 |

Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92, (b) the test score of a worker whose productivity index is 75.

*[Delhi Univ., MBA, 2005]*

**14.6** A company wants to assess the impact of R&D expenditure (₹ in 1000s) on its annual profit; (₹ in 1000's). The following table presents the information for the last eight years:

| *Year* | *R & D expenditure* | *Annual profit* |
|---|---|---|
| 1991 | 9 | 45 |
| 1992 | 7 | 42 |
| 1993 | 5 | 41 |
| 1994 | 10 | 60 |
| 1995 | 4 | 30 |
| 1996 | 5 | 34 |
| 1997 | 3 | 25 |
| 1998 | 2 | 20 |

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of ₹1,00,000 as R&D expenditure.

*[Jodhpur Univ., MBA, 2008]*

**14.7** Obtain the two regression equations from the following bivariate frequency distribution:

| *Sales Revenue (₹ in lakh)* | *Advertising Expenditure (₹ in thousand)* | | | |
|---|---|---|---|---|
| | *5–15* | *15–25* | *25–35* | *35-45* |
| 75–125 | 3 | 4 | 4 | 8 |
| 125–175 | 8 | 6 | 5 | 7 |
| 175–225 | 2 | 2 | 3 | 4 |
| 225–275 | 3 | 3 | 2 | 2 |

Estimate (a) the sales corresponding to advertising expenditure of ₹50,000, (b) the advertising expenditure for a sales revenue of ₹300 lakh, and (c) the coefficient of correlation.*[Delhi Univ., MBA, 2007]*

**14.8** The personnel manager of an electronic manufacturing company devises a manual test for job applicants to predict their production rating in the assembly department. In order to do this he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. The results are as follows:

| Worker : | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Test score : | 53 | 36 | 88 | 84 | 86 | 64 | 45 | 48 | 39 | 69 |
| Production rating : | 45 | 43 | 89 | 79 | 84 | 66 | 49 | 48 | 43 | 76 |

Fit a linear least squares regression equation of production rating on test score.

[*Delhi Univ., MBA, 2008*]

**14.9** Find the regression equation showing the capacity utilization on production from the following data:

| | | *Average Deviation* | *Standard Deviation* |
|---|---|---|---|
| Production (in lakh units) | : | 35.6 | 10.5 |
| Capacity utilization (in percentage) | : | 84.8 | 8.5 |
| Correlation coefficient $r = 0.62$ | | | |

Estimate the production when the capacity utilization is 70 per cent.

[*Delhi Univ., MBA, 2007; Pune Univ., MBA, 2008*]

**14.10** Suppose that you are interested in using past expenditure on R&D by a firm to predict current expenditures on R&D. You got the following data by taking a random sample of firms, where $x$ is the amount spent on R&D (in lakh of rupees) 5 years ago and $y$ is the amount spent on R&D (in lakh of rupees) in the current year:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ : | 30 | 50 | 20 | 80 | 10 | 20 | 20 | 40 |
| $y$ : | 50 | 80 | 30 | 110 | 20 | 20 | 40 | 50 |

(a) Find the regression equation of $y$ on $x$.

(b) If a firm is chosen randomly and $x = 10$, can you use the regression to predict the value of $y$? Discuss. [*Madurai-Kamraj Univ., MBA, 2005*]

**14.11** The following data relates to the scores obtained by a salesman of a company in an intelligence test and their weekly sales (in ₹1000's):

| Salesman intelligence : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Test score : | 50 | 60 | 50 | 60 | 80 | 50 | 80 | 40 | 70 |
| Weekly sales : | 30 | 60 | 40 | 50 | 60 | 30 | 70 | 50 | 60 |

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

[*HP Univ., M.com.,2006*]

**14.12** Two random variables have the regression equations:

$$3x + 2y - 26 = 0 \text{ and } 6x + y - 31 = 0$$

(a) Find the mean values of $x$ and $y$ and coefficient of correlation between $x$ and $y$.

(b) If the variance of $x$ is 25, then find the standard deviation of $y$ from the data.

[*MD Univ., M.Com., 2007; Kumaun Univ., MBA, 2005*]

**14.13** For a given set of bivariate data, the following results were obtained

$$\bar{x} = 53.2,\ \bar{y} = 27.9,$$

Regression coefficient of $y$ on $x = -1.5$, and Regression coefficient of $x$ and $y = -0.2$.

Find the most probable value of $y$ when $x = 60$.

**14.14** In trying to evaluate the effectiveness in its advertising campaign, a firm compiled the following information:

Calculate the regression equation of sales on advertising expenditure. Estimate the probable sales when advertisement expenditure is ₹ 60 thousand.

| *Year* | *Adv. expenditure (₹ 1000's)* | *Sales (in lakhs ₹)* |
|---|---|---|
| 2003 | 12 | 5.0 |
| 2004 | 15 | 5.6 |
| 2005 | 17 | 5.8 |
| 2006 | 23 | 7.0 |
| 2007 | 24 | 7.2 |
| 2008 | 38 | 8.8 |
| 2009 | 42 | 9.2 |
| 2010 | 48 | 9.5 |

[*Bharathidasan Univ., MBA, 2003*]

## Hints and Answers

**14.1** $\bar{x} = \Sigma x/n = 580/12 = 48.33$;

$\bar{y} = \Sigma y/n = 370/12 = 30.83$

$$b_{xy} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma y^2 - n(\bar{y})^2} = \frac{11494 - 12 \times 48.33 \times 30.83}{17206 - 12(30.83)^2}$$

$= -1.102$

Regression equation of $x$ on $y$:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 48.33 = -1.102\,(y - 30.83)$$

or $\quad x = 82.304 - 1.102y$

**14.2** Given $\bar{x} = 172$, $\bar{y} = 47.8$, $\sigma_x = 63.15$, $\sigma_y = 22.98$, and $r = 0.57$

Regression equation of food and entertainment ($y$) on accomodation ($x$) is given by

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 47.8 = 0.57\,\frac{22.98}{63.15}(x - 173)$$

or $\quad y = 11.917 + 0.207x$

For $x = 200$, we have $y = 11.917 + 0.207(200)$

$= 53.317$

**14.3** Let the experience and performance rating be represented by $x$ and $y$ respectively.

$\bar{x} = \Sigma x/n = 80/8 = 10;\ \bar{y} = \Sigma y/n = 648/8 = 81$

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{247}{218} = 1.133;$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$

Regression equation of $y$ on $x$

$$y - \bar{y} = byx\,(x - \bar{x})$$

or $\quad y - 81 = 1.133\,(x - 10)$

or $\quad y = 69.67 + 1.133x$

When $\quad x = 7,\ y = 69.67 + 1.133\,(7) = 77.60 \cong 78$

**14.4** Let price at Mumbai and Delhi be represented by $x$ and $y$, respectively

(a) Regression equation of $y$ on $x$

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 2.463 = 0.774\ \frac{0.326}{0.207}(x - 2.797)$$

For $x$ = ₹2.334, the price at Delhi would be $y$ = ₹1.899.

(b) Regression on equation of $x$ on $y$

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

or $\quad x - 2.791 = 0.774\dfrac{0.207}{0.326}(y - 2.463)$

For $y$ = ₹3.052, the price at Mumbai would be $x$ = ₹3.086.

**14.5** Let aptitude score and productivity index be represented by $x$ and $y$ respectively.

$$\bar{x} = \Sigma x/n = 650/10 = 65;\ \bar{y} = \Sigma y/n = 650/10 = 65$$

$$b_{xy} = \frac{n\,\Sigma d_x\,d_y - (\Sigma d_x)\,(\Sigma d_y)}{n\,\Sigma d_y^2 - \Sigma(d_y)^2} = \frac{1044}{1752} = 0.596;$$

where $d_x = x - \bar{x}$; $d_y = y - \bar{y}$

(a) Regression equation of $x$ on $y$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

or $\quad x - 65 = 0.596\,(y - 65)$

or $\quad x = 26.26 + 0.596y$

When $\quad y = 75,\ x = 26.26 + 0.596(75) = 70.96 \cong 71$

(b) $b_{yx} = \dfrac{n\,\Sigma d_x d_y - (\Sigma d_x)\,(\Sigma d_y)}{n\,\Sigma d_x^2 - (\Sigma d_x)^2} = \dfrac{1044}{894} = 1.168$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

or $\quad y - 65 = 1.168(x - 65)$

or $\quad y = -10.92 + 1.168x$

When $x = 92,\ y = -10.92 + 1.168(92) = 96.536 \cong 97$

**14.6** Let R&D expenditure and annual profit be denoted by $x$ and $y$ respectively

$$\bar{x} = \Sigma x/n = 40/8 = 5.625;\ \bar{y} = \Sigma y/n = 297/8 = 37.125$$

$$b_{yx} = \frac{n\,\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\,\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 238 - (-3)\,(1)}{8 \times 57 - (-3)^2} = 4.266\ ;$$

where $d_x = x - 6,\ d_y = y - 37$

Regression equation of annual profit on R&D expenditure

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 37.125 = 4.26\,(x - 5.625)$$

or $\quad y = 13.163 + 4.266x$

For $x$ = ₹1,00,000 as R&D expenditure, we have from above equation $y$ = ₹439.763 as annual profit.

**14.7** Let sales revenue and advertising expenditure be denoted by $x$ and $y$ respectively

$$\bar{x} = \text{A} + \frac{\Sigma\, fd_x}{n} \times h = 150 + \frac{12}{66} \times 50 = 159.09$$

$$\bar{y} = \text{B} + \frac{\Sigma\, fd_y}{n} \times k = 30 - \frac{26}{66} \times 10 = 26.06$$

$$b_{xy} = \frac{n\,\Sigma\, fd_x d_y - (\Sigma\, fd_x)\,(\Sigma\, fd_y)}{n\,\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2} \times \frac{h}{k}$$

$$= \frac{66\,(-14) - 12(-26)}{66(100) - (-26)^2} \times \frac{50}{10} = -0.516$$

(a) Regression equation of $x$ on $y$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 159.09 = -0.516(y - 26.06)$$

or $\quad x = 172.536 - 0.516y$

For $y = 50,\quad x = 147.036$

(b) Regression equation of $y$ on $x$

$$b_{yx} = \frac{n\,\Sigma\, fd_x d_y - (\Sigma\, fd_x)\,(\Sigma\, fd_y)}{n\,\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2} \times \frac{k}{h}$$

$$= \frac{66\,(-14) - 12(-26)}{66\,(70) - (12)^2} \times \frac{10}{50} = -\,0.027.$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 26.06 = -0.027(x - 159.09)$$

$$y = 30.355 - 0.027x$$

For $x = 300,\ y = 22.255$

(c) $r = \sqrt{b_{xy} \times b_{yx}} = -\,\sqrt{0.516 \times 0.027} = -0.1180$

**14.8** Let test score and production rating be denoted by $x$ and $y$ respectively.

$$\bar{x} = \Sigma x/n = 612/10 = 61.2;$$

$$\bar{y} = \Sigma y/n = 622/10 = 62.2$$

$$b_{yx} = \frac{n\,\Sigma\, d_x d_y - (\Sigma\, d_x)\,(\Sigma d_y)}{n\,\Sigma\, d_x^2 - (\Sigma\, d_x)^2} = \frac{10 \times 3213 - 2 \times 2}{10 \times 3554 - (2)^2}$$

$$= 0.904$$

Regression equation of production rating ($y$) on test score ($x$) is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 62.2 = 0.904(x - 61.2)$$

$$y = 6.876 + 0.904x$$

**14.9** Let production and capacity utilization be denoted by $x$ and $y$, respectively.

(a) Regression equation of capacity utilization ($y$) on production ($x$)

$$y - \bar{y} = r\ \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 84.8 = 0.62\ \frac{8.5}{10.5}(x - 35.6)$$

$$y = 66.9324 + 0.5019x$$

(b) Regression equation of production ($x$) on capacity utilization ($y$)

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$x - 35.6 = 0.62\ \frac{10.5}{8.5}(y - 84.8)$$

$$x = -29.3483 + 0.7659y$$

When $y = 70, x = -29.3483 + 0.7659(70) = 24.2647$

Hence the estimated production is 2,42,647 units when the capacity utilization is 70 per cent.

**14.10** $\bar{x} = \Sigma x/n = 270/8 = 33.75$; $\bar{y} = \Sigma y/n = 400/8 = 50$

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8 \times 4800 - 6 \times 0}{8 \times 3592 - (6)^2}$$

$$= 1.338;$$

where $d_x = x - 33$ and $d_y = y - 50$

Regression equation of $y$ on $x$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50 = 1.338(x - 33.75)$$

$$y = 4.84 + 1.338x$$

For $x = 10, y = 18.22$

**14.11** Let intelligence test score be denoted by $x$ and weekly sales by $y$

$$\bar{x} = 540/9 = 60;$$

$$\bar{y} = 450/9 = 50,$$

$$b_{yx} = \frac{n\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9 \times 1200}{9 \times 1600} = 0.75$$

Regression equation of $y$ on $x$:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50 = 0.75\,(x - 60)$$

$$y = 5 + 0.75x$$

For $x = 65, y = 5 + 0.75\,(65) = 53.75$

**14.12** (a) Solving two regression lines:

$$3x + 2y = 6 \text{ and } 6x + y = 31$$

we get mean values as = 4 and = 7

(b) Re-writing regression lines as follows:

$$3x + 2y = 26 \text{ or } y = 13 - (3/2)x,$$

So $\quad b_{yx} = -3/2$

$$6x + y = 31 \text{ or } x = 31/6 - (1/6)y,$$

So $\quad b_{xy} = -1/6$

Correlation coefficient,

$$r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{(3/2)(1/6)} = -0.5$$

Given, Var($x$) = 25, so $\sigma_x = 5$. Calculate $\sigma_y$ using the formula:

$$b_{yx} = r\frac{\sigma_y}{\sigma_y}$$

or $\quad -\dfrac{3}{2} = 0.5\dfrac{\sigma_y}{5}$ or $\sigma_y = 15$

**14.13** The regression equation of $y$ on $x$ is stated as:

$$y - \bar{y} = b_{xy}(x - \bar{x}) = r \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Given, $\bar{x} = 53.20$; $\bar{y} = 27.90$, $b_{yx} = -1.5$; $b_{xy} = -0.2$

Thus $y - 27.90 = -1.5(x - 53.20)$

or $\quad y = 107.70 - 1.5x$

For $x = 60$, we have $y = 107.70 - 1.5(60) = 17.7$

Also $\quad r = \sqrt{b_{yx} \times b_{xy}} = -\sqrt{1.5 \times 0.2} = -0.5477$

**14.14** Let advertising expenditure and sales be denoted by $x$ and $y$ respectively.

$$\bar{x} = \Sigma x/n = 217/8 = 27.125;$$

$$\bar{y} = \Sigma y/n = 58.2/8 = 7.26$$

$$b_{yx} = \frac{n\sum dx\,dy - (\sum dx)(\sum dy)}{n\sum d_x^2 - (\sum dx)^2}$$

$$= \frac{8(172.2) - (25)(2.1)}{8(1403) - (25)^2} = \frac{1325.1}{10599} = 0.125$$

Thus regression equation of $y$ on $x$ is:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

or $\quad y - 7.26 = 0.125(x - 27.125)$

$$y = 3.86 + 0.125x$$

When $x = 60$, the estimated value of $y = 3.869 + 0.125(60) = 11.369$

## 14.8 STANDARD ERROR OF ESTIMATE AND PREDICTION INTERVALS

The distribution of expected values of dependent variable, $y$, about a least squares regression line for given values of independent variable $x$ indicates the strength (or extent) and direction of relationship between these two variables. For example, wide pattern of dot points indicates a poor relationship while a very close pattern of dot points indicates a close relationship between two variables. The variability among observed values of dependent variable, $y$, about the regression line is measured in terms of *residuals*. A residual is defined as the difference between an observed value of dependent variable $y$ and its estimated (or fitted) value for a given value of the independent variable $x$. The residual about the regression line is given by

$$\text{Residual, } e_i = y_i - \hat{y}_i$$

The residual values $e_i$ are plotted on a diagram with respect to the least squares regression line $\hat{y} = a + bx$. These residual values are the vertical distances of every observation (dot point) from the least squares line as shown in Fig. 14.3 and represent error of estimation for individual values of dependent variable.

Since sum of the residuals is zero, therefore it is not possible to determine the total amount of error by summing the residuals. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftarrow \text{Error or residual sum of squares}$$

This quantity is also called the *sum of squares of errors (SSE)*.

The *variance of error of estimate* $\sigma_e^2$ or $S_{y.x}^2$ is determined as follows:

$$S_{yx}^2 \text{ or } \hat{\sigma}_e^2 = \frac{\Sigma e_i^2}{n-2} = \frac{\Sigma (y_i - y_i)^2}{n-2} = \frac{SSE}{n-2}$$

The denominator $n - 2$ represents the *residual degrees of freedom* and is obtained by subtracting from sample size, $n$ is the number of parameters $\beta_0$ and $\beta_1$ that are estimated by the sample parameters $a$ and $b$ in the least squares equation.

The *standard error of estimate (or standard deviation of error term),* $S_{yx}$ measures the variability of the observed values around the regression line. The standard deviation of error term, $S_{yx}$, about the least squares line is defined as

$$S_{yx} \text{ or } \sigma_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} \text{ or } \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\,\Sigma xy}{n-2}} = \sqrt{\frac{\text{SSE}}{n-2}} \qquad (14\text{-}4)$$

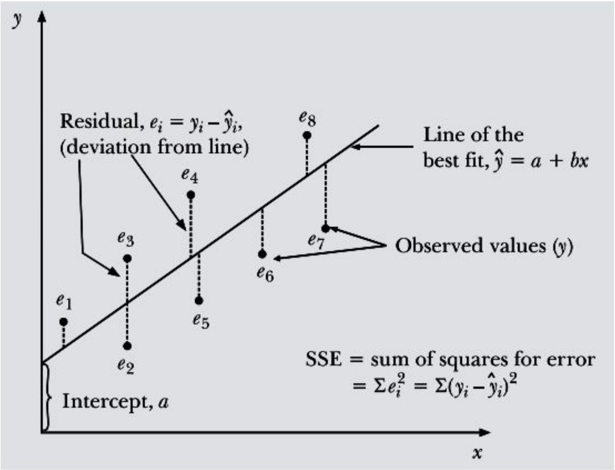

$y$
Residual, $e_i = y_i - \hat{y}_i$,
(deviation from line)
$e_1$
$e_2$
$e_3$
$e_4$
$e_5$
$e_6$
$e_7$
$e_8$
Line of the
best fit, $\hat{y} = a + bx$
Observed values ($y$)
SSE = sum of squares for error
$= \Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2$
Intercept, $a$
$x$

**Figure 14.3**
Residuals

The variance $S_{yx}^2$ measures how the least squares line 'best fits' the sample $y$-values. A large *variance and standard error of estimate* indicate a large amount of dispersion of sample $y$-values (dot points) around the regression line. Smaller the value of $S_{yx}$, closer the $y$-values (dot points) fall around the regression line and better the line fits the data and describes the better average relationship between the two variables. If all dot points fall on the line, then value of $S_{yx}$ is zero, and the relationship between the two variables is said to be perfect.

Smaller the value of $S_{yx}$, is considered useful in predicting the value of a dependent variable $y$. Dispersion of sample $y$-values (dot points) around the regression line needs to be measured because

(i) it facilitate in predicting the value of the dependent variable.
(ii) value of $Syx$ is used for interval estimates of the dependent variable so as to draw statistical inferences results.

The distribution of expected values of dependent variable $y$ about a least squares regression line for given values of independent variable $x$ is shown in Fig. 14.4. Suppose the amount of deviation in the values of $y$ given any particular value of $x$ follow normal distribution. Since average value of $y$ changes with the value of $x$, we have different normal distributions of $y$-values for every value of $x$, each having same standard deviation. When a relationship between two variables $x$ and $y$ exists, the standard deviation (also called *standard error of estimate*) is less than the standard deviation of all the $x$-values in the population computed about their mean.

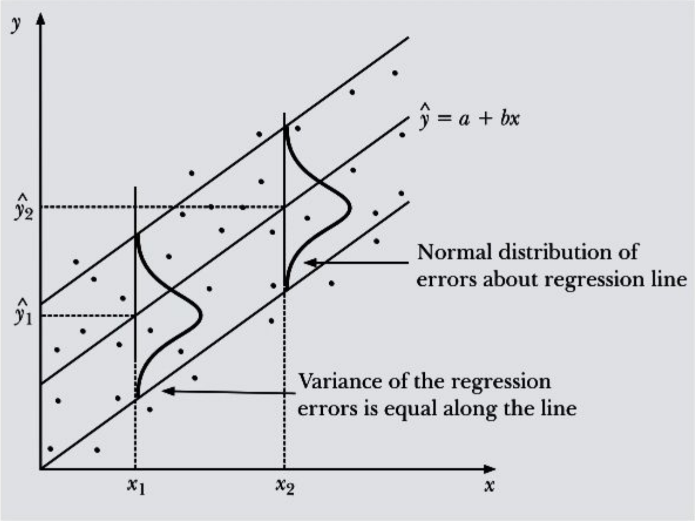

y
$\hat{y} = a + bx$
$\hat{y}_2$
$\hat{y}_1$
Normal distribution of errors about regression line
Variance of the regression errors is equal along the line
$x_1$
$x_2$
x

**Figure 14.4**
Regression Line Showing the Error Variance

The standard error of estimate can also be used to determine an interval estimate (also called *prediction interval*) based on sample data ($n < 30$) for the value of the dependent variable, $y$, for a given value of the independent variable, $x$, as follows:

Approximate interval estimate = $\hat{y} \pm t_{df} S_{yx}$

where value of *t-statistics* is obtained from $t$-distribution table at given level of significance and degree of freedom.

**Example 14.17:** The following data relate to advertising expenditure (₹ in lakh) and their corresponding sales (₹ in crore)

| | | | | | | |
|---|---|---|---|---|---|---|
| Advertising expenditure | : | 10 | 12 | 15 | 23 | 20 |
| Sales | : | 14 | 17 | 23 | 25 | 21 |

(a) Find the equation of the least-squares line fitting the data.
(b) Estimate the value of sales corresponding to advertising expenditure of ₹30 lakh.
(c) Calculate the standard error of estimate of sales on advertising expenditure.

**Solution:** Let the advertising expenditure be denoted by $x$ and sales by $y$. The calculations for the least squares line are shown in Table 14.7

**Table 14.7:** Calculations for Least-squares Line

| *Advt. Expenditure, x* | $d_x = x - 16$ | $d_x^2$ | *Sales y* | $d_y = y - 20$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 10 | −6 | 36 | 14 | −6 | 36 | 36 |
| 12 | −4 | 16 | 17 | −3 | 9 | 12 |
| 15 | −1 | 1 | 23 | 3 | 9 | −3 |
| 23 | 7 | 49 | 25 | 5 | 25 | 35 |
| 20 | 4 | 16 | 21 | 1 | 1 | 4 |
| 80 | 0 | 118 | 100 | 0 | 80 | 84 |

$$\bar{x} = \Sigma x/n = 80/5 = 16; \quad \bar{y} = \Sigma y/n = 100/5 = 20$$

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{5 \times 84}{5 \times 118} = 0.712$$

(a) Regression equation of $y$ on $x$ is

$$
\begin{aligned}
y - \bar{y} &= b_{yx}(x - \bar{x}) \\
y - 20 &= 0.712\,(x - 16) \\
y &= 8.608 + 0.712\,x
\end{aligned}
$$

where $a = 8.608$ and $b = 0.712$.

The fitted values and the residuals for the sample data in Table 14.7 are shown in Table 14.8. The fitted values are obtained by substituting values of $x$ in the least squares line (regression equation). For example, 8.608 + 0.712(10) = 15.728. The residuals that indicate how well the least squares line fits the actual data are equal to the actual value minus fitted value.

**Table 14.8:** Fitted Values and Residuals for Sample Data

| *Value, x* | *Fitted Value*<br>$y = 8.608 + 0.712x$ | *Residuals* |
|---|---|---|
| 10 | 15.728 | –5.728 |
| 12 | 17.152 | –5.152 |
| 15 | 19.288 | –4.288 |
| 23 | 24.984 | –1.984 |
| 20 | 22.848 | –2.848 |

(b) The least squares line (equation) obtained in part (a) may be used to estimate the sales turnover corresponding to the advertising expenditure of ₹ 30 lakh as:

$$\hat{y} = 8.608 + 0.712x = 8.608 + 0.712\,(30) = ₹29.968 \text{ crore}$$

(c) Calculations for standard error of estimate, $S_{yx}$ of sales ($y$) on advertising expenditure ($x$) are shown in Table 14.9.

**Table 14.9:** Calculations for Standard Error of Estimate

| $x$ | $y$ | $y^2$ | $xy$ |
|---|---|---|---|
| 10 | 14 | 196 | 140 |
| 12 | 17 | 289 | 204 |
| 15 | 23 | 529 | 345 |
| 23 | 25 | 625 | 575 |
| 20 | 21 | 441 | 420 |
| 80 | 100 | 2080 | 1684 |

$$
\begin{aligned}
S_{yx} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} &= \sqrt{\frac{2080 - 8.608 \times 100 - 0.712 \times 1684}{5-2}} \\
&= \sqrt{\frac{2080 - 860.8 - 1199}{3}} = 2.594
\end{aligned}
$$

### 14.8.1 Coefficient of Determination: Partitioning of Total Variation

It is desired that the residual variance should be as small as possible but its value depends on the unit in which values of dependent variable, $y$, are measured. Consequently, another measure of fit called *coefficient of determination* is needed that is independent of the unit in which values of dependent variable, $y$, are measured. The *coefficient of determination is the proportion of variability of the dependent variable y accounted for or explained by the independent variable x.* In other words, it measures how well (i.e. strength) the regression line fits the data. The coefficient of determination is denoted by $r^2$ and its value ranges from 0 to 1. A particular of value $r^2$ should be interpreted as high or low in accordance of the use and context with which the regression model is developed. The coefficient of determination is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$= 1 - \frac{\text{Residual variation in response variable } y\text{-values from least-squares line}}{\text{Total variance of } y\text{-values}}$$

where SST = total sum of square deviations (or total variance) of actual values of variable, $y$ from its mean value.

$$= \text{S}_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

SSE = sum of squares of error (*unexplained variation)* in the values of dependent variable, $y$ from the least squares line due to sampling errors (i.e. amount of residual variation in the data that is not explained by independent variable, $x$)

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

SSR = sum of squares of regression (*or explained variation*) is the actual values of dependent variable $y$ accounted for or explained by variation among values of independent variable, $x$

= SST – SSE

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = a \sum_{i=1}^{n} y_i + b \sum_{i=1}^{n} x_i y_i - n(\bar{y})^2$$

These three variations noted during regression analysis of a data set are shown in Fig 14.5. Thus,

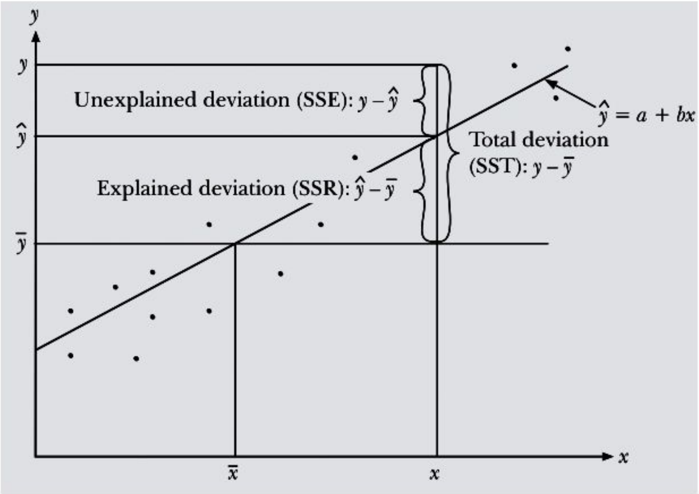
$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{\text{S}_{yx}^2}{\text{S}_y^2}\ ;\ \text{S}_{y \cdot x} = \text{S}_y\sqrt{1 - r^2}$$

where $\dfrac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$ = fraction of the total variation that is explained or accounted for

$\text{S}y \cdot x = \dfrac{\Sigma(y - \hat{y})^2}{n - 2}$, variance in the values of independent variable, $y$ from the least squares line

$\text{S}_y^2 = \dfrac{1}{n - 2}\Sigma(y - \bar{y})^2$ , total variance in the values of independent variable,

**Figure 14.5**
Relationship Between Three Types of Variations


y
y
$\hat{y}$
$\bar{y}$
Unexplained deviation (SSE): $y - \hat{y}$
Explained deviation (SSR): $\hat{y} - \bar{y}$
Total deviation (SST): $y - \bar{y}$
$\hat{y} = a + bx$
$\bar{x}$
x
x

An easy formula of coefficient of determination, $r^2$, is given by

$$r^2 = \frac{a\Sigma y + b\Sigma xy - n(\overline{y})^2}{\Sigma y^2 - n(\overline{y})^2} \quad \leftarrow \text{Short-cut method}$$

For example, the extent of relationship between sales revenue ($y$) and advertising expenditure ($x$) using data of Example 14.1 is computed as follows:

$$r^2 = \frac{a\,\Sigma y + b\Sigma xy - n(\overline{y})^2}{\Sigma y^2 - n(\overline{y})^2} = \frac{0.072 \times 40 + 0.704 \times 373 - 8(5)^2}{270 - 8(5)^2}$$

$$= \frac{2.88 + 262.592 - 200}{270 - 200} = \frac{65.47}{70} = 0.9352$$

The value $r^2 = 0.9352$ indicates that 93.52 per cent of the variance in sales revenue is on account of or statistically explained by advertising expenditure.

A comparison between bivariate correlation and regression analysis is summarized in Table 14-10.

**Table 14.10:** Comparison Between Linear Correlation and Regression

| | *Correlation* | *Regression* |
|---|---|---|
| • Measurement level | Interval or ratio scale | Interval or ratio scale |
| • Nature of variables | Both continuous, and linearly related | Both continuous, and linearly related |
| • $x - y$ relationship | $x$ and $y$ are symmetric | $y$ is dependent, $x$ is independent; regression of $x$ on $y$ differs from $y$ on $x$ |
| • Correlation | $b_{xy} = b_{yx}$ | Correlation between $x$ and $y$ is the same as the correlation between $y$ and $x$ |
| • Coefficient of determination | Explains common variance of $x$ and $y$ | Proportion of variability of $x$ explained by its least-squares regression on $y$ |

## Conceptual Questions 14A

1. (a) Explain the concept of regression and point out its usefulness in dealing with business problems. *[Delhi Univ., MBA, 2003]*

   Distinguish between correlation and regression. Also point out the properties of regression coefficients.

2. Explain the concept of regression and point out its importance in business forecasting. *[Delhi Univ., MBA, 2000, 2008]*

3. Under what conditions can there be one regression line? Explain. *[HP Univ., MBA, 2006]*

4. Why should a residual analysis always be done as part of the development of a regression model?

5. What are the assumptions of simple linear regression analysis and how can they be evaluated?

6. What is the meaning of the standard error of estimate?

7. What is the interpretation of $y$-intercept and the slope in a regression model?

8. What are regression lines? With the help of an example illustrate how they help in business decision-making. *[Delhi Univ., MBA, 2008]*

9. Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients? *[Osmania Univ., MBA; Delhi Univ., MBA, 2007]*

10. (a) Distinguish between correlation and regression analysis. *[Dipl in Mgt., AIMA, Osmania Univ., MBA, 2008]*

    (b) The coefficient of correlation and coefficient of determination are available as measures of association in correlation analysis. Describe the different uses of these two measures of association.

11. What are regression coefficients? State some of the important properties of regression coefficients. *[Dipl in Mgt., AIMA, Osmania Univ., MBA, 2001]*

12. What is regression? How is this concept useful to business forecasting? *[Jodhpur Univ., MBA, 2008]*

**13.** What is the difference between a prediction interval and a confidence interval in regression analysis?

**14.** Explain what is required to establish evidence of a cause-and-effect relationship between $y$ and $x$ with regression analysis.

**15.** What technique is used initially to identify the kind of regression model that may be appropriate?

**16.** (a) What are regression lines? Why is it necessary to consider two lines of regression?

(b) In case the two regression lines are identical, prove that the correlation coefficient is either + 1 or – 1. If two variables are independent, show that the two regression lines cut at right angles.

**17.** What are the purpose and meaning of the error terms in regression?

**18.** Give examples of business situations where you believe a straight line relationship exists between two variables. What would be the uses of a regression model in each of these situations?

**19.** 'The regression lines give only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in the estimate by calculating a quantity known as the standard error of estimate'. Elucidate.

**20.** Explain the advantages of the least-squares procedure for fitting lines to data. Explain how the procedure works.

## Formulae Used

1. Simple linear regression model

$$y = \beta_0 + \beta_1 x + e$$

2. Simple linear regression equation based on sample data

$$y = a + bx$$

3. Regression coefficient in sample regression equation

$$b = \hat{y}$$
$$a = \overline{y} - b\overline{x}$$

4. Residual representing the difference between an observed value of dependent variable $y$ and its fitted value

$$e = y - \hat{y}$$

5. Standard error of estimate based on sample data

- Deviations formula

$$S_{y.x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$$

- Computational formula

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\,\Sigma y - b\,\Sigma xy}{n - 2}}$$

6. Coefficient of determination based on sample data

- Sums of squares formula

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \overline{y})^2}$$

- Computational formula

$$r^2 = \frac{a\,\Sigma y + b\,\Sigma xy - n(\overline{y})^2}{\Sigma y^2 - n(\overline{y})^2}$$

7. Regression sum of squares

$$S_{y.x} = S_y \sqrt{1 - r^2}$$

8. Interval estimate based on sample data: $\hat{y} \pm t_{df}\, S_{yx}$

## Chapter Concepts Quiz

### True or False

**1.** [T] [F] A statistical relationship between two variables does not indicate a perfect relationship.

**2.** [T] [F] A dependent variable in a regression equation is a continuous random variable.

**3.** [T] [F] The residual value is required to estimate the amount of variation in the dependent variable with respect to the fitted regression line.

**4.** [T] [F] Standard error of estimate is the conditional standard deviation of the dependent variable.

**5.** [T] [F] Standard error of estimate is a measure of scatter of the observations about the regression line.

**6.** [T] [F] If one of the regression coefficients is greater than one the other must also be greater than one.

**7.** [T] [F] The signs of the regression coefficients are always same.

**8.** [T] [F] Correlation coefficient is the geometric mean of regression coefficients.

**9.** [T] [F] If the sign of two regression coefficients is negative, then sign of the correlation coefficient is positive.

**10.** [T] [F] Correlation coefficient and regression coefficient are independent.

# Chapter 14

*The cause is hidden, but the result is known.*

—Ovid

*I never think of the future, it comes soon enough.*

—Albert Einstein

# Regression Analysis

## LEARNING OBJECTIVES

After studying this chapter, you should be able to

- use simple linear regression for building models to business data.
- understand how the method of least squares is used to predict values of a dependent (or response) variable based on the values of an independent (or explanatory) variable.
- measure the variability (residual) of the dependent variable about a straight line (also called regression line) and examine whether regression model fits to the data.

## 14.1 INTRODUCTION

In Chapter 13, we introduced the concept of statistical relationship between two variables such as level of sales and amount of advertising; yield of a crop and the amount of fertilizer used; price of a product and its supply, and so on. Such statistical relationship indicates the degree (strength) and direction of association between two variables but fails to ascertain whether there is any functional (or algebraic) relationship between two variables? If yes, can it be used to estimate the most likely value of one variable, given the value of other variable?

The statistical technique that expresses a functional (or algebraic) relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called *regression analysis*. The variable whose value is to be estimated is called *dependent* (or *response*) *variable* and the variable whose value is used to estimate this value is called *independent* (regressor or *predictor*) *variable*. The linear algebraic equations that express a dependent variable in terms of an independent variable are called *linear regression equation*.

Sir Francis Galton in 1877, while studying the relationship between the height of father and sons found that though 'tall father has tall sons', the average height of sons of tall father is $x$ above the general height and the average height of sons is $2x/3$ above the general height. He described such a fall in the average height as 'regression to mediocrity'. The term regression in the literary sense is also referred as ***'moving backward'***.

### Difference Between Correlation and Regression Analysis

1. Developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable is referred to as regression analysis.

2. Measuring the strength (or degree) and direction of the relationship between two variables is referred as correlation analysis. The direction (direct or inverse) of the relationship is indicated by the correlation coefficient, and the absolute value of correlation coefficient indicates the extent (strength or degree) of the relationship.
3. Correlation analysis determines the strength (or degree) of association between two variables $x$ and $y$ but does not establish a cause-and-effect relationship. Regression analysis establishes the cause-and-effect relationship between $x$ and $y$, that is, a change in the value of independent variable $x$ *causes* a change (*effect*) in the value of dependent variable, $y$ assuming that all other factors that may affect $y$ remain unchanged.
4. In linear regression analysis one variable is considered as dependent variable and other as independent variable, while in correlation analysis both variables are considered to be independent.
5. *The coefficient of determination $r^2$ indicates the proportion of total variance in the dependent variable that is explained or accounted for due to variation in the independent variable.* Since value of $r^2$ is determined from a sample, its value is subject to sampling error.

## 14.2 ADVANTAGES OF REGRESSION ANALYSIS

The following are few advantages of regression analysis:

1. Regression analysis helps in developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable around the regression line. Closer the pair of values ($x$, $y$) fall around the regression line, better the line fits the data and hence smaller the variance and error of estimate. Thus, a good estimate can be made of the value of variable $y$ when all the points fall on the line, i.e. standard error of estimate equals zero.
3. If the sample size is large ($n \geq 30$), then interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either $x$ or $y$. The magnitude of $r^2$ remains the same regardless of the values of the two variables.

## 14.3 TYPES OF REGRESSION MODELS

A regression model is an algebraic equation between two variables based on the given data and estimating the value of a dependent variable based on the known values of one or more independent variables. A particular form of regression model depends upon the nature of the problem under study and the type of data available.

### 14.3.1 Simple and Multiple Regression Models

If a regression model represents the relationship between a dependent, $y$, and only one independent variable, $x$, then such a regression model is called a *simple regression model*. But if more than one independent variable is associated with a dependent variable, then such a regression model is called a *multiple regression model*. For example, sales turnover of a product (a dependent variable) is associated with more than one independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors and so on. Thus, estimation of possible sales turnover with respect to only one of these independent variables is an example of a simple regression model, otherwise multiple regression model.

### 14.3.2 Linear and Non-linear Regression Models

If the change (increase or decrease) in the values of a dependent (response) variable $y$ in a regression model is directly proportional to a unit change (increase or decrease) in the values of independent (predictor) variable $x$, then such a model is called a **linear regression model**. Thus, the relationship between these two variables can be represented by a straight-line relationship in terms of population parameters $\beta_0$ and $\beta_1$ as follows:

$$E(y) = \beta_0 + \beta_1 x \qquad (14\text{-}1)$$

where $\beta_0$ = $y$-intercept that represents mean (or average) value of the dependent variable $y$ when $x = 0$.

$\beta_1$ = slope of the regression line that represents the expected change (positive or negative) in the value of dependent variable, $y$ for a unit change in the value of independent variable, $x$.

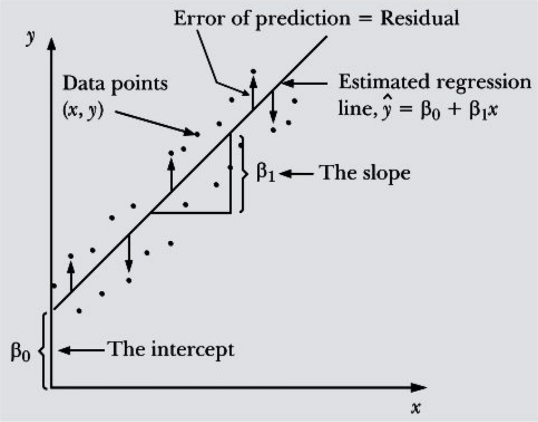


**Figure 14.1**
Straight Line Relationship

The intercept $\beta_0$ and the slope $\beta_1$ are *unknown regression coefficients*. The value of both $\beta_0$ and $\beta_1$ is to be calculated to predict average value of $y$ for a given value of $x$ by substituting these values in equation (14-1).

Figure 14.1 presents a scatter diagram of each pair of values $(x_i, y_i)$ around the regression line. Although, mean (or average) value of dependent variable, $y$, is a linear function of independent variable, $x$, but not all values of $y$ fall exactly on the straight line. Since few points do not fall on the regression line, therefore values of $y$ are not exactly equal to the values obtained by equation (14-1). Thus, such a straight line is also called ***line of mean deviations*** of observed $y$ value from the regression line. This situation arises due to *random error* (also called *residual variation or residual error*) in the prediction of the value of dependent variable $y$ for given value of independent variable, $x$. This implies that the variable, $x$, is not alone responsible for all variability in the value of variable, $y$. For example, sales volume is related to the level of expenditure on advertisement, but if other factors related to sales such as price of the product, quality of the product, competitors, etc., are ignored, then a regression equation to predict the sales volume ($y$) based on budget of advertising ($x$) only may cause an error. Thus for a fixed value of independent variable, $x$, the actual value of dependent variable, $y$, is determined by the ***mean value function plus a random error term,*** $e$, as follows:

$$\begin{aligned} y &= \text{Mean value function} + \text{Deviation} \\ &= \beta_0 + \beta_1 x + e \qquad (14\text{-}2) \end{aligned}$$

The equation (14-2) is referred to as **simple probabilistic linear regression model**. The error term, $e$, in equation (14-2) is called *random error* because its value associated with each value of variable, $y$, is assumed to vary unpredictably. The extent of random error associated with each value of variable, $y$, for a given value of $x$ is measured by the error variance. Lower the value of, $e$, better the regression model fit to a sample data.

The random errors corresponding to different observations $(x_i, y_i)$ for all $i$ are assumed to follow a normal distribution with mean zero and (unknown) constant standard deviation.

If the line passing through the pair of values of variables $x$ and $y$ is not linear, then the relationship between variables $x$ and $y$ is *non-linear*. A non-linear relationship implies that expected change (positive or negative) in the value of dependent variable, $y$, is not directly proportional to a unit change in the value of independent variable, $x$. A non-linear relationship is not very useful for predictions.

In this chapter, we will discuss methods of simple linear regression analysis involving single independent variable, whereas those involving two or more independent variables will be discussed in Chapter 15.

## 14.4 ESTIMATION: THE METHOD OF LEAST SQUARES

A sample of $n$ pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ is drawn from the population under study to estimate the values of regression coefficients $\beta_0$ and $\beta_1$. The method that provides the best linear unbiased estimate of the values of $\beta_0$ and $\beta_1$ is called the **method of least squares**. The estimated values of $\beta_0$ and $\beta_1$ should result in a straight line where most pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ fall very close (*best fit*) to it. Such a straight line is referred to as *'best fitted' (least squares or estimated) regression line because the sum of the squares of the vertical deviations (difference between the actual values of y and the estimated values predicted from the fitted line) is as small as possible.*

Rewriting equation (14-2) as follows:

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ or } e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for all } i$$

Mathematically, we intend to minimize

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let $b_0$ and $b_1$ be the least-squares estimators of $\beta_0$ and $\beta_1$, respectively. The least-squares estimators $b_0$ and $b_1$ must satisfy following equations:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{b_0 b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{b_0 b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\,x_i = 0$$

After simplifying these two equations, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \tag{14-3}$$

$$\sum_{i=1}^{n} x_i\, y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Equation (14-3) is called the *least-squares normal equation*. The values of least squares estimators $b_0$ and $b_1$ can be obtained by solving equation (14-3). Hence, the *fitted* or *estimated regression line* is given by

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ (called $y$ hat) is the *predicted value of* $y$ falling on the fitted regression line for a given value of $x$ and $e_i = y_i - \hat{y}_i$ is called the *residual* that describes the amount of error in fitting of the regression line to the values of $y_i$.

**Remark:** $\sum e_i = 0$, i.e. sum of the residuals is zero for any least-squares regression line.

## 14.5 ASSUMPTIONS FOR A SIMPLE LINEAR REGRESSION MODEL

To form a basis for application of simple linear regression models, certain assumptions about the population from which a sample of observations is drawn and the way in which observations are made to draw statistical inference are illustrated in Figure 14.2.

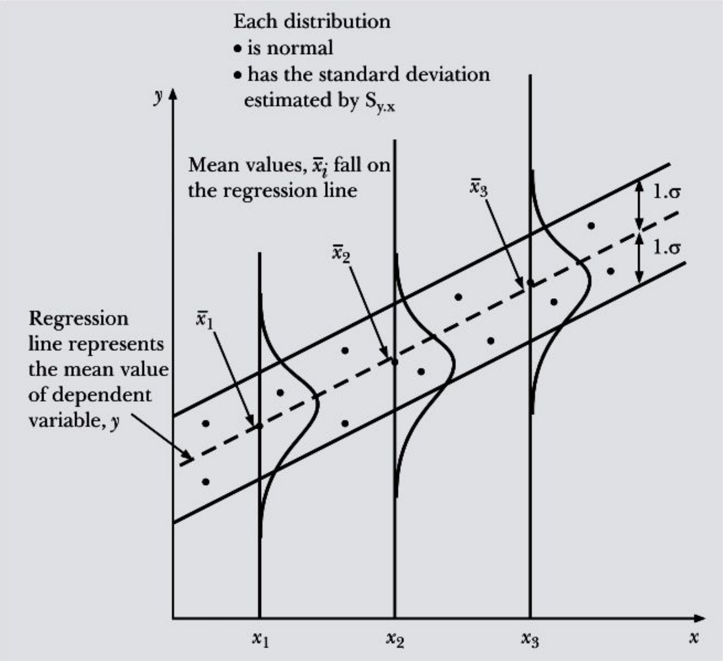

Each distribution
• is normal
• has the standard deviation estimated by $S_{y.x}$
Mean values, $\bar{x}_i$ fall on the regression line
Regression line represents the mean value of dependent variable, $y$
$y$
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
1.σ
1.σ
$x_1$
$x_2$
$x_3$
$x$

**Figure 14.2**
Assumptions in Regression Analysis

### Assumptions

1. There is a linear relationship between the dependent variable $y$ and independent variable $x$. This relationship can be described by a linear regression equation $y = a + bx + e$, where $e$ represents the deviation in the value of dependent variable, $y$, from its expected value for a given value of independent variable, $x$.
2. The set of expected (or mean) values of the dependent variable, $y$, for given values of independent variable, $x$, are normally distributed. The mean of these normally distributed values falls on the line of regression.
3. The dependent variable $y$ is a continuous random variable, whereas values of the independent variable $x$ are fixed and not random.
4. The sampling error associated with the expected value of the dependent variable $y$ is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The amount of deviation (error) in the value of dependent variable, $y$, may be different in successive observations.
5. The standard deviation and variance of expected values of the dependent variable about the regression line are constant for all values of the independent variable $x$ for the set of observations in a sample.
6. The expected value of the dependent variable cannot be obtained for a value of an independent variable falling outside the range of values in the sample.

## 14.6 PARAMETERS OF SIMPLE LINEAR REGRESSION MODEL

The objective regression analysis is to ascertain that a regression equation (line) should provide the best fit of sample data to the population data so that the error of variance is as small as possible. J. R. Stockton stated that *the device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression*.

The two variables $x$ and $y$ which are correlated can be expressed in terms of each other in the form of straight line equations called *regression equations* as follows:

- The regression equation of $y$ on $x$

$$y = a + bx$$

is used for estimating the value of $y$ for given values of $x$.

- Regression equation of $x$ on $y$

$$x = c + dy$$

is used for estimating the value of $x$ for given values of $y$.

### Remarks

1. Regression lines coincide (overlap) when variables $x$ and $y$ are perfectly correlated (either positive or negative).
2. Higher the degree of correlation, the two regression lines come closer to each other. Lesser the degree of correlation, the two regression lines are away from each other. Also, when variables $x$ and $y$ are not correlated, i.e. $r = 0$, the two regression lines are at right angle to each other.
3. The point of interaction of two linear regression lines represents the average value of variables $x$ and $y$.

## 14.6.1 Regression Coefficients

To estimate values of population parameter $\beta_0$ and $\beta_1$, the fitted (or estimated) simple linear regression model (equation or line) is written as

$$\hat{y} = a + bx$$

where $\hat{y}$ is estimated average (mean) value of dependent variable $y$ for a given value of independent variable $x$; $a$ or $b_0$ is $y$-intercept that represents average value of $\overline{y}$; and $b$ is the slope of regression line that represents the expected change in the value of $y$ for unit change in the value of $x$.

To determine the value of for a given value of $x$, this equation requires the determination of two unknown constants $a$ (intercept) and $b$ (also called regression coefficient). Once these constants are calculated, the regression line can be used to compute an estimated value of the dependent variable $y$ for a given value of independent variable $x$.

The particular values of $a$ and $b$ define a specific linear relationship between $x$ and $y$ based on sample data. The coefficient '$a$' represents the *level of fitted line* (i.e., the distance of the line above or below the origin) when $x$ equals zero, whereas coefficient '$b$' represents the *slope of the line* (a measure of the change in the estimated value of $y$ for a one-unit change in $x$).

The regression coefficient '$b$' is also denoted as

- $b_{yx}$ (*regression coefficient of y on x*) in the regression line, $y = a + bx$
- $b_{xy}$ (*regression coefficient of x on y*) in the regression line, $x = c + dy$

### Properties of Regression Coefficients

1. The correlation coefficient is the geometric mean of two regression coefficients, i.e., $r = \sqrt{b_{yx} \times b_{xy}}$.
2. If one regression coefficient is greater than one, then other regression coefficient must be less than one because the value of correlation coefficient $r$ cannot exceed one. However, both the regression coefficients may be less than one.
3. Both regression coefficients must have the same sign (either positive or negative). This property rules out the case of opposite sign of two regression coefficients.
4. The correlation coefficient will have the same sign (either positive or negative) as that of the two regression coefficients. For example, if $b_{yx} = -0.664$ and $b_{xy} = -0.234$, then $r = -\sqrt{0.664 \times 0.234} = -0.394$.
5. The arithmetic mean of regression coefficients $b_{xy}$ and $b_{yx}$ is more than or equal to the correlation coefficient, $r$, i.e., $(b_{yx} + b_{xy})/2 \geq r$. For example, if $b_{yx} = -0.664$ and

$b_{xy}$ = – 0.234, then the arithmetic mean of these two values is (– 0.664 – 0.234)/2 = – 0.449, and this value is more than the value of $r$ = – 0.394.

6. Regression coefficients are independent of origin but not of scale.

## 14.7 METHODS TO DETERMINE REGRESSION COEFFICIENTS

The following are the methods to determine the parameters of a fitted regression equation.

### 14.7.1 Least Squares Normal Equations

Let $\hat{y} = a + bx$ be the least squares line of $y$ on $x$, where $\hat{y}$ is the estimated average value of dependent variable $y$. Since best-fitted least squares line minimizes the sum of squares of the deviations of the observed values of $y$ from $\hat{y}$, therefore sum of residuals for any least-square line is minimum, i.e.,

$$L = \Sigma\,(y - \hat{y})^2 = \Sigma\{y - (a + bx)\}^2 = \text{minimum, where } a, b = \text{constants}$$

Differentiating $L$ with respect to $a$ and $b$ and equating to zero, we have

$$\frac{\partial \mathrm{L}}{\partial a} = -2\Sigma\{y - (a + bx)\} = 0$$

$$\frac{\partial \mathrm{S}}{\partial b} = -2\Sigma\{y - (a + bx)\}x = 0$$

Solving these two equations, we get the same set of equations as equations (14-3)

$$\begin{aligned} \Sigma\, y &= n\,a + b\Sigma\, x \\ \Sigma\, xy &= a\,\Sigma x + b\Sigma\, x^2 \end{aligned} \qquad (14\text{-}4)$$

where $n$ is the total number of pairs of values of $x$ and $y$ in a sample data. The equation (14-4) is called *normal equations* with respect to the regression line of $y$ on $x$. After solving these equations for $a$ and $b$, the values of $a$ and $b$ are substituted in the regression equation, $y = a + bx$.

Similarly if a least squares line is $x = c + dy$ of $x$ on $y$, where $x$ is the estimated average value of dependent variable $x$, then the normal equations will be

$$\begin{aligned} \Sigma\, x &= nc + d\,\Sigma\, y \\ \Sigma\, xy &= n\,\Sigma\, y + d\,\Sigma\, y^2 \end{aligned}$$

These equations are solved for constants $c$ and $d$. The values of these constants are substituted to the regression equation $x = c + dy$.

#### Alternative Method to Calculate Value of Constants

Instead of using the algebraic method to calculate values of constants $a$ and $b$ or $c$ and $d$, we may directly use the results of the solutions of these normal equations.

The gradient '$b$' (regression coefficient of $y$ on $x$) and '$d$' (regression coefficient of $x$ on $y$) are calculated as

$$b = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}}, \quad \text{where } \mathrm{S}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$\mathrm{S}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

and

$$d = \frac{\mathrm{S}_{yx}}{\mathrm{S}_{yy}}, \quad \text{where } \mathrm{S}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y\right)^2$$

Since the regression line passes through the point ($\bar{x}, \bar{y}$), the mean values of $x$ and $y$ and the regression equations can be used to find the value of constants $a$ and $c$ as follows:

$$a = \bar{y} - b\bar{x} \text{ for regression equation of } y \text{ on } x$$

$$c = \bar{x} - d\,\bar{y} \text{ for regression equation of } x \text{ on } y$$

The calculated values of $a$, $b$ and $c$, $d$ are substituted in the regression line $y = a + bx$ and $x = c + dy$, respectively, to determine the exact relationship.

**Example 14.1:** Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 per cent in advertising expenditure.

| *Firm* | *Annual Percentage Increase in Advertising Expenditure* | *Annual Percentage Increase in Sales Revenue* |
|---|---|---|
| A | 1 | 1 |
| B | 3 | 2 |
| C | 4 | 2 |
| D | 6 | 4 |
| E | 8 | 6 |
| F | 9 | 8 |
| G | 11 | 8 |
| H | 14 | 9 |

**Solution:** Assume sales revenue ($y$) is dependent on advertising expenditure ($x$). Calculations for regression line using following normal equations are shown in Table 14.1

$$\Sigma y = na + b\Sigma x \text{ and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

**Table 14.1** Calculation for Normal Equations

| *Sales Revenue* $y$ | *Advertising Expenditure, x* | $x^2$ | $xy$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 3 | 9 | 6 |
| 2 | 4 | 16 | 8 |
| 4 | 6 | 36 | 24 |
| 6 | 8 | 64 | 48 |
| 8 | 9 | 81 | 72 |
| 8 | 11 | 121 | 88 |
| 9 | 14 | 196 | 126 |
| 40 | 56 | 524 | 373 |

*Normal Equations Approach:*

$$\Sigma y = na + b\Sigma x \quad \text{or} \quad 40 = 8a + 56b$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{or} \quad 373 = 56a + 524b$$

Solving these equations, we get $a = 0.072$ and $b = 0.704$

Substituting these values in the regression equation

$$y = a + bx = 0.072 + 0.704x$$

For $x = 7.5$ per cent or 0.075 an increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704\,(0.075) = 0.1248 \text{ or } 12.48\%$$

*Short-cut Method*

$$b = \frac{S_{xy}}{S_{xx}} = \frac{93}{132} = 0.704,$$

where $S_{xy} = \Sigma xy - \dfrac{\Sigma x \Sigma y}{n} = 373 - \dfrac{40 \times 56}{8} = 93$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 524 - \frac{(56)^2}{8} = 132$$

The intercept '$a$' on the $y$-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{40}{8} - 0.704 \times \frac{56}{8} = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of $a = 0.072$ and $b = 0.704$ in the regression equation, we get

$$y = a + bx = 0.072 + 0.704\,x$$

For $x = 0.075$, we have $y = 0.072 + 0.704\ (0.075) = 0.1248$ or 12.48 per cent.

**Example 14.2:** The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in ₹1000's and analysed the results

| Week : | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Sales : | 2.69 | 2.62 | 2.80 | 2.70 | 2.75 | 2.81 |

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

**Solution:** Assume sales ($y$) are dependent on weeks ($x$). Then the normal equations for regression equation: $y = a + bx$ are written as

$$\Sigma y = n\,a + b\,\Sigma x \text{ and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

Calculations for sales during various weeks are shown in Table 14.2.

**Table 14.2** Calculations of Normal Equations

| *Week* ($x$) | *Sales* ($y$) | $x^2$ | $xy$ |
|---|---|---|---|
| 1 | 2.69 | 1 | 2.69 |
| 2 | 2.62 | 4 | 5.24 |
| 3 | 2.80 | 9 | 8.40 |
| 4 | 2.70 | 16 | 10.80 |
| 5 | 2.75 | 25 | 13.75 |
| 6 | 2.81 | 36 | 16.86 |
| 21 | 16.37 | 91 | 57.74 |

The gradient '$b$' is calculated as

$$b = \frac{S_{xy}}{S_{xx}} = \frac{0.445}{17.5} = 0.025; \quad S_{xy} = \Sigma xy - \frac{\Sigma x\,\Sigma y}{n} = 57.74 - \frac{21 \times 16.37}{6} = 0.445$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 91 - \frac{(21)^2}{6} = 17.5$$

The intercept '$a$' on the $y$-axis is calculated as

$$a = \bar{y} - b\bar{x} = \frac{16.37}{6} - 0.025 \times \frac{21}{6}$$

$$= 2.728 - 0.025 \times 3.5 = 2.64$$

Substituting the values $a = 2.64$ and $b = 0.025$ in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For $x = 7$, we have $y = 2.64 + 0.025(7) = 2.815$

Hence, the expected sales during the 7th week are likely to be ₹2.815 (in ₹1000's).

### 14.7.2 Deviations Method

Computation time while using least squares normal equations method becomes lengthy when values of $x$ and $y$ are in more than two digits. The computational time may be reduced by using following two methods:

(a) **Deviations Taken from Actual Mean Values of *x* and *y*** If deviations of actual values of variables $x$ and $y$ are taken from their mean values, then regression equations can be written as

- Regression equation of $y$ on $x$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where $b_{yx}$ = regression coefficient of $y$ on $x$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

- Regression equation of $x$ on $y$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

where $b_{xy}$ = regression coefficient of $x$ on $y$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

(b) **Deviations Taken from Assumed Mean Values for *x* and *y*** If mean value of either $x$ or $y$ or both are not integer, then prefer to take deviations of actual values of variables $x$ and $y$ from their assumed means.

- Regression equation of $y$ on $x$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{where } b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2}$$

$n$ = number of observations
$d_x = x - A$; A is assumed mean of $x$
$d_y = y - B$; B is assumed mean of $y$

- Regression equation of $x$ on $y$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\text{where } b_{xy} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_y^2 - (\Sigma d_y)^2}$$

$n$ = number of observations
$dx = x - A$; A is assumed mean of $x$
$dy = y - B$; B is assumed mean of $y$

(c) **Regression Coefficients in Terms of Correlation Coefficient** If deviations are taken from actual mean values, then the values of regression coefficients can be calculated as follows:

$$byx = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\text{Covariance}(x, y)}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$bxy = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\text{Covariance}(x, y)}{\sigma_y^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

**Example 14.3:** The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (₹ in 1000's)

| Salesmen : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Test scores : | 50 | 60 | 50 | 60 | 80 | 50 | 80 | 40 | 70 |
| Weekly sales : | 30 | 60 | 40 | 50 | 60 | 30 | 70 | 50 | 60 |

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales. *[HP Univ., M.Com.,2006 ]*

**Solution:** Assume weekly sales ($y$) as dependent variable and test scores ($x$) as independent variable. Calculations for the following regression equation are shown in Table 14.3.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

**Table 14.3** Calculation for Regression Equation

| *Weekly Sales, x* | $dx = x - 60$ | $d_x^2$ | *Test Score, y* | $dy = y - 50$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 50 | −10 | 100 | 30 | −20 | 400 | 200 |
| 60 | 0 | 0 | 60 | 10 | 100 | 0 |
| 50 | −10 | 100 | 40 | −10 | 100 | 100 |
| 60 | 0 | 0 | 50 | 0 | 0 | 0 |
| 80 | 20 | 400 | 60 | 10 | 100 | 200 |
| 50 | −10 | 100 | 30 | −20 | 400 | 200 |
| 80 | 20 | 400 | 70 | 20 | 400 | 400 |
| 40 | −20 | 400 | 50 | 0 | 0 | 0 |
| 70 | 10 | 100 | 60 | 10 | 100 | 100 |
| 540 | 0 | 1600 | 450 | 0 | 1600 | 1200 |

(a) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{540}{9} = 60;\quad \bar{y} = \dfrac{\Sigma y}{n} = \dfrac{450}{9} = 50$

$$b_{yx} = \frac{\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{1200}{1600} = 0.75$$

Substituting values in the regression equation, we have

$$y - 50 = 0.75\,(x - 60) \text{ or } y = 5 + 0.75x$$

For test score $x = 65$ of salesman, we have

$$y = 5 + 0.75\,(65) = 53.75$$

Hence, we conclude that the weekly sales are expected to be ₹53.75 (₹ in 1000's) for a test score of 65.

**Example 14.4:** A company is introducing a job evaluation scheme in which all jobs are graded by points for skill, responsibility, and so on. Monthly pay scales (₹ in 1000's) are then drawn up according to the number of points allocated and other factors such as experience and local conditions. To date the company has applied this scheme to 9 jobs:

| Job | : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Points | : | 5 | 25 | 7 | 19 | 10 | 12 | 15 | 28 | 16 |
| Pay (₹) | : | 3.0 | 5.0 | 3.25 | 6.5 | 5.5 | 5.6 | 6.0 | 7.2 | 6.1 |

(a) Find the least-squares regression line for linking pay scales to points.
(b) Estimate the monthly pay for a job graded by 20 points.

**Solution:** Assume monthly pay ($y$) as the dependent variable and job grade points ($x$) as the independent variable. Calculations for the following regression equation are shown in Table 14.4.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

**Table 14.4** Calculations for Regression Equation

| *Grade Points, x* | $d_x = x - 15$ | $d_x^2$ | *Pay Scale, y* | $d_y = y - 5$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 5 | −10 | 100 | 3.0 | −2.0 | 4 | 20 |
| 25 | 10 | 100 | 5.0 ← B | 0 | 0 | 0 |
| 7 | −8 | 64 | 3.25 | −1.75 | 3.06 | 14 |
| 19 | 4 | 16 | 6.5 | 1.50 | 2.25 | 6 |
| 10 | −5 | 25 | 5.5 | 0.50 | 0.25 | −2.5 |
| 12 | −3 | 9 | 5.6 | 0.60 | 0.36 | −1.8 |
| 15 ← A | 0 | 0 | 6.0 | 1.00 | 1.00 | 0 |
| 28 | 13 | 169 | 7.2 | 2.2 | 4.84 | 28.6 |
| 16 | 1 | 1 | 6.1 | 1.1 | 1.21 | 1.1 |
| 137 | 2 | 484 | 48.15 | 3.15 | 16.97 | 65.40 |

(a) $\bar{x} = \frac{\Sigma x}{n} = \frac{137}{9} = 15.22;\ \bar{y} = \frac{\Sigma y}{n} = \frac{48.15}{9} = 5.35$

Since mean values $\bar{x}$ and $\bar{y}$ are non-integer value, therefore deviations are taken from assumed mean as shown in Table 14.4.

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9 \times 65.40 - 2 \times 3.15}{9 \times 484 - (2)^2} = \frac{582.3}{4352} = 0.133$$

Substituting values in the regression equation, we have

$$y - \bar{y} = b_{yx}(x - \bar{x}) \text{ or } y - 5.35 = 0.133(x - 15.22) = 3.326 + 0.133x$$

(b) For job grade point $x = 20$, the estimated average pay scale is given by

$$y = 3.326 + 0.133x = 3.326 + 0.133\,(20) = 5.986$$

Hence, likely monthly pay for a job with grade points 20 is ₹5986.

**Example 14.5:** The following data, based on 450 students, are given for marks is Statistics and Economics at a certain examination:

| | | |
|---|---|---|
| Mean marks in Statistics | : | 40 |
| Mean marks in Economics | : | 48 |
| S.D. of marks in Statistics | : | 12 |
| The variance of marks in Economics | : | 256 |
| Sum of the product of deviation of marks from their respective mean | : | 42075 |

Obtain equations of the two lines of regression and estimate the average marks in Economics of candidates who obtained 50 marks in Statistics. [*Nagpur Univ., M.Com., 1996*]

**Solution:** Let the marks in Statistics be denoted by $x$ and marks in Economics by $y$. Then given that

$$\bar{x} = 40,\ \bar{y} = 48,\ \sigma_x = 12,\ \sigma_y = \sqrt{256} = 16.$$

Regression equation of $x$ on $y$ : $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

$$x - 40 = 0.487\frac{12}{16}(y - 48),$$

$$= 0.365\,y - 17.52 \text{ or } x = 22.48 + 0.365y.$$

where, $r = \frac{\Sigma d_x d_y}{n\,\sigma_x \sigma_y} = \frac{42075}{450 \times 12 \times 16} = 0.487$

Regression equation of $y$ on $x$ : $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$y - 48 = 0.487\frac{16}{12}(x - 40)$$

$$= 0.649x - 25.96 \text{ or } y = 22.04 + 0.649x.$$

The estimated marks in Economics for a candidate who has obtained $x = 50$ marks in Statistics will be

$$y = 22.04 + 0.649\,(50) = 54.49.$$

**Example 14.6:** The following data give the ages and blood pressure of 10 women.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | : | 56 | 42 | 36 | 47 | 49 | 42 | 60 | 72 | 63 | 55 |
| Blood pressure | : | 147 | 125 | 118 | 128 | 145 | 140 | 155 | 160 | 149 | 150 |

(a) Find the correlation coefficient between age and blood pressure.
(b) Determine the least-squares regression equation of blood pressure on age.
(c) Estimate the blood pressure of a woman whose age is 45 years.

[*Ranchi Univ. MBA*, 2003; *South Gujarat Univ., MBA, 2007*]

**Solution:** Assume blood pressure ($y$) as the dependent variable and age ($x$) as the independent variable. Calculations for regression equation of blood pressure on age are shown in Table 14.5.

**Table 14.5** Calculations for Regression Equation

| *Age, x* | $d_x = x - 49$ | $d_x^2$ | *Blood, y* | $d_y = y - 145$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 56 | 7 | 49 | 147 | 2 | 4 | 14 |
| 42 | −7 | 49 | 125 | −20 | 400 | 140 |
| 36 | −13 | 169 | 118 | −27 | 729 | 351 |
| 47 | −2 | 4 | 128 | −17 | 289 | 34 |
| $\boxed{49} \leftarrow$ A | 0 | 0 | $\boxed{145} \leftarrow$ B | 0 | 0 | 0 |
| 42 | −7 | 49 | 140 | −5 | 25 | 35 |
| 60 | 11 | 121 | 155 | 10 | 100 | 110 |
| 72 | 23 | 529 | 160 | 15 | 225 | 345 |
| 63 | 14 | 196 | 149 | 4 | 16 | 56 |
| 55 | 6 | 36 | 150 | 5 | 25 | 30 |
| 522 | 32 | 1202 | 1417 | −33 | 1813 | 1115 |

(a) Coefficient of correlation between age and blood pressure is given by

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{10(1115) - (32)(-33)}{\sqrt{10(1202) - (32)^2}\sqrt{10(1813) - (-33)^2}}$$

$$= \frac{11150 + 1056}{\sqrt{12020 - 1024}\sqrt{18130 - 1089}} = \frac{12206}{13689} = 0.892$$

We may conclude that there is a high degree of positive correlation between age and blood pressure.

(b) The regression equation of blood pressure on age is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{522}{10} = 52.2; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{1417}{10} = 141.7$$

and

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{10(1115) - 32(-33)}{10(1202) - (32)^2} = \frac{12206}{10996} = 1.11$$

Substituting these values in the above equation, we have

$$y - 141.7 = 1.11\ (x - 52.2) \text{ or } y = 83.758 + 1.11x$$

This is the required regression equation of $y$ on $x$.

($c$) For a women whose age is 45, the estimated average blood pressure will be

$$y = 83.758 + 1.11(45) = 83.758 + 49.95 = 133.708$$

Hence, the likely blood pressure of a woman of 45 years is 134.

**Example 14.7:** The General Sales Manager of Kiran Enterprises—an enterprise dealing in the sale of readymade men's wear—is toying with the idea of increasing his sales to ₹80,000. On checking the records of sales during the last 10 years, it was found that the annual sale proceeds and advertisement expenditure were highly correlated to the extent of 0.8. It was further noted that the annual average sale has been ₹45,000 and annual average advertisement expenditure ₹30,000, with a variance of ₹1600 and ₹625 in advertisement expenditure respectively.

In view of the above, how much expenditure on advertisement would you suggest the General Sales Manager of the enterprise to incur to meet his target of sales?

[*Kurukshetra Univ., MBA, 2008*]

**Solution:** Assume advertisement expenditure ($y$) as the dependent variable and sales ($x$) as the independent variable. Then the regression equation advertisement expenditure on sales is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Given $r = 0.8, \sigma_x = 40, \sigma_y = 25, \bar{x} = 45{,}000, \bar{y} = 30{,}000$. Substituting these values in the above equation, we have

$$(y - 30{,}000) = 0.8\,\frac{25}{40}(x - 45{,}000) = 0.5(x - 45{,}000)$$

$$y = 30{,}000 + 0.5x - 22{,}500 = 7500 + 0.5x$$

When a sales target is fixed at $x = 80{,}000$, the estimated amount likely to the spent on advertisement would be

$$y = 7500 + 0.5 \times 80{,}000 = 7500 + 40{,}000 = ₹47{,}500$$

**Example 14.8:** Find the most likely production corresponding to a rainfall of 40″ from the following data:

| | *Rainfall (x)* | *Production (y)* |
|---|---|---|
| Average | 30″ | 500 kg |
| Standard deviation | 5″ | 100 kg |

Coefficient of correlation, $r = 0.8$ *[GGSIP Univ., MBA, 2002]*

**Solution:** Given that, $r = 0.8, \bar{x} = 30, \bar{y} = 500, \sigma_y = 100$, and $\sigma_x = 5$.

The regression equation of production ($y$) on rainfall ($x$) is:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 500 = 0.8\frac{100}{5}(x - 30)$$

$$= 16x - 480 \text{ or } y = 20 + 16x.$$

When rainfall, $x = 40''$, the likely production will be $y = 16(40) + 20 = 660$ kg.

**Example 14.9:** You are given the following information about advertising expenditure and sales:

| | *Advertisement (x)*<br>*(Rs in lakh)* | *Sales (y)*<br>*(Rs in lakh)* |
|---|---|---|
| Arithmetic mean, $\bar{x}$ | 10 | 90 |
| Standard deviation, $\sigma$ | 3 | 12 |

Correlation coefficient = 0.8

(a) Obtain the two regression equations.

(b) Find the likely sales when advertisement budget is ₹15 lakh.

(c) What should be the advertisement budget if the company wants to attain sales target of ₹120 lakh? *[Kumaon Univ., MBA, 2000, MBA, Delhi Univ., 2002]*

**Solution:** (*a*) Regression equation of $x$ on $y$ is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Given $\bar{x} = 10, r = 0.8, \sigma_x = 3, \sigma_y = 12, \bar{y} = 90$. Substituting these values in the above regression equation, we have

$$x - 10 = 0.8\,\frac{3}{12}(y - 90) \text{ or } x = -8 + 0.2y$$

Regression equation of $y$ on $x$ is given by

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 90 = 0.8\ \frac{12}{3} \quad \text{or } y = 58 + 3.2x$$

(b) Substituting $x = 15$ in regression equation of $y$ on $x$. The likely average sales volume would be

$$y = 58 + 3.2(15) = 58 + 48 = 106$$

Thus the likely sales for advertisement budget of ₹15 lakh is ₹106 lakh.

(c) Substituting $y = 120$ in the regression equation of $x$ on $y$. The likely advertisement budget to attain desired sales target of ₹120 lakh would be

$$x = -8 + 0.2\ y = -8 + 0.2(120) = 16$$

Hence, the likely advertisement budget of ₹16 lakh should be sufficient to attain the sales target of ₹120 lakh.

**Example 14.10:** In a partially destroyed laboratory record of an analysis of regression data, the following results only are legible:

Variance of $x = 9$

Regression equations: $8x - 10y + 66 = 0$ and $40x - 18y = 214$

Find on the basis of the above information:

(a) The mean values of $x$ and $y$.

(b) Coefficient of correlation between $x$ and $y$.

(c) Standard deviation of $y$. *[Pune Univ., MBA, 2006]*

**Solution:** (a) Since two regression lines always intersect at a point $(\bar{x}, \bar{y})$ representing mean values of the variables involved, solving given regression equations to get the mean values $\bar{x}$ and $\bar{y}$ as shown below:

$$8x - 10y = -66$$
$$40x - 18y = 214$$

Multiplying the first equation by 5 and subtracting from the second, we have

$$32y = 544 \text{ or } y = 17, \text{ i.e. } \bar{y} = 17$$

Substituting the value of $y$ in the first equation, we get

$$8x - 10(17) = -66 \text{ or } x = 13, \text{ i.e. } \bar{x} = 13$$

(b) To find correlation coefficient $r$ between $x$ and $y$, we need to determine the regression coefficients $b_{xy}$ and $b_{yx}$.

Rewriting the given regression equations in such a way that the coefficient of dependent variable is less than one at least in one equation.

$$8x - 10y = -66 \text{ or } 10y = 66 + 8x \text{ or } y = \frac{66}{10} + \frac{8}{10}x$$

That is, $b_{yx} = 8/10 = 0.80$

$$40x - 18y = 214 \text{ or } 40x = 214 + 18y \text{ or } x = \frac{214}{40} + \frac{18}{40}y$$

That is, $b_{xy} = 18/40 = 0.45$

Hence coefficient of correlation $r$ between $x$ and $y$ is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.80} = 0.60$$

(c) To determine the standard deviation of $y$, consider the formula:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \text{ or } \sigma_y = \frac{b_{yx}\,\sigma_x}{r} = \frac{0.80 \times 3}{0.6} = 4$$

**Example 14.11:** There are two series of index numbers, P for price index and S for stock of a commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4, respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off values of S for various values of P?

**Solution:** The regression equation to read off values of P for various values S is given by

$$\text{P} = a + b\text{S} \text{ or } (\text{P} - \overline{\text{P}}) = r\frac{\sigma_p}{\sigma_s}(\text{S} - \overline{\text{S}})$$

Given $\overline{\text{P}} = 100$, $\overline{\text{S}} = 103$, $\sigma_p = 8$, $\sigma_s = 4$, $r = 0.4$. Substituting these values in the above equation, we have

$$\text{P} - 100 = 0.4\frac{8}{4} \text{ or } \text{P} = 17.6 + 0.8\text{ S}$$

This equation cannot be used to read off values of S for various values of P. Thus to read off values of S for various values of P, we use another regression equation of the form:

$$\text{S} = c + d\text{P} \text{ or } \text{S} - \overline{\text{S}} = \frac{\sigma_s}{\sigma_p}(\text{P} - \overline{\text{P}})$$

Substituting given values in this equation, we have

$$\text{S} - 103 = 0.4\ \frac{4}{8}\ (\text{P} - 100) \text{ or } \text{S} = 83 + 0.2\text{P}$$

**Example 14.12:** For certain $x$ and $y$ series which are correlated, the two lines of regression are:

$$5x - 6y + 90 = 0 \text{ and } 15x - 8y - 130 = 0.$$

Find the means of the two series and the correlation. *[MD Univ., M.Com., 2001]*

**Solution:** Solving two simultaneous regression equations to find mean value, we get $\overline{x} = 30$ and $\overline{y} = 40$.

Rewriting first regression equation as follows to find correlation coefficient, $r$:

$$6y = 5x + 90, \text{ i.e., } y = x\frac{5}{6} + 15 \text{ or } b_{yx} = \frac{5}{6}.$$

Also, $$15x = 8y + 130, \text{ i.e., } x = \frac{8}{15}y + \frac{130}{15} \text{ or } b_{xy} = \frac{8}{15}.$$

Since both regression coefficients $b_{xy}$ and $b_{yx}$ are less than one, applying following formula to get correlation coefficient:

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = 0.667.$$

**Example 14.13:** From 10 observations of price ($x$) and supply ($y$) of a commodity, the following summary figures were obtained (in appropriate units):

$$\Sigma x = 130;\ \Sigma y = 220;\ \Sigma x^2 = 2288;\ \Sigma y^2 = 5506; \text{ and } \Sigma xy = 3467$$

Compute a regression line of $y$ on $x$ and estimate the supply when the price is 16.

**Solution:** Given that $\overline{y} = \frac{1}{n}\Sigma y = \frac{220}{10} = 22$ and $\overline{x} = \frac{1}{n}\Sigma x = \frac{130}{10} = 13.$

Regression line of $y$ on $x$ is given by

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x}), \text{ where } r\frac{\sigma_y}{\sigma_x} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{607}{598} = 1.015.$$

$$\begin{aligned} y - 22 &= 1.015\ (x - 13) \\ &= 1.015\,x - 13.195 \\ n &= 1.015\,x + 8.805. \end{aligned}$$

When price $x = 16$, the corresponding supply $y$ becomes

$$y = 1.015\ (16) + 8.805 = 25.045.$$

Thus, the estimated supply is of 25.45 units when price is 16 units.

**Example 14.14:** For a given set of bivariate data, the following results were obtained:

$$\overline{x} = 53.2,\ \overline{y} = 27.9, \text{ Regression coefficient of } y \text{ on } x = -1.5$$

Regression coefficient of $x$ on $y = -0.2$. Find the most probable value of $y$ when $x$ is 60.

[*Mysore Univ., B.Com., 2002*]

**Solution:** For finding the most probable value of $y$ when $x = 60$, a regression equation of $y$ on $x$ is written as:

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x}),$$

$$y - 27.9 = -1.5(x - 53.2)$$

$$= -1.5\,x + 79.8 \quad \text{or} \quad y = 107.7 - 1.5\,x.$$

For $x = 60$, $y = 107.7 - 1.5\,(60) = 17.7$ and $r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{-0.2 \times -1.5} = -0.548$.

**Example 14.15:** The two regression lines obtained in a correlation analysis of 60 observations are

$$5x = 6x + 24 \text{ and } 1000y = 768x - 3708$$

What is the correlation coefficient and what is its probable error? Show that the ratio of the coefficient of variability of $x$ to that of $y$ is 5/24. What is the ratio of variances of $x$ and $y$?

**Solution:** Rewriting the regression equations

$$5x = 6y + 24 \text{ or } x = \frac{6}{5}y + \frac{24}{5}$$

That is, $b_{xy} = 6/5$. Also

$$1000y = 768x - 3708 \text{ or } y = \frac{768}{1000}x - \frac{3708}{1000}$$

That is, $b_{yx} = 768/1000$. Since

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ and } b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{768}{1000},$$

therefore $\quad b_{xy}b_{yx} = r^2 = \frac{6}{5} \times \frac{768}{1000} = 0.9216 \quad \text{or} \quad r = \sqrt{0.9216} = 0.96.$

$$\text{Probable error of } r = 0.6745\frac{1 - r^2}{\sqrt{n}} = 0.6745\ \frac{1 - (0.96)^2}{\sqrt{60}}$$

$$= \frac{0.0528}{7.7459} = 0.0068$$

Solving the given regression equations for $x$ and $y$, we get $\overline{x} = 6$ and $\overline{y} = 1$. Also

$$r\,\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } 0.96\frac{\sigma_x}{\sigma_y} = \frac{6}{5} \text{ or } \frac{\sigma_x}{\sigma_y} = \frac{6}{5 \times 0.96} = \frac{5}{4}$$

$$\text{Ratio of coefficient of variability} = \frac{\sigma_x/\overline{x}}{\sigma_y/\overline{y}} = \frac{\overline{y}}{\overline{x}}\cdot\frac{\sigma_x}{\sigma_y} = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}.$$

### 14.7.3 Regression Coefficients for Grouped Sample Data

If data set is grouped or classified into frequency distribution of either variable $x$ or $y$ or both, then values of regression coefficients $b_{xy}$ and $b_{yx}$ are calculated by using the formulae:

$$b_{xy} = \frac{n\Sigma\, d_x d_y - \Sigma\, fd_x\; \Sigma\, fd_y}{n\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2} \times \frac{h}{k}$$

$$b_{yx} = \frac{n\Sigma\, fd_x d_y - \Sigma\, fd_x\; \Sigma\, fd_y}{n\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2} \times \frac{k}{h}$$

where $h$ is width of the class interval of sample data on $x$ variable and $k$ is the width of the class interval of sample data on $y$ variable.

**Example 14.16:** The following bivariate frequency distribution relates to sales turnover (₹ in lakh) and money spent on advertising (₹ in 1000's). Obtain the two regression equations

| *Sales Turnover (₹ in lakh)* | *Advertising Budget (₹ in 1000's)* | | | |
|---|---|---|---|---|
| | *50–60* | *60–70* | *70–80* | *80–90* |
| 20– 50 | 2 | 1 | 2 | 5 |
| 50– 80 | 3 | 4 | 7 | 6 |
| 80–110 | 1 | 5 | 8 | 6 |
| 110–140 | 2 | 7 | 9 | 2 |

Estimate (a) the sales turnover corresponding to advertising budget of ₹1,50,000, and (b) the advertising budget to achieve a sales turnover of ₹200 lakh.

**Solution:** Let $x$ and $y$ represent sales turnover and advertising budget, respectively. Then the regression equation for estimating the sales turnover ($x$) on advertising budget ($y$) is expressed as

$$x - \bar{x} = b_{xy}\,(y - \bar{y})$$

where $b_{xy} = \dfrac{n\Sigma\, fd_x d_y - \Sigma\, fd_x\, \Sigma\, fd_y}{n\Sigma\, fd_y^2 - (\Sigma\, fd_y)^2}$

Similarly, the regression equation for estimating the advertising budget ($y$) on sales turnover of ₹200 lakh is written as

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where $b_{yx} = \dfrac{n\Sigma\, fd_x d_y - (\Sigma\, fd_x)\,(\Sigma\, fd_y)}{n\Sigma\, fd_x^2 - (\Sigma\, fd_x)^2}$

The calculations for regression coefficients $b_{xy}$ and $b_{yx}$ are shown in Table 14.6.

$$\bar{x} = \text{A} + \frac{\Sigma\, fd_x}{n} \times h = 65 + \frac{50}{70} \times 30 = 65 + 21.428 = 86.428$$

$$\bar{y} = \text{B} + \frac{\Sigma\, fd_y}{n} \times k = 75 - \frac{14}{70} \times 10 = 75 - 2 = 73$$

**Table 14.6:** Calculations for Regression Coefficients

| | | | *Advertising Budget* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Sales*<br>$x$ | $m.v.$ | $y$<br>$m.v$<br>$d_y$<br>$d_x$ | *50–60*<br>*55*<br>*−2* | *60–70*<br>*65*<br>*−1* | *70–80*<br>*75*<br>*0* | *80–90*<br>*85*<br>*1* | $f$ | $fd_x$ | $fd_x^2$ | $fd_x d_y$ |
| 20–50 | 35 | 1 | 2<br>4 | 1<br>1 | 2<br>— | 5<br>−5 | 10 | −10 | 10 | 0 |
| 50–80 | 65 | 0 | 3<br>— | 4<br>— | 7<br>— | 6<br>— | 20 | 0 | 0 | 0 |
| 80–110 | 95 | 1 | 1<br>−2 | 5<br>−5 | 8<br>— | 6<br>6 | 20 | 20 | 20 | −1 |
| 110–140 | 125 | 2 | 2<br>−8 | 7<br>−14 | 9<br>— | 2<br>4 | 20 | 40 | 80 | −18 |
| | | $f$ | 8 | 17 | 26 | 19 | $n$ = 70 | 50 = $\Sigma fd_x$ | 110 = $\Sigma fd_x^2$ | −19 $\Sigma fd_x d_y$ |
| | | $fd_y$ | −16 | −17 | 0 | 19 | −14 = $\Sigma fd_y$ | | | |
| | | $fd_y^2$ | 32 | 17 | 0 | 19 | 68 = $\Sigma fd_y^2$ | | | |
| | | $fd_x d_y$ | −6 | −18 | 0 | 5 | −19 = $\Sigma fd_x d_y$ | | | |

$$b_{xy} = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_y^2 - (\Sigma fd_y)^2} \times \frac{h}{k} = \frac{70\times -19-(50)(-14)}{70\times 68-(-14)^2}\times\frac{30}{10}$$

$$= \frac{-1330+700}{4760-196}\times\frac{30}{10} = \frac{-18,900}{45,640} = -0.414$$

$$b_{yx} = \frac{n\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{n\Sigma fd_x^2 - (\Sigma fd_x)^2} \times \frac{k}{h} = \frac{70\times -19-(50)(-14)}{70\times 110-(50)^2}\times\frac{10}{30}$$

$$= \frac{-1330+700}{7700-2500}\times\frac{10}{30} = \frac{-6300}{1,56,000} = -0.040$$

Substituting these values in the two regression equations, we get

(a) Regression equation of sales turnover ($x$) to advertising budget ($y$) is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 86.428 = -0.414\,(y - 73), \text{ or } x = 116.65 - 0.414y$$

For $y = 150$, we have $x = 116.65 - 0.414 \times 150 =$ ₹54.55 lakh

(b) Regression equation of advertising budget ($y$) on sales turnover ($x$) is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 73 = -0.040\,(x - 86.428) \text{ or } y = 76.457 - 0.04x$$

For $x = 200$, we have $y = 76.457 - 0.04\,(200) =$ ₹68.457 thousand.

## Self-practice Problems 14A

**14.1** The following calculations have been made for prices of twelve stocks ($x$) at the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares ($y$). From these calculations find the regression equation of price of stocks on the volume of sales of shares.

$$\Sigma x = 580, \quad \Sigma y = 370, \quad \Sigma xy = 11494,$$
$$\Sigma x^2 = 41658, \; \Sigma y^2 = 17206.$$

*[Rajasthan Univ., M.Com., 2005]*

**14.2** A survey was conducted to study the relationship between expenditure (in ₹) on accommodation ($x$) and expenditure on food and entertainment ($y$) and the following results were obtained:

| | *Mean* | *Standard Deviation* |
|---|---|---|
| • Expenditure on accommodation | 173 | 63.15 |
| • Expenditure on food and entertainment | 47.8 | 22.98 |

Coefficient of correlation $r = 0.57$

Write down the regression equation and estimate the expenditure on food and entertainment if the expenditure on accommodation is ₹200.

*[Bangalore Univ., B.Com.,2008]*

**14.3** The following data give the experience of machine operators and their performance ratings given by the number of good parts turned out per 100 pieces:

| Operator | : 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| experience ($x$) | : 16 | 12 | 18 | 4 | 3 | 10 | 5 | 12 |
| Performance ratings ($y$) | : 87 | 88 | 89 | 68 | 78 | 80 | 75 | 83 |

Calculate the regression lines of performance ratings on experience and estimate the probable performance if an operator has 7 years experience.

*[Jammu Univ., M.Com.; Lucknow Univ., MBA, 2006]*

**14.4** A study of prices of a certain commodity at Delhi and Mumbai yield the following data:

| | *Delhi* | *Mumbai* |
|---|---|---|
| • Average price per kilo (Rs) | 2.463 | 2.797 |
| • Standard deviation | 0.326 | 0.207 |
| • Correlation coefficient between prices at Delhi and Mumbai $r = 0.774$ | | |

Estimate from the above data the most likely price (a) at Delhi corresponding to the price of ₹2.334 per kilo at Mumbai (b) at Mumbai corresponding to the price of 3.052 per kilo at Delhi.

**14.5** The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

| Aptitude scores ($x$) | : 60 | 62 | 65 | 70 | 72 | 48 | 53 | 73 | 65 | 82 |
|---|---|---|---|---|---|---|---|---|---|---|
| Productivity index ($y$) | : 68 | 60 | 62 | 80 | 85 | 40 | 52 | 62 | 60 | 81 |

Calculate the two regression equations and estimate (a) the productivity index of a worker whose test score is 92, (b) the test score of a worker whose productivity index is 75.

*[Delhi Univ., MBA, 2005]*

**14.6** A company wants to assess the impact of R&D expenditure (₹ in 1000s) on its annual profit; (₹ in 1000's). The following table presents the information for the last eight years:

| *Year* | *R & D expenditure* | *Annual profit* |
|---|---|---|
| 1991 | 9 | 45 |
| 1992 | 7 | 42 |
| 1993 | 5 | 41 |
| 1994 | 10 | 60 |
| 1995 | 4 | 30 |
| 1996 | 5 | 34 |
| 1997 | 3 | 25 |
| 1998 | 2 | 20 |

Estimate the regression equation and predict the annual profit for the year 2002 for an allocated sum of ₹1,00,000 as R&D expenditure.

*[Jodhpur Univ., MBA, 2008]*

**14.7** Obtain the two regression equations from the following bivariate frequency distribution:

| *Sales Revenue (₹ in lakh)* | *Advertising Expenditure (₹ in thousand)* | | | |
|---|---|---|---|---|
| | *5–15* | *15–25* | *25–35* | *35-45* |
| 75–125 | 3 | 4 | 4 | 8 |
| 125–175 | 8 | 6 | 5 | 7 |
| 175–225 | 2 | 2 | 3 | 4 |
| 225–275 | 3 | 3 | 2 | 2 |

Estimate (a) the sales corresponding to advertising expenditure of ₹50,000, (b) the advertising expenditure for a sales revenue of ₹300 lakh, and (c) the coefficient of correlation.*[Delhi Univ., MBA, 2007]*

**14.8** The personnel manager of an electronic manufacturing company devises a manual test for job applicants to predict their production rating in the assembly department. In order to do this he selects a random sample of 10 applicants. They are given the test and later assigned a production rating. The results are as follows:

| Worker | : A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Test score | : 53 | 36 | 88 | 84 | 86 | 64 | 45 | 48 | 39 | 69 |
| Production rating | : 45 | 43 | 89 | 79 | 84 | 66 | 49 | 48 | 43 | 76 |

Fit a linear least squares regression equation of production rating on test score.

[*Delhi Univ., MBA, 2008*]

**14.9** Find the regression equation showing the capacity utilization on production from the following data:

| | | *Average Deviation* | *Standard Deviation* |
|---|---|---|---|
| Production (in lakh units) | : | 35.6 | 10.5 |
| Capacity utilization (in percentage) | : | 84.8 | 8.5 |
| Correlation coefficient $r = 0.62$ | | | |

Estimate the production when the capacity utilization is 70 per cent.

[*Delhi Univ., MBA, 2007; Pune Univ., MBA, 2008*]

**14.10** Suppose that you are interested in using past expenditure on R&D by a firm to predict current expenditures on R&D. You got the following data by taking a random sample of firms, where $x$ is the amount spent on R&D (in lakh of rupees) 5 years ago and $y$ is the amount spent on R&D (in lakh of rupees) in the current year:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x$ : | 30 | 50 | 20 | 80 | 10 | 20 | 20 | 40 |
| $y$ : | 50 | 80 | 30 | 110 | 20 | 20 | 40 | 50 |

(a) Find the regression equation of $y$ on $x$.

(b) If a firm is chosen randomly and $x = 10$, can you use the regression to predict the value of $y$? Discuss. [*Madurai-Kamraj Univ., MBA, 2005*]

**14.11** The following data relates to the scores obtained by a salesman of a company in an intelligence test and their weekly sales (in ₹1000's):

| Salesman intelligence | : | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Test score | : | 50 | 60 | 50 | 60 | 80 | 50 | 80 | 40 | 70 |
| Weekly sales | : | 30 | 60 | 40 | 50 | 60 | 30 | 70 | 50 | 60 |

(a) Obtain the regression equation of sales on intelligence test scores of the salesmen.

(b) If the intelligence test score of a salesman is 65, what would be his expected weekly sales?

[*HP Univ., M.com.,2006*]

**14.12** Two random variables have the regression equations:

$$3x + 2y - 26 = 0 \text{ and } 6x + y - 31 = 0$$

(a) Find the mean values of $x$ and $y$ and coefficient of correlation between $x$ and $y$.

(b) If the variance of $x$ is 25, then find the standard deviation of $y$ from the data.

[*MD Univ., M.Com., 2007; Kumaun Univ., MBA, 2005*]

**14.13** For a given set of bivariate data, the following results were obtained

$$\bar{x} = 53.2,\ \bar{y} = 27.9,$$

Regression coefficient of $y$ on $x = -1.5$, and Regression coefficient of $x$ and $y = -0.2$.

Find the most probable value of $y$ when $x = 60$.

**14.14** In trying to evaluate the effectiveness in its advertising campaign, a firm compiled the following information:

Calculate the regression equation of sales on advertising expenditure. Estimate the probable sales when advertisement expenditure is ₹ 60 thousand.

| *Year* | *Adv. expenditure (₹ 1000's)* | *Sales (in lakhs ₹)* |
|---|---|---|
| 2003 | 12 | 5.0 |
| 2004 | 15 | 5.6 |
| 2005 | 17 | 5.8 |
| 2006 | 23 | 7.0 |
| 2007 | 24 | 7.2 |
| 2008 | 38 | 8.8 |
| 2009 | 42 | 9.2 |
| 2010 | 48 | 9.5 |

[*Bharathidasan Univ., MBA, 2003*]

## Hints and Answers

**14.1** $\bar{x} = \Sigma x/n = 580/12 = 48.33;$

$\bar{y} = \Sigma y/n = 370/12 = 30.83$

$$b_{xy} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma y^2 - n(\bar{y})^2} = \frac{11494 - 12 \times 48.33 \times 30.83}{17206 - 12(30.83)^2}$$

$= -1.102$

Regression equation of $x$ on $y$:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 48.33 = -1.102\,(y - 30.83)$$

or $\quad x = 82.304 - 1.102y$

**14.2** Given $\bar{x} = 172$, $\bar{y} = 47.8$, $\sigma_x = 63.15$, $\sigma_y = 22.98$, and $r = 0.57$

Regression equation of food and entertainment ($y$) on accomodation ($x$) is given by

$$y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 47.8 = 0.57\,\frac{22.98}{63.15}(x - 173)$$

or $\quad y = 11.917 + 0.207x$

For $x = 200$, we have $y = 11.917 + 0.207(200)$

$= 53.317$

**14.3** Let the experience and performance rating be represented by $x$ and $y$ respectively.

$\bar{x} = \Sigma x/n = 80/8 = 10;\ \bar{y} = \Sigma y/n = 648/8 = 81$

$$b_{yx} = \frac{n\,\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\,\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{247}{218} = 1.133;$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$

Regression equation of $y$ on $x$

$$y - \overline{y} = byx\,(x - \overline{x})$$

or $\quad y - 81 = 1.133\,(x - 10)$

or $\quad y = 69.67 + 1.133x$

When $\quad x = 7,\ y = 69.67 + 1.133\,(7) = 77.60 \cong 78$

**14.4** Let price at Mumbai and Delhi be represented by $x$ and $y$, respectively

(a) Regression equation of $y$ on $x$

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x})$$

$$y - 2.463 = 0.774\ \frac{0.326}{0.207}(x - 2.797)$$

For $x$ = ₹2.334, the price at Delhi would be $y$ = ₹1.899.

(b) Regression on equation of $x$ on $y$

$$x - \overline{x} = r\frac{\sigma_x}{\sigma_y}(y - \overline{y})$$

or $\quad x - 2.791 = 0.774\,\dfrac{0.207}{0.326}(y - 2.463)$

For $y$ = ₹3.052, the price at Mumbai would be $x$ = ₹3.086.

**14.5** Let aptitude score and productivity index be represented by $x$ and $y$ respectively.

$\overline{x} = \Sigma x/n = 650/10 = 65;\ \overline{y} = \Sigma y/n = 650/10 = 65$

$$b_{xy} = \frac{n\,\Sigma\,d_x\,d_y - (\Sigma\,d_x)\,(\Sigma\,d_y)}{n\,\Sigma\,d_y^2 - \Sigma\,(d_y)^2} = \frac{1044}{1752} = 0.596;$$

where $d_x = x - \overline{x}$; $d_y = y - \overline{y}$

(a) Regression equation of $x$ on $y$

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

or $\quad x - 65 = 0.596\,(y - 65)$

or $\quad x = 26.26 + 0.596y$

When $\quad y = 75,\ x = 26.26 + 0.596(75) = 70.96 \cong 71$

(b) $b_{yx} = \dfrac{n\,\Sigma\,d_x d_y - (\Sigma d_x)\,(\Sigma d_y)}{n\,\Sigma d_x^2 - (\Sigma d_x)^2} = \dfrac{1044}{894} = 1.168$

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

or $\quad y - 65 = 1.168(x - 65)$

or $\quad y = -10.92 + 1.168x$

When $x = 92, y = -10.92 + 1.168(92) = 96.536 \cong 97$

**14.6** Let R&D expenditure and annual profit be denoted by $x$ and $y$ respectively

$\overline{x} = \Sigma x/n = 40/8 = 5.625;\ \overline{y} = \Sigma y/n = 297/8 = 37.125$

$$b_{yx} = \frac{n\,\Sigma\,d_x d_y - (\Sigma\,d_x)(\Sigma\,d_y)}{n\,\Sigma\,d_x^2 - (\Sigma\,d_x)^2} = \frac{8 \times 238 - (-3)\,(1)}{8 \times 57 - (-3)^2} = 4.266\ ;$$

where $d_x = x - 6,\ d_y = y - 37$

Regression equation of annual profit on R&D expenditure

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 37.125 = 4.26\,(x - 5.625)$$

or $\quad y = 13.163 + 4.266x$

For $x$ = ₹1,00,000 as R&D expenditure, we have from above equation $y$ = ₹439.763 as annual profit.

**14.7** Let sales revenue and advertising expenditure be denoted by $x$ and $y$ respectively

$$\overline{x} = \text{A} + \frac{\Sigma\,fd_x}{n} \times h = 150 + \frac{12}{66} \times 50 = 159.09$$

$$\overline{y} = \text{B} + \frac{\Sigma\,fd_y}{n} \times k = 30 - \frac{26}{66} \times 10 = 26.06$$

$$b_{xy} = \frac{n\,\Sigma\,fd_x d_y - (\Sigma\,fd_x)\,(\Sigma\,fd_y)}{n\,\Sigma\,fd_y^2 - (\Sigma\,fd_y)^2} \times \frac{h}{k}$$

$$= \frac{66\,(-14) - 12(-26)}{66(100) - (-26)^2} \times \frac{50}{10} = -0.516$$

(a) Regression equation of $x$ on $y$

$$x - \overline{x} = b_{xy}(y - \overline{y})$$

$$x - 159.09 = -0.516(y - 26.06)$$

or $\quad x = 172.536 - 0.516y$

For $y = 50,\ \ x = 147.036$

(b) Regression equation of $y$ on $x$

$$b_{yx} = \frac{n\,\Sigma\,fd_x d_y - (\Sigma\,fd_x)\,(\Sigma\,fd_y)}{n\,\Sigma\,fd_x^2 - (\Sigma\,fd_x)^2} \times \frac{k}{h}$$

$$= \frac{66\,(-14) - 12(-26)}{66\,(70) - (12)^2} \times \frac{10}{50} = -0.027.$$

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 26.06 = -0.027(x - 159.09)$$

$$y = 30.355 - 0.027x$$

For $x = 300, y = 22.255$

(c) $r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{0.516 \times 0.027} = -0.1180$

**14.8** Let test score and production rating be denoted by $x$ and $y$ respectively.

$\overline{x} = \Sigma x/n = 612/10 = 61.2;$

$\overline{y} = \Sigma y/n = 622/10 = 62.2$

$$b_{yx} = \frac{n\,\Sigma\,d_x d_y - (\Sigma\,d_x)\,(\Sigma d_y)}{n\,\Sigma\,d_x^2 - (\Sigma\,d_x)^2} = \frac{10 \times 3213 - 2 \times 2}{10 \times 3554 - (2)^2} = 0.904$$

Regression equation of production rating ($y$) on test score ($x$) is given by

$$y - \overline{y} = b_{yx}(x - \overline{x})$$

$$y - 62.2 = 0.904(x - 61.2)$$

$$y = 6.876 + 0.904x$$

**14.9** Let production and capacity utilization be denoted by $x$ and $y$, respectively.

(a) Regression equation of capacity utilization ($y$) on production ($x$)

$$y - \overline{y} = r\ \frac{\sigma_y}{\sigma_x}(x - \overline{x})$$

$$y - 84.8 = 0.62\ \frac{8.5}{10.5}(x - 35.6)$$

$$y = 66.9324 + 0.5019x$$

(b) Regression equation of production ($x$) on capacity utilization ($y$)

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$x - 35.6 = 0.62\ \frac{10.5}{8.5}(y - 84.8)$$

$$x = -29.3483 + 0.7659y$$

When $y = 70$, $x = -29.3483 + 0.7659(70) = 24.2647$

Hence the estimated production is 2,42,647 units when the capacity utilization is 70 per cent.

**14.10** $\bar{x} = \Sigma x/n = 270/8 = 33.75$; $\bar{y} = \Sigma y/n = 400/8 = 50$

$$b_{yx} = \frac{n\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{8\times 4800 - 6\times 0}{8\times 3592 - (6)^2}$$
$$= 1.338;$$

where $d_x = x - 33$ and $d_y = y - 50$

Regression equation of $y$ on $x$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 50 = 1.338(x - 33.75)$$
$$y = 4.84 + 1.338x$$

For $x = 10$, $y = 18.22$

**14.11** Let intelligence test score be denoted by $x$ and weekly sales by $y$

$$\bar{x} = 540/9 = 60;$$
$$\bar{y} = 450/9 = 50,$$
$$b_{yx} = \frac{n\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{9\times 1200}{9\times 1600} = 0.75$$

Regression equation of $y$ on $x$:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 50 = 0.75\ (x - 60)$$
$$y = 5 + 0.75x$$

For $x = 65$, $y = 5 + 0.75\ (65) = 53.75$

**14.12** (a) Solving two regression lines:

$$3x + 2y = 6 \text{ and } 6x + y = 31$$

we get mean values as = 4 and = 7

(b) Re-writing regression lines as follows:

$$3x + 2y = 26 \text{ or } y = 13 - (3/2)x,$$

So $\quad b_{yx} = -3/2$

$$6x + y = 31 \text{ or } x = 31/6 - (1/6)y,$$

So $\quad b_{xy} = -1/6$

Correlation coefficient,

$$r = \sqrt{b_{xy}\times b_{yx}} = -\sqrt{(3/2)(1/6)} = -0.5$$

Given, Var($x$) = 25, so $\sigma_x = 5$. Calculate $\sigma_y$ using the formula:

$$b_{yx} = r\frac{\sigma_y}{\sigma_y}$$

or $\quad -\dfrac{3}{2} = 0.5\dfrac{\sigma_y}{5}$ or $\sigma_y = 15$

**14.13** The regression equation of $y$ on $x$ is stated as:

$$y - \bar{y} = b_{xy}(x - \bar{x}) = r\cdot\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Given, $\bar{x} = 53.20$; $\bar{y} = 27.90$, $b_{yx} = -1.5$; $b_{xy} = -0.2$

Thus $y - 27.90 = -1.5(x - 53.20)$

or $\quad y = 107.70 - 1.5x$

For $x = 60$, we have $y = 107.70 - 1.5(60) = 17.7$

Also $\quad r = \sqrt{b_{yx}\times b_{xy}} = -\sqrt{1.5\times 0.2} = -0.5477$

**14.14** Let advertising expenditure and sales be denoted by $x$ and $y$ respectively.

$$\bar{x} = \Sigma x/n = 217/8 = 27.125;$$
$$\bar{y} = \Sigma y/n = 58.2/8 = 7.26$$
$$b_{yx} = \frac{n\sum dx\,dy - (\sum dx)(\sum dy)}{n\sum d_x^2 - (\sum dx)^2}$$
$$= \frac{8(172.2) - (25)(2.1)}{8(1403) - (25)^2} = \frac{1325.1}{10599} = 0.125$$

Thus regression equation of $y$ on $x$ is:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

or $\quad y - 7.26 = 0.125(x - 27.125)$

$$y = 3.86 + 0.125x$$

When $x = 60$, the estimated value of $y = 3.869 + 0.125(60) = 11.369$

## 14.8 STANDARD ERROR OF ESTIMATE AND PREDICTION INTERVALS

The distribution of expected values of dependent variable, $y$, about a least squares regression line for given values of independent variable $x$ indicates the strength (or extent) and direction of relationship between these two variables. For example, wide pattern of dot points indicates a poor relationship while a very close pattern of dot points indicates a close relationship between two variables. The variability among observed values of dependent variable, $y$, about the regression line is measured in terms of *residuals*. A residual is defined as the difference between an observed value of dependent variable $y$ and its estimated (or fitted) value for a given value of the independent variable $x$. The residual about the regression line is given by

$$\text{Residual, } e_i = y_i - \hat{y}_i$$

The residual values $e_i$ are plotted on a diagram with respect to the least squares regression line $\hat{y} = a + bx$. These residual values are the vertical distances of every observation (dot point) from the least squares line as shown in Fig. 14.3 and represent error of estimation for individual values of dependent variable.

Since sum of the residuals is zero, therefore it is not possible to determine the total amount of error by summing the residuals. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leftarrow \text{Error or residual sum of squares}$$

This quantity is also called the *sum of squares of errors (SSE)*.

The *variance of error of estimate* $\sigma_e^2$ or $S_{y.x}^2$ is determined as follows:

$$S_{yx}^2 \text{ or } \hat{\sigma}_e^2 = \frac{\Sigma e_i^2}{n-2} = \frac{\Sigma (y_i - y_i)^2}{n-2} = \frac{SSE}{n-2}$$

The denominator $n - 2$ represents the *residual degrees of freedom* and is obtained by subtracting from sample size, $n$ is the number of parameters $\beta_0$ and $\beta_1$ that are estimated by the sample parameters $a$ and $b$ in the least squares equation.

The *standard error of estimate (or standard deviation of error term),* $S_{yx}$ measures the variability of the observed values around the regression line. The standard deviation of error term, $S_{yx}$, about the least squares line is defined as

$$S_{yx} \text{ or } \sigma_e = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n-2}} \text{ or } \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\,\Sigma xy}{n-2}} = \sqrt{\frac{\text{SSE}}{n-2}} \qquad (14\text{-}4)$$

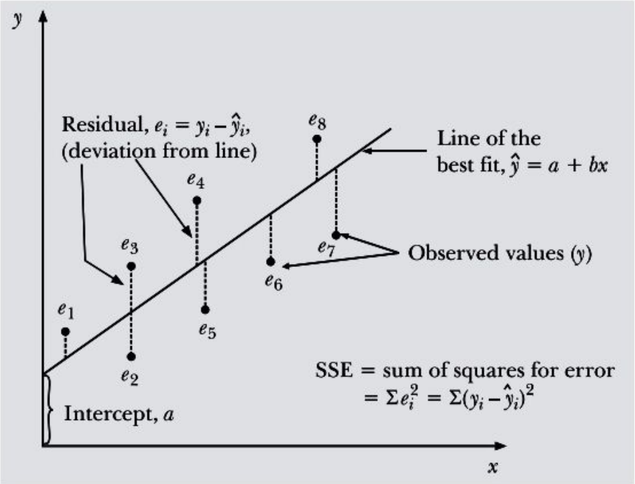

y
x
Residual, $e_i = y_i - \hat{y}_i$,
(deviation from line)
$e_1$
$e_2$
$e_3$
$e_4$
$e_5$
$e_6$
$e_7$
$e_8$
Line of the
best fit, $\hat{y} = a + bx$
Observed values (y)
SSE = sum of squares for error
$= \Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2$
Intercept, a

**Figure 14.3**
Residuals

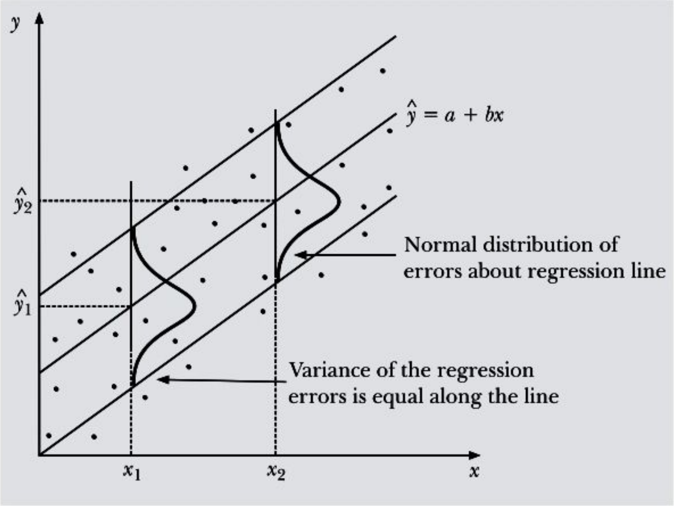
The variance $S_{yx}^2$ measures how the least squares line 'best fits' the sample $y$-values. A large *variance and standard error of estimate* indicate a large amount of dispersion of sample $y$-values (dot points) around the regression line. Smaller the value of $S_{yx}$, closer the $y$-values (dot points) fall around the regression line and better the line fits the data and describes the better average relationship between the two variables. If all dot points fall on the line, then value of $S_{yx}$ is zero, and the relationship between the two variables is said to be perfect.

Smaller the value of $S_{yx}$, is considered useful in predicting the value of a dependent variable $y$. Dispersion of sample $y$-values (dot points) around the regression line needs to be measured because

(i) it facilitate in predicting the value of the dependent variable.
(ii) value of $Syx$ is used for interval estimates of the dependent variable so as to draw statistical inferences results.

The distribution of expected values of dependent variable $y$ about a least squares regression line for given values of independent variable $x$ is shown in Fig. 14.4. Suppose the amount of deviation in the values of $y$ given any particular value of $x$ follow normal distribution. Since average value of $y$ changes with the value of $x$, we have different normal distributions of $y$-values for every value of $x$, each having same standard deviation. When a relationship between two variables $x$ and $y$ exists, the standard deviation (also called *standard error of estimate*) is less than the standard deviation of all the $x$-values in the population computed about their mean.

**Figure 14.4** Regression Line Showing the Error Variance


y
$\hat{y} = a + bx$
$\hat{y}_2$
$\hat{y}_1$
Normal distribution of errors about regression line
Variance of the regression errors is equal along the line
$x_1$
$x_2$
x

The standard error of estimate can also be used to determine an interval estimate (also called *prediction interval*) based on sample data ($n < 30$) for the value of the dependent variable, $y$, for a given value of the independent variable, $x$, as follows:

Approximate interval estimate $= \hat{y} \pm t_{df} S_{yx}$

where value of *t-statistics* is obtained from $t$-distribution table at given level of significance and degree of freedom.

**Example 14.17:** The following data relate to advertising expenditure (₹ in lakh) and their corresponding sales (₹ in crore)

| | | | | | | |
|---|---|---|---|---|---|---|
| Advertising expenditure | : | 10 | 12 | 15 | 23 | 20 |
| Sales | : | 14 | 17 | 23 | 25 | 21 |

(a) Find the equation of the least-squares line fitting the data.
(b) Estimate the value of sales corresponding to advertising expenditure of ₹30 lakh.
(c) Calculate the standard error of estimate of sales on advertising expenditure.

**Solution:** Let the advertising expenditure be denoted by $x$ and sales by $y$. The calculations for the least squares line are shown in Table 14.7

**Table 14.7:** Calculations for Least-squares Line

| *Advt. Expenditure, x* | $d_x = x - 16$ | $d_x^2$ | *Sales* $y$ | $d_y = y - 20$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 10 | −6 | 36 | 14 | −6 | 36 | 36 |
| 12 | −4 | 16 | 17 | −3 | 9 | 12 |
| 15 | −1 | 1 | 23 | 3 | 9 | −3 |
| 23 | 7 | 49 | 25 | 5 | 25 | 35 |
| 20 | 4 | 16 | 21 | 1 | 1 | 4 |
| 80 | 0 | 118 | 100 | 0 | 80 | 84 |

$$\bar{x} = \Sigma x/n = 80/5 = 16; \quad \bar{y} = \Sigma y/n = 100/5 = 20$$

$$b_{yx} = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{n\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{5 \times 84}{5 \times 118} = 0.712$$

(a) Regression equation of $y$ on $x$ is

$$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ y - 20 &= 0.712\,(x - 16) \\ y &= 8.608 + 0.712\,x \end{aligned}$$

where $a = 8.608$ and $b = 0.712$.

The fitted values and the residuals for the sample data in Table 14.7 are shown in Table 14.8. The fitted values are obtained by substituting values of $x$ in the least squares line (regression equation). For example, $8.608 + 0.712(10) = 15.728$. The residuals that indicate how well the least squares line fits the actual data are equal to the actual value minus fitted value.

**Table 14.8:** Fitted Values and Residuals for Sample Data

| *Value, x* | *Fitted Value*<br>$y = 8.608 + 0.712x$ | *Residuals* |
|---|---|---|
| 10 | 15.728 | –5.728 |
| 12 | 17.152 | –5.152 |
| 15 | 19.288 | –4.288 |
| 23 | 24.984 | –1.984 |
| 20 | 22.848 | –2.848 |

(b) The least squares line (equation) obtained in part (a) may be used to estimate the sales turnover corresponding to the advertising expenditure of ₹ 30 lakh as:

$$\hat{y} = 8.608 + 0.712x = 8.608 + 0.712\,(30) = ₹29.968 \text{ crore}$$

(c) Calculations for standard error of estimate, $S_{yx}$ of sales ($y$) on advertising expenditure ($x$) are shown in Table 14.9.

**Table 14.9:** Calculations for Standard Error of Estimate

| $x$ | $y$ | $y^2$ | $xy$ |
|---|---|---|---|
| 10 | 14 | 196 | 140 |
| 12 | 17 | 289 | 204 |
| 15 | 23 | 529 | 345 |
| 23 | 25 | 625 | 575 |
| 20 | 21 | 441 | 420 |
| 80 | 100 | 2080 | 1684 |

$$S_{yx} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}} = \sqrt{\frac{2080 - 8.608 \times 100 - 0.712 \times 1684}{5-2}}$$

$$= \sqrt{\frac{2080 - 860.8 - 1199}{3}} = 2.594$$

### 14.8.1 Coefficient of Determination: Partitioning of Total Variation

It is desired that the residual variance should be as small as possible but its value depends on the unit in which values of dependent variable, $y$, are measured. Consequently, another measure of fit called *coefficient of determination* is needed that is independent of the unit in which values of dependent variable, $y$, are measured. The *coefficient of determination is the proportion of variability of the dependent variable y accounted for or explained by the independent variable x.* In other words, it measures how well (i.e. strength) the regression line fits the data. The coefficient of determination is denoted by $r^2$ and its value ranges from 0 to 1. A particular of value $r^2$ should be interpreted as high or low in accordance of the use and context with which the regression model is developed. The coefficient of determination is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$= 1 - \frac{\text{Residual variation in response variable } y\text{-values from least-squares line}}{\text{Total variance of } y\text{-values}}$$

where SST = total sum of square deviations (or total variance) of actual values of variable, $y$ from its mean value.

$$= \text{S}_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

SSE = sum of squares of error (*unexplained variation*) in the values of dependent variable, $y$ from the least squares line due to sampling errors (i.e. amount of residual variation in the data that is not explained by independent variable, $x$)

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - a \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i y_i$$

SSR = sum of squares of regression (*or explained variation*) is the actual values of dependent variable $y$ accounted for or explained by variation among values of independent variable, $x$

= SST – SSE

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = a \sum_{i=1}^{n} y_i + b \sum_{i=1}^{n} x_i y_i - n(\bar{y})^2$$

These three variations noted during regression analysis of a data set are shown in Fig 14.5. Thus,

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{\text{S}_{yx}^2}{\text{S}_y^2} \; ; \; \text{S}_{y \cdot x} = \text{S}_y \sqrt{1 - r^2}$$

where $\dfrac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$ = fraction of the total variation that is explained or accounted for

$\text{S}y \cdot x = \dfrac{\Sigma(y - \hat{y})^2}{n - 2}$, variance in the values of independent variable, $y$ from the least squares line

$\text{S}_y^2 = \dfrac{1}{n-2} \Sigma(y - \bar{y})^2$ , total variance in the values of independent variable,

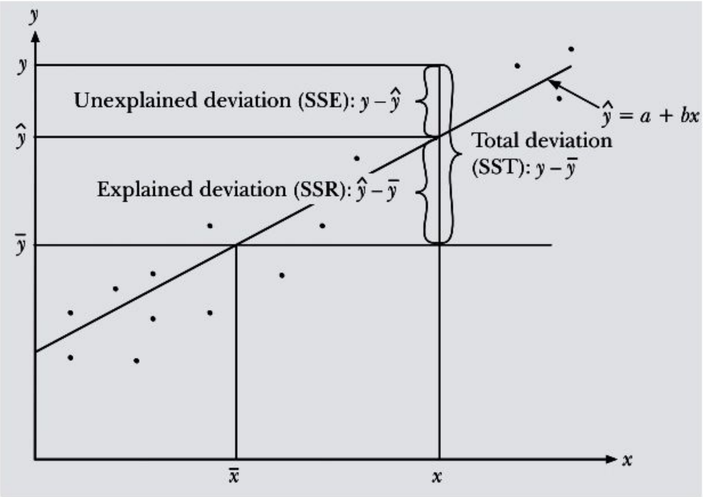

y
y
ŷ
ȳ
Unexplained deviation (SSE): y – ŷ
Explained deviation (SSR): ŷ – ȳ
Total deviation (SST): y – ȳ
ŷ = a + bx
x̄
x
x

**Figure 14.5**
Relationship Between Three Types of Variations

An easy formula of coefficient of determination, $r^2$, is given by

$$r^2 = \frac{a\,\Sigma\, y + b\,\Sigma\, xy - n\,(\overline{y})^2}{\Sigma\, y^2 - n\,(\overline{y})^2} \qquad \leftarrow \text{Short-cut method}$$

For example, the extent of relationship between sales revenue (*y*) and advertising expenditure (*x*) using data of Example 14.1 is computed as follows:

$$r^2 = \frac{a\,\Sigma\, y + b\,\Sigma\, xy - n\,(\overline{y})^2}{\Sigma\, y^2 - n\,(\overline{y})^2} = \frac{0.072 \times 40 + 0.704 \times 373 - 8(5)^2}{270 - 8(5)^2}$$

$$= \frac{2.88 + 262.592 - 200}{270 - 200} = \frac{65.47}{70} = 0.9352$$

The value $r^2 = 0.9352$ indicates that 93.52 per cent of the variance in sales revenue is on account of or statistically explained by advertising expenditure.

A comparison between bivariate correlation and regression analysis is summarized in Table 14-10.

**Table 14.10:** Comparison Between Linear Correlation and Regression

| | *Correlation* | *Regression* |
|---|---|---|
| • Measurement level | Interval or ratio scale | Interval or ratio scale |
| • Nature of variables | Both continuous, and linearly related | Both continuous, and linearly related |
| • $x - y$ relationship | *x* and *y* are symmetric | *y* is dependent, *x* is independent; regression of *x* on *y* differs from *y* on *x* |
| • Correlation | $b_{xy} = b_{yx}$ | Correlation between *x* and *y* is the same as the correlation between *y* and *x* |
| • Coefficient of determination | Explains common variance of *x* and *y* | Proportion of variability of *x* explained by its least-squares regression on *y* |

## Conceptual Questions 14A

1. (a) Explain the concept of regression and point out its usefulness in dealing with business problems. [*Delhi Univ., MBA, 2003*]

   Distinguish between correlation and regression. Also point out the properties of regression coefficients.

2. Explain the concept of regression and point out its importance in business forecasting. [*Delhi Univ., MBA, 2000, 2008*]

3. Under what conditions can there be one regression line? Explain. [*HP Univ., MBA, 2006*]

4. Why should a residual analysis always be done as part of the development of a regression model?

5. What are the assumptions of simple linear regression analysis and how can they be evaluated?

6. What is the meaning of the standard error of estimate?

7. What is the interpretation of *y*-intercept and the slope in a regression model?

8. What are regression lines? With the help of an example illustrate how they help in business decision-making. [*Delhi Univ., MBA, 2008*]

9. Point out the role of regression analysis in business decision-making. What are the important properties of regression coefficients? [*Osmania Univ., MBA; Delhi Univ., MBA, 2007*]

10. (a) Distinguish between correlation and regression analysis. [*Dipl in Mgt., AIMA, Osmania Univ., MBA, 2008*]

    (b) The coefficient of correlation and coefficient of determination are available as measures of association in correlation analysis. Describe the different uses of these two measures of association.

11. What are regression coefficients? State some of the important properties of regression coefficients. [*Dipl in Mgt., AIMA, Osmania Univ., MBA, 2001*]

12. What is regression? How is this concept useful to business forecasting? [*Jodhpur Univ., MBA, 2008*]

**13.** What is the difference between a prediction interval and a confidence interval in regression analysis?

**14.** Explain what is required to establish evidence of a cause-and-effect relationship between $y$ and $x$ with regression analysis.

**15.** What technique is used initially to identify the kind of regression model that may be appropriate?

**16.** (a) What are regression lines? Why is it necessary to consider two lines of regression?
(b) In case the two regression lines are identical, prove that the correlation coefficient is either + 1 or – 1. If two variables are independent, show that the two regression lines cut at right angles.

**17.** What are the purpose and meaning of the error terms in regression?

**18.** Give examples of business situations where you believe a straight line relationship exists between two variables. What would be the uses of a regression model in each of these situations?

**19.** 'The regression lines give only the best estimate of the value of quantity in question. We may assess the degree of uncertainty in the estimate by calculating a quantity known as the standard error of estimate'. Elucidate.

**20.** Explain the advantages of the least-squares procedure for fitting lines to data. Explain how the procedure works.

## Formulae Used

1. Simple linear regression model

$$y = \beta_0 + \beta_1 x + e$$

2. Simple linear regression equation based on sample data

$$y = a + bx$$

3. Regression coefficient in sample regression equation

$$b = \hat{y}$$
$$a = \overline{y} - b\overline{x}$$

4. Residual representing the difference between an observed value of dependent variable $y$ and its fitted value

$$e = y - \hat{y}$$

5. Standard error of estimate based on sample data
   - Deviations formula

$$S_{y.x} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$$

   - Computational formula

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\,\Sigma y - b\,\Sigma xy}{n-2}}$$

6. Coefficient of determination based on sample data
   - Sums of squares formula

$$r^2 = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\overline{y})^2}$$

   - Computational formula

$$r^2 = \frac{a\,\Sigma y + b\,\Sigma xy - n(\overline{y})^2}{\Sigma y^2 - n(\overline{y})^2}$$

7. Regression sum of squares

$$S_{y.x} = S_y\sqrt{1-r^2}$$

8. Interval estimate based on sample data: $\hat{y} \pm t_{df}\, S_{yx}$

## Chapter Concepts Quiz

### True or False

**1.** [T] [F] A statistical relationship between two variables does not indicate a perfect relationship.

**2.** [T] [F] A dependent variable in a regression equation is a continuous random variable.

**3.** [T] [F] The residual value is required to estimate the amount of variation in the dependent variable with respect to the fitted regression line.

**4.** [T] [F] Standard error of estimate is the conditional standard deviation of the dependent variable.

**5.** [T] [F] Standard error of estimate is a measure of scatter of the observations about the regression line.

**6.** [T] [F] If one of the regression coefficients is greater than one the other must also be greater than one.

**7.** [T] [F] The signs of the regression coefficients are always same.

**8.** [T] [F] Correlation coefficient is the geometric mean of regression coefficients.

**9.** [T] [F] If the sign of two regression coefficients is negative, then sign of the correlation coefficient is positive.

**10.** [T] [F] Correlation coefficient and regression coefficient are independent.