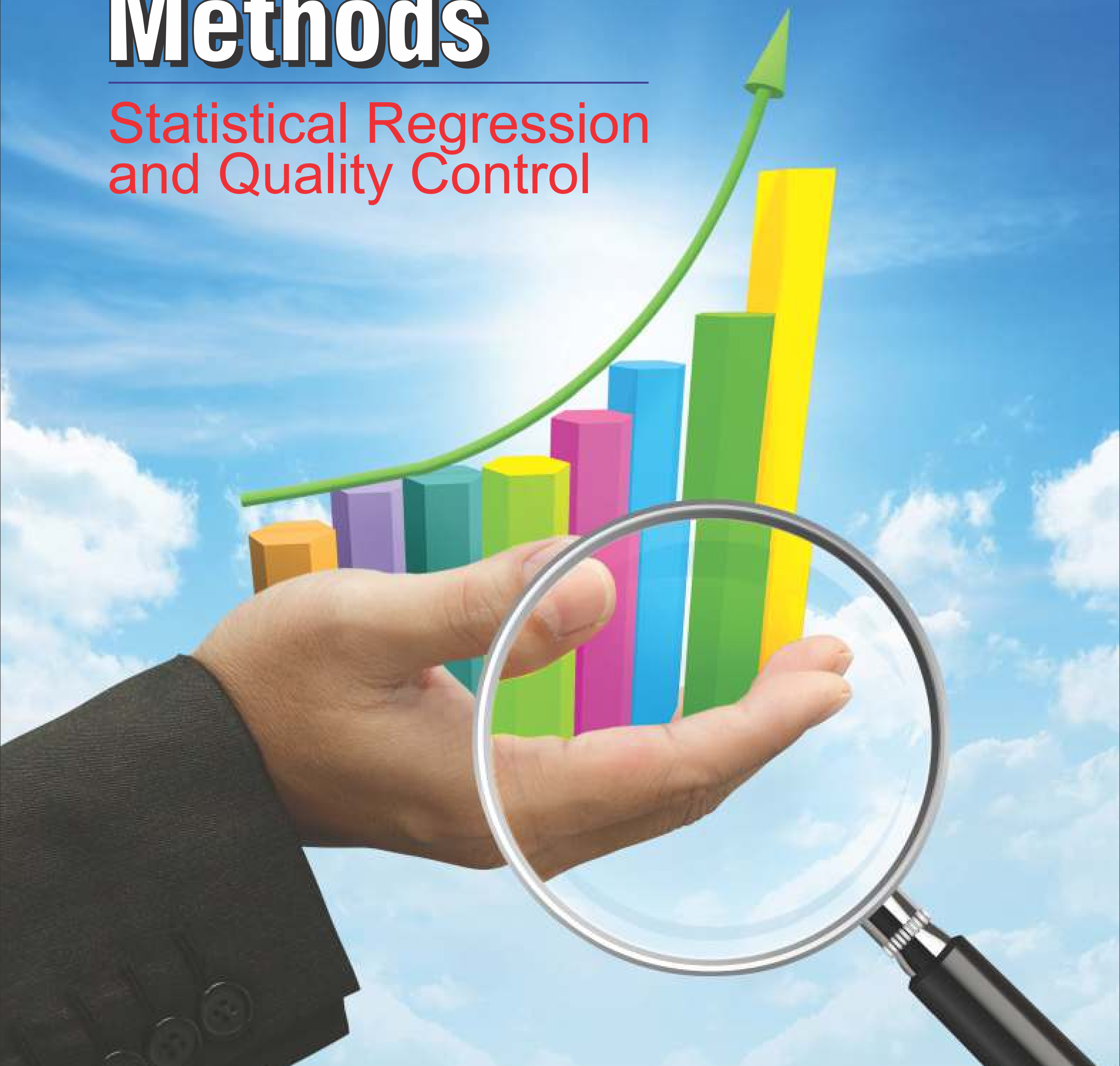
Block-III

Quantitative Methods

Statistical Regression and Quality Control



Block

III

STATISTICAL REGRESSION AND QUALITY CONTROL

Ref. No. QM SLM 102021B3

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

Our E-mail id: cwfeedback@icfaiuniversity.in

Center for Distance and Online Education (CDOE)
The ICFAI Foundation for Higher Education

(Deemed-to-be-University Under Section 3 of UGC Act, 1956)

Donthanapally, Shankarapalli Road, Hyderabad- 501203.

BLOCK III: STATISTICAL REGRESSION AND QUALITY CONTROL

Based on statistical relations between variables, the future of a business may be forecasted. In a world of uncertainty, it is never possible to forecast accurate future business returns, and the expected return may or may not materialize. The concept of probability and expected value of a random variable is used for taking decisions under conditions of uncertainty. The probability is chance of occurrence of an event. The expected value is the mean of a probability distribution. The mean is computed as the weighted average of the value that the random variable can assume. The probabilities assigned are used as weights. Thus, it is computed by summing up the random variables multiplied by their respective probabilities of occurrence. In this block, multiple regression, time series analysis and quality control are discussed.

Unit-7 Multiple Regression deals with multiple regression that consists more than two variables. It also deals with standard error, multicollinearity, multiple correlation coefficients, and partial correlation coefficients. Regression with two independent variables, standard error, estimating interval and testing the model are discussed.

Unit-8 Time Series Analysis discusses the concept of Time Series Analysis, where in the pattern of changes in the data collected over regular intervals of time is identified. The data collected can be at a periodical interval of days, weeks, months and years. After identifying the patterns, project them into future to get an estimate of the variable under consideration. The variations observed in time series can be classified as (i) the secular trend, (ii) the cyclical fluctuation, (iii) the seasonal variation, and (iv) the irregular variation.

Unit-9 Quality Control explains the relevance of quality control. It is the responsibility of the manufacturer to look into the product's quality so that it performs the way it was expected to be or as claimed by customers. This can only be achieved if the manufacturer monitors the production process at each stage and eventually all the constituent parts which go into the making of the product. Statistical process control will help to eliminate completely or minimize to the extent possible the variability or deviations from the established standards present in manufacturing process.

Unit 7

Multiple Regressions

Structure

7.1 Introduction

In the last unit, you studied the relationship between the dependent variable Y and independent variable X. But sometimes the dependent variable Y depends on more than one variable. A simple example is the idea of profit. The profit of a company would depend upon a number of revenues and costs. Hence, the simple regression equation appears to be inadequate in representing such a situation. Multiple regression equations help to form the relationship between one dependent variable and more than one independent variable. In this unit, you will learn multiple regression equations

7.2 Objectives

After going through the unit, you should be able to:

- Formulate regression with two independent variables;
- Define and calculation of standard error;
- Explain the concept of Multicollinearity; and
- Define multiple correlation coefficients and partial correlation coefficients.

7.3 Regression with Two Independent Variables

Multiple regression means estimating a single regression with more than one outcome or variable. In **multiple regression,** there is just one dependent variable, i.e. y. But, the predictor variables or parameters are multiple.

Multiple linear regression explains the relationship between one continuous dependent variable and independent variable. There can be two or more independent variables. Values of independent variables (x) are associated with dependent variables (y).

The following Exhibit 7.1 explains how multiple regression analysis can help Indian Railways to find out predictive maintenance regimes.

Exhibit 7.1: Indian Railways: Identifying Predictive Maintenance Regimes through Regression Analyses

Priya Agrawal, Deputy Chief Signal and Telecom Engineer at Indian Railways, in her article "The power of big data: Smart maintenance using data analytics" published in Global Railway Review explored how the fourth largest railway network in the world is taking a smart approach in shifting from periodic maintenance to condition-based and predictive maintenance regimes. Though there are projects in the pipeline to install a number of sensors to monitor different parameters of signalling gears installed at site, the power of data analytics can be helpful in finding asset failure trends based on past failure history, as well.

Consider, for example, the seasonal variation in failures. Assets such as track circuits, cables and some power equipment are more prone to fail during spring when there is rain, thunder and lightning, resulting in the water logging of tracks etc., which eventually leads to asset failures. Historical failure data can be analysed to run regression models to find a relationship between the level of rainfall at that time of year (week, days etc.) and the number or sites of failures, etc. Similarly, other machine learning algorithms (such as time series forecasting) can help in understanding the probability of the failure of a particular asset within a given time interval if it has undergone a specific number of cycles of operation.

Models developed through analysis can help to find a relationship between different variables and help us predict the outcome with a certain probability and degree of accuracy. Route maintenance teams can smartly plan the maintenance strategy for the route throughout the year and act accordingly. Decisions such as which stations and assets need greater attention, which maintenance personnel needs training etc., if aligned with such data-based models, can yield better results.

Adapted from https://www.globalrailwayreview.com/article/113415/big-data-smart-maintenance-data-analytics/, 4 January 2021 (Accessed on September 30th 2021)

The multiple regression equation takes the form:

$$\hat{Y} = A + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_kX_k$$

Where,

There are k independent variables $X_1, X_2, \dots X_k$.

If there are only two independent variables X_1 and X_2 the regression equation will be:

$$Y = A + B_1X_1 + B_2X_2$$

Least square method is a popular method to find the value of constant a, b_1 and b_2, where a, b_1, b_2 are estimators of A, B_1, B_2. The regression equation would then be:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

Adding over the set of given values, we get,

$$\Sigma\hat{Y} = na + b_1\ \Sigma X_1 + b_2\ \Sigma X_2 \qquad \text{... (1)}$$

The second equation is obtained by multiplying the two sides of the regression equation (1) by X_1 and then taking the summation.

$$\Sigma(X_1\hat{Y}) = a\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2 \qquad \text{... (2)}$$

Similarly, the third equation is obtained by multiplying the two sides of the regression equation (1) by X_2 and then taking the summation.

$$\Sigma(X_2\hat{Y}) = a\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2{}^2 \qquad \text{… (3)}$$

After establishing the "normal equations," we can calculate the value of a, b_1 and b_2 by using simultaneous equation methods discussed earlier.

Example 1

Modern Enterprises Ltd., manufactures and sells domestic products. The marketing manager of the company is concerned about the sales behavior of a kitchen-knife set manufactured and sold by the company. In his opinion, the number of advertisements and price per set are significant determinants of sales. He has collected the following data from the records of the company.

Sales ('000 sets sold)	Number of Advertisements	Price per set (Rs.)
30	2	120
58	5	110
67	9	135
79	12	125
14	8	140

Solution

The multiple regression equation that relates the sales with the number of advertisements and the price per set would be –

Let,

Y = Sales ('000 sets)

X_1 = Number of advertisements

X_2 = Price per set.

Then,

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

Hence, we have the normal equations:

$$\sum \hat{Y} = na + b_1 \sum X_1 + b_2 \sum X_2$$

$$\sum X_1 \hat{Y} = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1X_2$$

$$\sum X_2 \hat{Y} = a \sum X_2 + b_1 \sum X_1X_2 + b_2 \sum X_2^2$$

(figures in thousands)

Y	X_1	X_2	X_1X_2	X_1^2	X_2^2	X_1Y	X_2Y
30	2	120	240	4	14400	60	3600
58	5	110	550	25	12100	290	6380
67	9	135	1215	81	18225	603	9045
79	12	125	1500	144	15625	948	9875
14	8	140	1120	64	19600	112	1960
248	36	630	4625	318	79950	2013	30860

So, the normal equations are:

$$248 = 5a + 36b_1 + 630b_2 \quad \text{.... (i)}$$

$$2013 = 36a + 318b_1 + 4625b_2 \quad \text{... (ii)}$$

$$30860 = 630a + 4625b_1 + 79950b_2 \quad \text{... (iii)}$$

Multiplying equation (i) with 36 and equation (ii) with 5 and subtracting equation (ii) from equation (i).

We get

$$8928 = 180a + 1296b_1 + 22680b_2$$

$$\underline{-10065 = -180a \pm 1590b_1 \pm 23125b_2}$$

$$-1137 = -294\, b_1 - 445b_2$$

i.e. $1137 = 294\, b_1 + 445b_2$ ... (iv)

Multiplying equation (i) with 126 and subtracting it from equation (iii) we get –

$$\begin{array}{rcl} 30860 & = & 630a + 4625b_1 + 79950b_2 \\ -31248 & = & -630a \pm 4536b_1 \pm 79380b_2 \\ \hline -388 & = & 89b_1 + 570b_2 \end{array} \quad \ldots \text{(v)}$$

Multiplying equation (iv) with 89 and equation (v) with 294 and subtracting (v) from (iv) we get

$$\begin{array}{rcl} 101193 & = & 26166b_1 + 39605b_2 \\ -\,114072 & = & 26166b_1 + 167580b_2 \\ \hline 215265 & = & -127975b_2 \\ \text{or } b_2 & = & 215265/-127975 = -1.682 \end{array}$$

Putting the value of b_2 in equation (iv) we get

$$\begin{array}{rcl} 1137 & = & 294\, b_1 + 445\,(-1.682) \\ b_1 & = & 1885.49\,/294 = 6.413 \end{array}$$

Putting the values of b_1 and b_2 in equation (i) we get

$248 = 5a + 36\,(6.413) + 630\,(-\,1.682)$

or $248 = 5a + 230.86 - 1059.66$

$a = 1076/5 = 215.358$

$\therefore$ The multiple regression equation is:

$\hat{Y} = 215.358 + 6.413\, X_1 - 1.682X_2$

For $X_1 = 4$, and $X_2 = 126$

$\hat{Y} = 215.358 + 6.413\,(4) - 1.682\,(126) = 29.078$ (thousand sets)

i.e., 29,078 sets

$\therefore$ Sales will be equal to 29,078 sets.

7.4 Standard Error

The standard error of the estimate for a regression equation is given by:

$$S_e = \sqrt{\Sigma\,(Y - \hat{Y})^2/(n - k - 1)}$$

Where,

Y = The sample value of the dependent variable

$\hat{Y}$ = The corresponding estimate obtained by using the regression equation

n = Number of observations

k = Number of independent variables.

For the regression equation with two independent variables an alternative formula mainly utilizing data already collected for ascertaining the regression equation is:

$$S_e \quad = \quad \sqrt{\frac{\sum Y^2 - a\sum Y - b_1 \sum X_1 Y - b_2 \sum X_2 Y}{n - k - 1}}$$

Example 2

In our earlier example (Example 1),

$$S_e \quad = \quad \sqrt{\frac{15190 - 215.358 \text{ x } 248 - 6.413 \text{ x } 2013 - (-1.682) \text{ x } 30,860}{6 - 2 - 1}}$$

$$= \quad 10.66 \text{ thousands.}$$

7.4.1 Estimating Interval

Now suppose you want to predict a 95% interval of sales for four advertisements and at Rs.126 price per set. Estimated sales for this level is 29,078.

Here N < 30, so we can use 't' distribution for predicting the interval. The degree of freedom for t test is 2 = (5 – 2 – 1) because we have used three parameters a, b_1, and b_2 to estimate the 5 data point; hence, we have lost 3 degrees of freedom. For 2 degrees of freedom and 95% probability is +4.303 and –4.303 (t distribution standard table).

Requited interval will be:

$\hat{Y} + t(S_e)$ = 29,078 + (4.303 x 10.66) = Rs.29,124

$\hat{Y} - t(S_e)$ = 29,078 – (4.303 x 10.66) = Rs.29,032.

7.4.2 Testing the Model

As in simple linear regression, a, b_1, b_2, ..., b_k are unbiased estimators of A, B_1, B_2, ... B_k respectively.

As before, s_{b_i} are the estimators of the standard error of the regression coefficients. In multiple regression, these estimates are difficult to compute, and we will use computer outputs to make the necessary inferences later. Of course, as before, we will test the hypothesis.

$$H_0 : B_i = 0 \text{ and } \frac{b_i - B_i}{s_{b_i}} \sim t_{n-(k+1)} \text{ distribution.}$$

To make inference about the regression as a whole, we need to examine whether the sample implies that the Coefficient of Determination R^2 is 0 or not.

Check Your Progress - 1

1. A general multiple regression equation which has n-independent variables, will have

a. n constants

b. n + 1 constants

c. n – 1 constants

d. n dependent variables

e. No constants.

2. How many dependent variables will a multiple regression analysis have?

a. 0

b. 1

c. 2

d. More than 1

e. More than 2

3. If no. of samples are less than 30, which distribution is used to estimate interval?

a. t distribution

b. Z distribution

c. F distribution

d. Gamma distribution

e. No distribution is used.

4. In a super market the sales of various products in their retail outlet are analysed. The supermarket collected data on product category (X) and volumes of sales (Y). Which of the following is appropriate to establish the relationship between a product's sales volume, price, number of advertisements to promote it, and seasonal effect on its demand?

a. Multiple regression

b. Standard error

c. Simple regression

d. Rank correlation

e. Standard deviation

5. Based on past data, a coffee shop owner builds the regression equation. The equation is $Y = 2.5 + 0.5X$

Where,

X= Price of the coffee type

Y = Glasses sold (in 100s)

Sum $Y^2 = 68900$

Sum Y = 510

Sum X*Y = 49750

What is the standard error computed using short cut method

a. 103.38

b. 94.34

c. 23.45

d. 10.338

e. 94.68

7.5 Multicollinearity

Multicollinearity is a problematic condition that rises in the situation where two or more independent variables of a multiple regression model are highly correlated. If there is a high level of correlation between some of the independent variables, the regression coefficients become less reliable. Due to this reason, sometimes the model fits the data well even though none of the X variables has a statistically significant impact on predicting Y. This is because both X and Y will convey the same message and if we take a single variable alone it may not contribute to the variable but if both variables are taken at a time it may contribute a lot. So, the overall model fits the data well, but X variable makes a significant contribution when it is added to the last model.

7.6 Multiple Correlation Coefficients

Multiple correlation coefficients between two independent variables X_1, X_2 and dependent variable Y can be measured using the following formula:

$$R_{Y.X_1X_2} = \sqrt{1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

This can be also expressed as:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

Where,

r_{12} = The correlation coefficient between Y and X_1

r_{23} = The correlation coefficient between X_1 and X_2

r_{13} = The correlation coefficient between Y and X_2

$R_{1.23}$ = Will take values between 0 and 1 i.e., $0 \le R_{1.23} \le 1$.

Coefficient of multiple determinations can be represented as $R^2_{Y.X_1X_2}$, which is the square of the above coefficient of multiple correlation. $R^2_{Y.X_1X_2}$ also takes values in the range 0 to 1.

Example 3

Consider the following matrix of correlation coefficients between 3 variables Y, X_1 and X_2.

		(1)	(2)	(3)
		Y	X_1	X_2
(1)	Y	1	0.9	0.6
(2)	X_1	0.6	1	0.9
(3)	X_2	0.7	0.8	1

Find the multiple correlation coefficients between Y and both X_1 and X_2.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.9^2 + 0.6^2 - [2 \text{ x } 0.9 \text{ x } 0.9 \text{ x } 0.6]}{1 - 0.9^2}} = \sqrt{1.042} = 1.02$$

7.7 Partial Correlation Coefficients

Partial correlation coefficient measure the correlation between each X and Y after the effects of the other variables have been removed. In other words, partial correlation coefficients measure how much of the variation in the dependent variable is explained by one independent variable if all the other independent variables are kept constant. The general formula for calculating partial correlation between X and Y after removing effects of other variable is as follows:

$$R_{12.3} = \frac{r_{12} - r_{13} . r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

The regression equation is: $Y = a + b_1 X_1 + b_2 X_2$. $R_{12.3}$ is partial correlation coefficient of Y on X_1 if X_2 is kept constant.

Example 4

The following is the matrix of correlation coefficient among Y, X_1 and X_2.

		(1)	(2)	(3)
		Y	X_1	X_2
(1)	Y	1	0.8	0.6
(2)	X_1	0.9	1	0.9
(3)	X_2	0.7	0.5	1

The partial correlation coefficient between Y and X_1 using our simplified subscripts is given by:

$$R_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.8-(0.6 \text{ x } 0.9)}{\sqrt{(1-0.6^2)(1-0.9^2)}} = 0.745$$

$$\therefore R_{12.3} = 0.745$$

Hence, 74.5% of the variation in Y is explained by X_1 when X_2 is kept constant.

Note that here $R_{12.3}$ is not equal to r_{12}. This is because r_{13} and r_{23} are not equal to zero.

<u>Check Your Progress - 2</u>

6. Multicollinearity is the
   a. Existence of correlation among independent variables
   b. Absence of correlation among independent variables
   c. Existence of correlation among dependent variables
   d. Absence of correlation among dependent variables
   e. Predictive power of the constants.

7. If it is given that the correlation coefficients between X_1 and X_2 is 0.36, between X_2 and X_3 is 0.64, between X_3 and X_1 is 0.49, the multiple correlation coefficient between X_1 and both X_2 and X_3 will be given by
   a. 0.36
   b. 0.49
   c. 0.64
   d. 0.73
   e. None of the above.

8. A multiple regression plane passes through the points (2, 3, –1), (1, 5, 3) and (1, –1, 1).

   What is the equation of the multiple regression plane?

   a. $Y = \frac{13}{3} - \frac{10}{3} X_1 + \frac{1}{3} X_2$.
   b. $Y = \frac{14}{3} - \frac{11}{3} X_1 + \frac{1}{3} X_2$.
   c. $Y = \frac{14}{3} + \frac{10}{3} X_1 + \frac{2}{3} X_2$.
   d. $Y = \frac{17}{3} - \frac{13}{3} X_1 + \frac{1}{3} X_2$.
   e. $Y = \frac{14}{3} - \frac{10}{3} X_1 + \frac{1}{3} X_2$.

9. A multiple regression relationship contains two independent variables. The standard error of estimate is 4.8. Error sum of squares = 576.

   What is the number of data points?

   a. 24.
   b. 25.
   c. 26.
   d. 27.
   e. 28.

10. Which of the following is true in partial correlation coefficient?

    a. One dependent variable is kept constant
    b. All Independent variables are kept constant
    c. Independent variables are kept constant except one independent variable
    d. All dependent variables are kept constant
    e. No variable is kept constant.

7.8 Summary

- A multiple regression equation has only one dependent variable and multiple independent variables. The value of the dependent variable is derived from the values of the independent variables.
- All the independent variables need not have the same coefficients.
- Coefficient of multiple correlations describes the correlation between the dependent variable and all the independent variables.
- Partial correlation coefficient is a measure of correlation between two variables after the effect of the other variables has been removed.

7.9 Glossary

Multicollinearity: Multicollinearity refers to the correlation among two or more variables in linear regression model.

Multiple Correlation Coefficients: Coefficient of correlation which will measure the correlation between dependent variable Y and more than one independent variable (X_1, X_2….).

Multiple Regressions: A statistical process in which more than two variables are used to predict another variable.

7.10 Suggested Readings/Reference Material

1. Gupta, S. P. Statistical Methods. 46th Revised ed. New Delhi: Sultan Chand & Sons. 2021.
2. I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management. Pearson Education; Eighth edition, 2017.

3. Gerald Keller. Statistics for Management and Economics. Cengage, 2017.
4. Arora, P. N., and Arora, S. CA Foundation Course Statistics. 6th ed. S Chand Publishing, 2018.
5. Mario F Triola. Elementary Statistics. 13th ed., 2018.
6. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran. Statistics for Business and Economics. 13th Edition, Cengage Learning India Pvt. Ltd., 2019.
7. S D Sharma. Operations Research. Kedar Nath Ram Nath, 2018.
8. Hamdy A. Taha. Operations Research: An Introduction. 10th ed., Pearson, 2016.
9. Malhotra, N. (2012), Marketing Research: An Applied Orientation, 7th ed., Pearson, 2019.
10. Cooper, D.R. and Schindler, P.S. and J. K. Sharma (2018), Business Research Methods, 12th edition, McGraw-Hill Education.

7.11 Self-Assessment Questions

1. Describe the process of solving regression equation which has more than two independent variables.

2. What is standard error? How can we calculate standard error in multiple correlation of dependent and two independent variables?

3. Write short notes on Coefficient of determination.

4. Define the term “Multicollinearity”.

5. What do mean by Partial and Total Correlation?

7.12 Answers to Check Your Progress Questions

1. (b) n+1 constants

2. (b) 1. In multiple regression, there is just one dependent variable.

3. (a) t distribution

4. (a) Multiple regression. It establishes the relationship among all four parameters of a product. Sales volume is the dependent variable and other three are independent variables.

5. (a) 103.38

Sqrt(sum Y square-a * sumY -b*sum X*Y)/n-2

6. (a) Existence of correlation among independent variables

7. (b) 0.49

8. (e) $Y = \frac{14}{3} - \frac{10}{3}X_1 + \frac{1}{3}X_2$

9. (e) 28

(sqrt (576/(n-k-1)) = 4.8. n-k-1 = 25

10. (c) Independent variables are kept constant except one independent variable

Partial correlation coefficients measure how much of the variation in the dependent variable is explained by one independent variable if all the other independent variables are kept constant.

Unit 8

Time Series Analysis

Structure

8.1 Introduction

In the previous unit, you studied the statistical relation between more than two variables where one variable is an independent variable, while two or more are dependent variables. Here, the statistical relation between time and variables are studied. There are various quantities in nature which fluctuate with time. Examples are, sales, price of stocks, consumption, etc. Earlier, it was assumed that such fluctuations are a consequence of random and unpredictable events. Now most statisticians, and analysts show that some of these cases follow a particular trend and hence are predictable in the short-term and are amenable to simple modeling. The basic objective behind predicting future is to frame the future objective of the firm to meet the forthcoming challenges.

A time series is a sequence of observations which are ordered in time; and the technique used for forecasting future events is called time series forecasting. Time series models are especially suitable for evaluating short-term effects of time-varying exposures. In time-series studies, a single population is assessed with reference to its change over time.

8.2 Objectives

After going through the unit, you should be able to:

- Define Time series analysis;
- Explain Secular trend;
- Identify Cyclical variation;

- Analyze Seasonal variation;
- Demonstrate Moving average; and
- Contrast Irregular variation.

8.3 Uses of Time Series

Time series is the sequential arrangement of data. In the time series, time is a variable which can relate the entire situation to suitable location points. Time can be hours, days, months or years.

A time series depicts the relationship between two variables. Time is one of those variables and the second is any quantitative variable. It is not necessary that the relationship always shows increment in the change of the variable with reference to time. The relation is not always decreasing too. Here time is an independent variable.

The most important use of studying time series is that it helps us in forecasting the future behavior of the variable based on past experience or past data which is called as historical data. It compares the actual performance which is the current performance with the expected performance or the predicted performance of the business and helps in planning.

The purpose of time series is to study the past behavior of the phenomenon or the variable under consideration. Then the changes in the values of different variables at different times or places, etc. can be related.

The numerous motives or the forces which influence the values of an observation in a time series are the components of a time series. The four categories of the components of time series are

- Trend (T)
- Seasonal Variations (S)
- Cyclic Variations (C)
- Random or Irregular movements (I)

Seasonal and Cyclic Variations are the periodic changes or short-term variations.

These components may be combined in different ways. It is usually a multiplicative or an additive model, i.e., $y_t = T \times C \times S \times I$ or $y_t = T + C + S + I$. To correct for the trend in the first case one divides the first expression by the trend (T). In the second case it is subtracted.

Additive Model for Time Series Analysis

If y_t is the time series value at time t. T_t, S_t, C_t, and R_t are the trend value, seasonal, cyclic and random fluctuations at time t respectively. According to the Additive Model, a time series can be expressed as

$y_t = T_t + S_t + C_t + R_t$.

This model assumes that all four components of the time series act independently of each other.

Multiplicative Model for Time Series Analysis

The multiplicative model assumes that the various components in a time series operate proportionately to each other. According to this model

$y_t = T_t \times S_t \times C_t \times R_t$

Mixed models

Different assumptions lead to different combinations of additive and multiplicative models as

$y_t = T_t + S_t + C_t \times R_t$

The time series analysis can also be done using the model yt = $T_t + S_t \times C_t \times R_t$ or $y_t = T_t \times C_t + S_t \times R_t$ etc.

Trend component: The trend is the long-term pattern of a time series. In this case the time factor is linearly related with the dependent variable. A trend can be positive or negative depending on whether the time series displays an increasing long-term pattern or a decreasing long-term pattern depending on the coefficient of time. If a time series does not show an increasing or decreasing pattern then the series is stationary in the mean. If the time series values are plotted on a graph in accordance with time t, the pattern of the data clustering shows the type of trend. If the set of data cluster is more or less around a straight line, then the trend is linear otherwise it is non-linear (Curvilinear).

Seasonal component: Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or during the same quarter every year. For instance, retail sales peak during the month of December.

These are the recurring forces which operate in a regular and periodic manner over a duration of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations can be natural variations or man-made variations. The various seasons or climatic conditions play an important role in seasonal variations. For example, production of crops depends on seasons, the sale of umbrella and raincoats is high in the rainy season, and the sale of electric fans and A.C. increase in summer.

The effect of man-made variations is seen because of some festivals, customs, habits, fashions, and some occasions like marriage.. They repeat themselves every year. An increase in a season should not be taken as an indicator of better business conditions.

Cyclical component: Any pattern showing an up and down movement around a given trend is identified as a cyclical pattern. The duration of a cycle depends on the type of business or industry being analyzed.

The variations in a time series which operate themselves over a period of more than one year are the cyclic variations. This oscillatory movement has a period of fluctuation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.

It is a four-phase cycle containing the stages of prosperity, recession, depression, and recovery. The cyclic variation may be regular but not periodic. The upswings and the downswings in business depend upon the combined effect of the economic factors and the collaboration between them.

Irregular component: This component is unpredictable. Every time series has some unpredictable component that makes it a random variable. In prediction, the objective is to "model" all the components to the point that the only component that remains unexplained is the random component.

There are some other factors which causes the variation in the time series data. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

The objective of time series analysis is to:

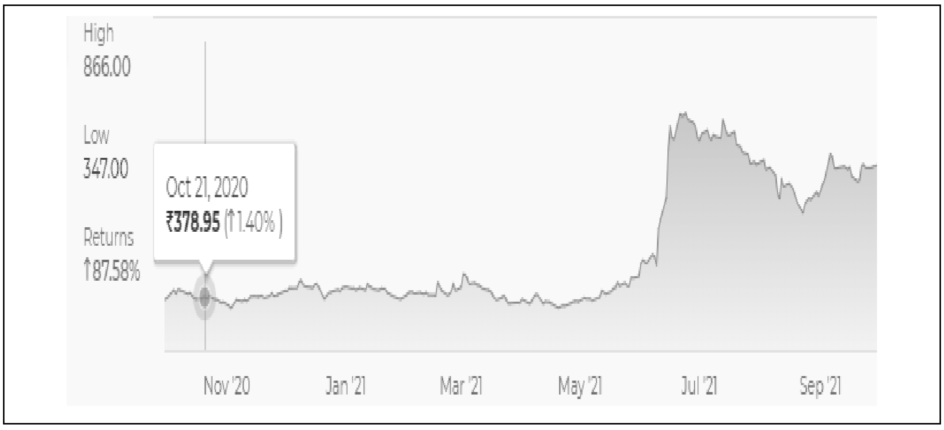
- Find the particular method to use in forecasting the future.
- Put that method to use to forecast the future.

8.4 Secular Trend

A secular trend is a long-term trend having a relatively smooth pattern. The time period normally considered is one year and above. For example, the share price of Reliance Industries infra shows a long-term trend as it grows steadily from Rs.378 in October 2020 to Rs.801 in June 2021.

Figure 8.1: Reliance Industries Infra Ltd.

(Share price 2020-2021)



Adapted from https://economictimes.indiatimes.com/reliance-industrial-infrastructure-ltd/stocks/companyid-12097.cms

The basic advantages of studying a secular trend are:

- It helps to project future movement of data.
- It helps to examine the performance of policy implemented.

The secular trend is not always linear and may be curvilinear. The linear trend line or curvilinear trend can be determined by least square method. Whether it is a straight line or curvilinear can be judged by studying the scatter diagram of the data.

8.4.1 Fitting the Linear Trend by the Least Squares Method

The least squares method is used mostly for data fitting. The best fit result minimizes the sum of squared errors or residuals which are said to be the differences between the observed or experimental value and corresponding fitted value given in the model. There are two basic kinds of the least squares' methods – ordinary or linear least squares and nonlinear least squares.

Linear Trend by the ordinary Least Squares Method is as follows:

The equation of a straight line is given by

$$Y = a + bX.$$

Where,

X	=	Independent variable
Y	=	Dependent variable
a	=	Y-intercept
b	=	Slope of the trend line.

Slope of the best fitting regression line can be calculated by using the formula

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}, \quad a = \overline{Y} - b\overline{X}.$$

Process of Solving a Time Series Problem

Step I

Convert traditional data into coding data or translated time.

Normally, an independent variable in the time series data is used in terms of month, year or week. It is a common practice to convert the traditional data into another form that can simplify the calculation. In order to make it easier, we take the mean of the X data points and subtract it from actual X values and represent that column as "x".

Step II

If mean of X is a whole number, write $x = X - \overline{X}$

If mean of X is in fraction like 2004.5 or 2004.25 etc., then deviation will also be a fraction. To convert the fraction into a whole number multiply the deviation with a suitable number, say $x = Z(X - \overline{X})$.

Step III

In this case mean will be zero so the value of a and b can be found using the equation given below:

$$b = \frac{\sum xY}{\sum x^2} \text{ and } a = \overline{Y}$$

Step IV

Now the regression equation will be $Y = a + bx$, where $x = X - \overline{X}$.

Example 1

The Traffic Police Department in a city is studying the number of fatal accidents that have taken place over the period of six years. The following data were collected from their records:

Years (X)	Number of Accidents (Y)	$Z = X - \overline{x}$	x = 2.Z	(2) x (4)	x^2
(1)	(2)	(3)	(4)	(5)	(6)
2015	200	–2.5	–5	–1,000	25
2016	185	–1.5	–3	–555	9
2017	180	– 0.5	–1	–180	1
2018	190	0.5	1	190	1
2019	175	1.5	3	525	9
2020	185	2.5	5	925	25
Total 12,105	1,115			–95	70

$$\overline{X} = 12105/6 = 2017.5$$

We have, $b = \frac{\sum xY}{\sum x^2} = \frac{-95}{70} = -1.357$ and $a = \overline{Y} = \frac{1,115}{6} = 185.83$

Therefore, the regression equation describing the secular trend is given by

$$\hat{Y} = 185.83 - 1.357x$$

To estimate the number of accidents for the year 2021 you have to convert the time data as follows:

$$z = 2021 - 2017.5 = 3.5;\ x = 2z = 2 \text{ x } 3.5 = 7$$

Substituting this in the estimating equation, we have

$$\hat{Y} = 185.83 - 1.357 \text{ x } 7 = 176.33$$

So, the number of accidents for the year 2021 would be 176.

8.3.1 Fitting the Curvilinear Trend by Least Square Method

The curvilinear trend equation is $Y = a + bX + cX^2$.

The corresponding estimating equation is given by $\hat{Y} = a + bx + cx^2$

Where,

"x" represents the coded time variable.

As we have seen in the case of a curvilinear trend, the values of the constants a, b and c for the curvilinear of the best fit are given by solving these equations.

$$\sum Y = an + c\sum x^2$$

$$\sum x^2 Y = a\sum x^2 + c\sum x^4 \text{ and } b = \frac{\sum xY}{\sum x^2}.$$

Example 2

The number of color computers sold during the years 2016 to 2020 are as shown below.

Year (X)	Number of sets sold (in thousands) (Y)
2016	30
2017	45
2018	60
2019	80
2020	100

For this data, we fit a curvilinear trend. The calculations pertaining to this example are shown below:

	Years (X)	Number of sets sold (Y) (in thousands)	$x = X - \bar{X}$	x^2	x^4	xY	x^2Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2016	30	–2	4	16	–60	120
	2017	45	–1	1	1	–45	45
	2018	60	0	0	0	0	0
	2019	80	1	1	1	80	80
	2020	100	2	4	16	200	400
Total	10,090	315		10	34	175	645

$\bar{x}$ = 10,090/5= 2018

We have, $\sum Y = an + c\sum x^2$

$\sum x^2 Y = a\sum x^2 + c\sum x^4$ and $b = \dfrac{\sum xY}{\sum x^2}$

Substituting the values in these equations, we have

$315 = 5a + 10c$ ... (i)

$645 = 10a + 34c$ ... (ii)

$b = \dfrac{175}{10} = 17.5$ ... (iii)

Multiplying (i) by 2, we have $630 = 10a + 20c$ … (iv)

We solve (iv) and (ii) $630 = 10a + 20c$

$645 = 10a + 34c$

On subtracting, we have $-15 = -14c$

$$c = \frac{15}{14} = 1.07$$

Substituting the value of c = 1.07 in (i), we have

$308 = 5a + 10(1.07)$

$308 = 5a + 10.7$

$308 - 10.7 = 5a$

i.e., $5a = 297.3$

$a = 297.3/5 = 59.46$

On substituting the values of a, b and c in the equation of the parabola, we have

$$\hat{Y} = 59.46 + 17.5x + 1.07x^2$$

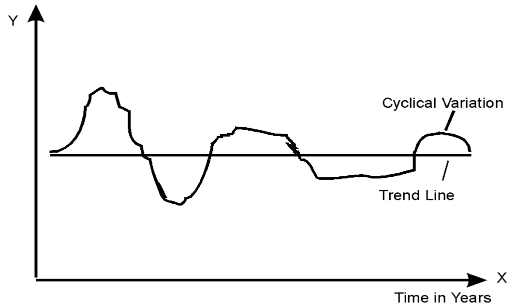
Limitation

Students should remember that these equations cannot be used across any range of time for forecasting, because the life cycle of business keeps changing due to various reasons like advancement in economy, technology, etc.

8.5 Cyclical Variation

Cyclical variation is also called the long-term pattern that is normally apparent over a number of years, resulting in a cyclical effect with a duration of more than one year. The best example of cyclical variation is the business cycle that records the period of economic recession and inflation. The cyclical variation is shown in the Figure 8.2.

Figure 8.2: Comparison of Cyclical Variation and Trend line



8.5.1 Residual Method Comparison of Cyclical Variation and Trend Line

As discussed earlier, we can describe the secular trend using the trend line. We can segregate the other variations like cyclical variation and irregular variation from the trend line. Here, we have assumed that the major percentage component is unexplained by the trend line. Residual method segregates the cyclical and irregular components using two methods:

(i) Percent of Trend Measure, and (ii) Relative Cyclical Residual Measure.

PERCENT OF TREND MEASURE

Mathematically, we can express it as $((Y/\hat{Y}) \times 100)$

Where, Y is the actual dependable value, and $\hat{Y}$ is corresponding estimated values.

We express the cyclical variation component as a percentage.

Relative Cyclical Residual Measure

Relative cyclical residual measure is the ratio of the difference between Y and the corresponding $\hat{Y}$ values. Mathematically, it is expressed as: $\frac{Y - \hat{Y}}{\hat{Y}} \times 100$

Relative cyclical residual measure can also be computed if 100 is subtracted from the values of percentage trend.

Example 3

The number of cartons of cereal sold by a departmental store is shown below.

Years (X)	Number of Cartons Sold (Y)
2015	8
2016	15
2017	14
2018	16
2019	30
2020	30

a. Find out the linear trend equation from the above data.

b. Identify the cyclical variation using the percent of trend measure.

Calculation of Percent of Trend

	X	Y	$Z = X - \overline{X}$	x = 2Z	x^2	xY
	(1)	(2)	(3)	(4)	(5)	(6)
	2015	8	–2.5	–5	25	–40
	2016	15	–1.5	–3	9	–45
	2017	14	–0.5	–1	1	–14
	2018	16	0.5	1	1	16
	2019	30	1.5	3	9	90
	2020	30	2.5	5	25	150
Total	12,105	113			70	157

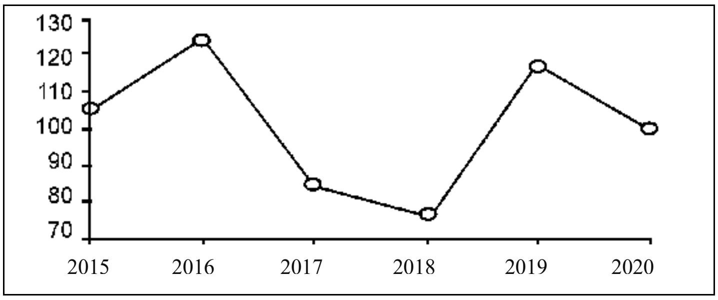
$$\overline{X} = \frac{\sum X}{n} = 12105/6 = 2{,}017.5;\ \overline{Y} = \frac{\sum Y}{n} = \frac{113}{6} = 18.83$$

$$a = 18.83,\ b = \frac{\sum xY}{\sum x^2} = \frac{157}{70} = 2.243$$

So, linear trend equation is Y = 18.83 + 2.243 x

X	X	$\hat{Y}$	Y	$(Y - \hat{Y})^2$	Percent trend $((Y/\hat{Y})x\ 100)$	Relative Cyclical Residual $\frac{Y - \hat{Y}}{\hat{Y}} x100$
2015	–5	7.615	8	0.148	105.06	5.05
2016	–3	12.101	15	8.404	123.96	23.95
2017	–1	16.587	14	6.692	84.40	15.59
2018	1	21.073	16	25.735	75.93	24.07
2019	3	25.559	30	19.722	117.38	17.36
2020	5	30.045	30	0.002	99.85	0.148
Total				60.703		

The graph below is given for providing clarity using the percent of trend values shown in the table.



Check Your Progress - 1

1. The level of the working capital needed by a small firm during the six years is shown below.

Year	1	2	3	4	5	6
Working Capital (in Rs. lakh)	5.10	5.85	6.35	6.65	7.10	7.35

For this data, which of the following represent at a linear trend?

a. $\hat{Y} = 6.98 + 0.2186x$.

b. $\hat{Y} = 6.4 + 0.2186x$.

c. $\hat{Y} = 6.4 + 0.2925x$.

d. $\hat{Y} = 6.25 + 0.2225x$.

e. $\hat{Y} = 6.55 + 0.2045x$.

2. Based on above question, estimate the level of working capital required for the next two years.

a. Rs.8.367 lakh and Rs.8.58 lakh. respectively.

b. Rs.8.367 lakh and Rs.8.88 lakh. respectively.

c. Rs.8.25 lakh and Rs.8.75 lakh. respectively.

d. Rs.8.45 lakh and Rs.8.88 lakh. respectively.

e. Rs.8.102 lakh and Rs.8.40lakh. respectively.

3. The following data pertains to the estimated demand and actual demand for a product for the last 8 periods.

Period	Estimated Demand ('000 units)	Actual Demand ('000 units)
1	714	750
2	546	526
3	568	652
4	739	667
5	577	682
6	662	859
7	788	1022
8	1097	981

You are required to identify the cyclical variation in demand using percent of trend measure and state in which period fluctuation is maximum.

a. 3

b. 4

c. 5

d. 6

e. 7

4. A retail store had collected data on number of cartons of a soft drink sold in a month and fit a linear trend equation. The actual data value of cartons sold was 28 in February month and the computed value from linear trend equation is 26.54. What is the relative cyclical residual value?

   a. 5.5
   b. 5.21
   c. 105.5
   d. 94.78
   e. 92.78

5. An electronics sales shop has compiled data of sales of washing machines of different types for over 5 years and decided to fit a curvilinear trend equation for the data. The fit equation is

   $Y = 12.78 +\mathbf{?}x + 3.5x^2$. If the computed value for the third year is 57.84, what is the constant at **?** Place?

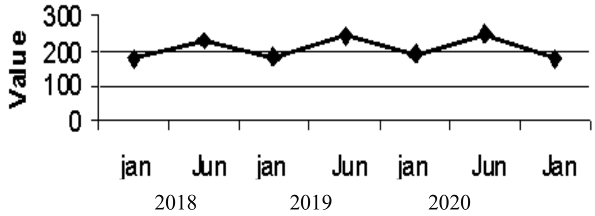
   a. 4.25
   b. 4.52
   c. 4.12
   d. 4.25
   e. 4.53

8.6 Seasonal Variation

When a repetitive pattern is observed over some time horizon, the series is said to have seasonal variation. Seasonal variation may occur within a year or month, week or day. The best example of seasonal variation is food grain price which is low just after the harvest and increases thereafter.

The seasonal variation is depicted in the figure below. We can observe the uniformity in the pattern during every second quarter of the year.

Figure 8.3: Sale Price of Food grains



Seasonal analysis is useful because:

i. It is useful to various categories in policy decision and plan formulation.

ii. It is useful to project the past patterns into future.

8.7 Moving Averages

For better forecast in seasonal variations, it is important to determine the component that actually exists in a time series data. However, for random variation, it is often difficult to identify the proper component. Moving average is one of the methods used to smoothen the random data. This can be especially helpful in volatile markets.

Moving average is a series of arithmetic means over time. The process of calculating moving average is discussed below:

Type I: When Period is Odd

When the period number is odd, the successive value of the moving average is placed simply against the middle value of the concerned group item.

For example,

Consider the following set of data: y_1, y_2, y_3 , y_4 , y_5 for 3-Year Moving Average

First average = $\dfrac{y_1 + y_2 + y_3}{3}$; Second average = $\dfrac{y_2 + y_3 + y_4}{3}$

You can note that in each subsequent calculation, we exclude the first number and include the number which comes immediately after the last number of the first group.

Type II: When Period is Even

When the given data set is odd we do not have to center them further because they are already at center; but when even number of data points are present we have to center the moving average. To calculate moving average and the component of seasonal variation in case of even data, follow these steps.

Step 1: Compute the moving total by considering the values in the four quarters of all the years.

Step 2: Calculate the average of the moving totals calculated in step 1.

Step 3: Since the number of data points is even, we need to center the moving averages. Take the average of first two moving averages and associate it with the third data point.

Step 4: Calculate the percentage of the actual data to the centered Moving Average Values (MAV).

Example 4

Given below is the data, regarding the amount of the food grains exported during 2004-07 on a quarterly basis. For this data determine the seasonal component.

Years	Quarter I	Quarter II	Quarter III	Quarter IV
2017	5	10	10	25
2018	15	10	20	30
2019	30	35	40	40
2020	20	25	50	30

Year	Quarter	Amount Exported	Four Quarter Moving Total	Four Quarter Moving Average	Centered Moving Average	Percentage of Actual to MAV
	(1)	(2)	(3)	(4)	(5)	(6)
2017	I	5	–			
	II	10				
			50	12.50		
	III	10			13.750	137.500
			60	15.00		
	IV	25			15.000	60.000
			60	15.00		
2018	I	15			16.250	108.330
			70	17.50		
	II	10			18.125	181.250
			75	18.75		
	III	20			20.625	103.125
			90	22.50		
	IV	30			24.125	99.160
			115	25.75		
2019	I	30			29.750	80.416
			135	33.75		
	II	35			35.000	100.000
			145	36.25		
	III	40			35.000	87.500
			135	33.75		
	IV	40			32.500	81.250
			125	31.25		
2020	I	20			32.500	162.500
			135	33.75		
	II	25			32.500	130.000

Year	Quarter	Amount Exported	Four Quarter Moving Total	Four Quarter Moving Average	Centered Moving Average	Percentage of Actual to MAV
	(1)	(2)	(3)	(4)	(5)	(6)
			125	31.25		
	III	50				
			–			
	IV	30				

Calculation of Modified Mean

For calculating the modified mean for each quarter, first we have to arrange data quarter-wise and then cancel the highest and lowest data in each quarter as shown in table below. The canceling out of the highest and the lowest values reduces the effect of extreme cyclical and irregular variations and further, the averaging smoothens the series for the same. The resultant values indicate the seasonal component.

Then the mean of the remaining values gives us the modified mean for that quarter respectively. In our example, after the cancelation of the highest and the lowest values, we are left with a single value. Therefore, this happens to be the mean value.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2017	–	–	~~137.5~~	~~60.00~~
2018	108.33	~~181.25~~	103.125	~~99.16~~
2019	~~80.416~~	~~100.00~~	~~87.50~~	81.25
2020	~~162.5~~	130.00	–	–

Modified Means

Quarter I	:	108.33/1	=	108.330
Quarter II	:	130.00/1	=	130.000
Quarter III	:	103.125/1	=	103.125
Quarter IV	:	81.25/1	=	81.250
Total			=	422.700

Calculation of Adjusting Factor

It is the ratio of 400 and the modified mean calculated above.

Adjusting factor = 400/422.70 = 0.946

Calculation of Seasonal Indices

Quarter	Unadjusted Means	Adjusting Constant	Seasonal Index
(i)	(ii)	(iii)	(iv) = (ii) x (iii)
I	108.330	0.946	102.48
II	130.000	0.946	122.98
III	103.125	0.946	97.56
IV	81.250	0.946	76.86
Total			400.00

8.7.1 Deseasonalizing a Time Series

The ratio to average method allows us to identify the component of the seasonal variation in time series data and the indices themselves help us to nullify the effects of seasonality on the time series. The use of indices to nullify the seasonal effects in the common parlance is referred to as deseasonalizing time series. It involves dividing the original data points with the relevant seasonal index expressed as a percentage.

In our example, the deseasonalization process is carried out as follows:

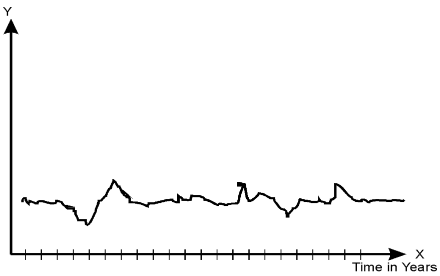
Year	Quarter	Actual Data	Seasonal Index/100	Deseasonalized Data
(1)	(2)	(3)	(4)	(5) = (3)/(4)
2017	I	1	112.73/100	0.887
2018	II	2	98.41/100	2.032
2019	III	2	118.10/100	1.693
2020	IV	1	70.77/100	1.413

The removal of the seasonal component helps us to analyze the components of secular, cyclical and irregular variations. The secular trend then obtained can be utilized for projecting the trend into the future.

8.8 Irregular Variation

Irregular variation is caused by irregular and unpredictable changes in a time series that are not caused by any other component. Because irregular variation is random in nature, it is called as random variation. For example, sometimes the share price of an unpredictable stock increases or decreases due to sudden good news or bad news. Due to random nature it is difficult to form a suitable trend.

Figure 8.4: Diagrammatic Presentation of Irregular Variation



Check Your Progress - 2

6. In the third quarter of 2019 it was found that the amount of rice exported by a firm was Rs.3 crore. The centered moving average for that period was found to be 2.76. The value of the seasonal component, if expressed in terms of percentage, will be

   a. 108.69

   b. 102.67

   c. 95.84

   d. 91.23

   e. 87.65.

7. The adjusting factor for a particular year, when four quarters were considered was 0.9234. The sum of modified means in this case is

   a. 433.00
   b. 433.18
   c. 433.21
   d. 433.25
   e. 433.30.

8. An agricultural products company had collected data of sale of one of its manures over five years (2014-2019). The table presents the sale values in ('00). Using 3 year moving average compute the third average.

Year (X)	Sale ('00)
2014-15	24
2015-16	28
2016-17	27
2017-18	32
2018-19	34

   a. 27

   b. 29

   c. 31

   d. 30

   e. 22

9. A pharmaceutical sales company has collected sales data for over 4 quarters and arrived at these modified means.

   Q1-98.5

   Q2-101.4

   Q3-103.8

   Q4-108.3

What is the seasonal index for the third quarter?

a. 95.62

b. 105.13

c. 98.43

d. 100.77

e. 99.77

10. What is referred to as deseasonalizing time series?

a. Nullifying effects

b. Multiplying effects

c. Adding up effects

d. No. role of effects

e. Subtracting an effect from other

8.9 Summary

- Time Series analysis is used to detect pattern of change in statistical information over regular interval of time. Statisticians have classified the four types of trends as – secular trend, cyclical fluctuation, seasonal variation and irregular variation. Secular trend represents the long-term direction of a time series.
- Cyclical variation also describes the long-term pattern that is normally apparent over a number of years. Normally, the duration of cyclical variation is more than one year. The series is said to have seasonal variation when a repetitive pattern is observed over some time horizon. Moving average is one of the methods that is used to smoothen the random data of time series.

8.10 Glossary

Cyclical Fluctuation: A type of variation in a time series, in which the value of the variable fluctuates above and below a secular trend line.

Irregular Variation: A condition in a time series in which the value of a variable is completely unpredictable.

Seasonal Variation: Patterns of change in a time series within a year; patterns that tend to be repeated from year to year.

Secular Trend: A type of variation in a time series, the value of the variable tending to increase or decrease over a long period of time.

Time Series: Information accumulated at regular intervals and the statistical methods used for determining patterns in such data.

8.11 Suggested Readings/Reference Material

1. Gupta, S. P. Statistical Methods. 46th Revised ed. New Delhi: Sultan Chand & Sons. 2021.
2. I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management. Pearson Education; Eighth edition, 2017.
3. Gerald Keller. Statistics for Management and Economics. Cengage, 2017.
4. Arora, P. N., and Arora, S. CA Foundation Course Statistics. 6th ed. S Chand Publishing, 2018.
5. Mario F Triola. Elementary Statistics. 13th ed., 2018.
6. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran. Statistics for Business and Economics. 13th Edition, Cengage Learning India Pvt. Ltd., 2019.
7. S D Sharma. Operations Research. Kedar Nath Ram Nath, 2018.
8. Hamdy A. Taha. Operations Research: An Introduction. 10th ed., Pearson, 2016.
9. Malhotra, N. (2012), Marketing Research: An Applied Orientation, 7th ed., Pearson, 2019.
10. Cooper, D.R. and Schindler, P.S. and J. K. Sharma (2018), Business Research Methods, 12th edition, McGraw-Hill Education.

8.12 Self-Assessment Questions

1. What is moving average? What are its uses in time series?

2. Distinguish between the seasonal component and trend component in time series.

3. What do you mean by cyclical fluctuation in time series? Give an example.

4. Explain the use of percentage of trend method in time series analysis.

5. What measures are used for seasonal indices in time series analysis?

8.13 Answers to Check Your Progress Questions

1. (b) $\hat{Y} = 6.4 + 0.2186x$

2. (a) Rs.10.00 and Rs.10.34 respectively

3. (d) 6

(AD/ED *100 highest)

4. (a) 5.5

(28-26.54/26.54)*100

5. (b) 4.52

(Substitute 3 for x and get? value)

6. (a) 108.69

7. (b) 433.16

(400/.9234)

8. (c) 31

(27+32+34/3)

9. (d) 100.77

(103.8*400/412)

10. (a) The use of indices to nullify the seasonal effects in the common parlance is referred to as deseasonalizing time series.

Unit 9
Quality Control

Structure

9.1 Introduction

In the previous unit, we learnt an overview of time series analysis. In this unit, we try to learn about controlling the quality of a product. When we want to purchase a product, we often talk about quality. This aspect becomes an important decision-making component at the time of buying a product. From the business point of view, quality conformance means meeting the needs of the customer. In these times of globalization of the world of today, market competition has become stiffer. Only those companies that produce consistently a product of high quality can survive. In these circumstances, the manufacturer is responsible for ensuring the product's quality so that it can fulfill the customer's expectation. One way of doing this is to test regularly a sample of the product to verify that its characteristics of interest have not gone out of control. To make it easier, tests are represented by a graph to check whether the process has gone out of control or is under control. Four techniques that a quality control program uses to check the overall quality of its products are:

Control charts for variables comprising, $\bar{X}$ Charts, R Charts; and Control charts for attributes comprising, P Charts.

The first two charts deal with quantitative variables, while the attributes charts deal with qualitative factors.

9.2 Objectives

After going through the unit, you should be able to:

- Explain Statistical Process Control;
- Demonstrate $\bar{X}$ Charts;
- Analyze R Charts; and
- Write about P Charts.

9.3 Statistical Process Control

Statistical Process Control (SPC) is a quality control methodology first adapted by W A Shewart. Quality means conformance to requirement. Any variability present in the manufacturing process can lead to defect. Practically, the variability cannot be eliminated so the principal objective of quality control is to reduce the variability.

Variations may be either due to chance or an assignable cause. Chance variations (also called random variation) are small variations in the product or production process due to natural differences. Every product is slightly different due to chance variation. Assignable cause variation is a non-random variation or special cause variation and it is a variation outside of the norm due to some specific cause that can be identified and corrected. Inferior inputs, inexperienced worker, and machine wear could be the possible causes of assignable variation. It is important to note that systematic non-random variation can be reduced or eliminated by dealing with a specific issue but in random variation it is mandatory that the entire process be redesigned. Therefore, if the case where the process is out-of-control, it indicates the presence of non-random patterns the management should first identify the cause of that variation and eliminate it. This process of reduction results in the process being brought “in-control”. Later, the entire process can be redesigned to improve or reduce the incidence of random or inherent variability.

Control charts are used in a manufacturing process to detect variations as early as possible. All control charts have upper limits, which are three standard errors above the acceptable level, and lower limits, which are three standard errors below the acceptable level. If a sample has a value outside of the upper or lower limits, it means the quality control process has detected an assignable cause variation. The control charts may be plotted using mean, variability, or attributes. Control charts that use mean are known as $\bar{X}$ charts and control charts that use variations are known as R charts.

The following Exhibit 9.1 shows the benefits of controls charts in healthcare sector

Exhibit 9.1: Control Charts to Quality Improvement in Healthcare

In an article in Dovepress, a scientific and medical research journal (May 2021), it was found that ariable control charts have been applied in 11 different healthcare contexts, using 17 different variables, at various levels within healthcare organizations. The main reason for applying variable control charts is to demonstrate a process change, usually following a specific change or quality intervention. The study identified various limitations and benefits of applying variable control charts. The charts are visually easy to understand for both management and employees, but they are limited by their requirement for potentially complex and resource-intensive data collection.

Contd….

Control charts contribute to quality improvement in healthcare by enabling visualization and monitoring of variations and changes in healthcare processes. The areas include surgery or pressure ulcer prevention, detecting and monitoring hospital acquired infections, etc. In addition, the charts benefit different organizational levels, at the management level and front-line staff levels. For instance, for the service quality to meet patients' expectations, the process delivery must be stable and repeatable. In such context, quality is expressed through numerical measurements, such as the waiting time, length of stay or door-to-needle-time, where it is feasible to use variable control charts. These charts use those variables as a quality indicator.

Adapted from https://www.dovepress.com/the-contribution-of-variable-control-charts-to-quality-improvement-in--peer-reviewed-fulltext-article-JHL, September 10th 2021

9.4 $\bar{X}$ Charts when the Mean and the Standard Deviation are Known

As discussed earlier, the $\bar{X}$ Charts monitors changes in the means. It measures the variation of sample means around some acceptable level. There can be two cases – one, when mean and standard deviation is known and the other when standard deviation is not known. Let us start with the first case where mean and standard deviation are known.

9.4.1 Process of Making and Interpreting $\bar{X}$ Chart

Let us assume that mean and standard deviation are μ and σ. These values are referred to as the acceptable values (benchmark values).

Step 1: *Collect sample and plot a sampling distribution.*

The mean and the standard deviation of the sampling distribution will be

$$\mu_{\bar{x}} = \mu \,; \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

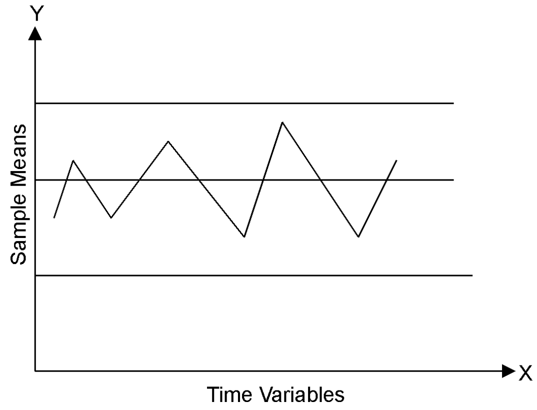
Step 2: *Plot chart by taking time variable (in days) on X-axis and the sample means on Y-axis, which is known as $\bar{X}$ Charts.*

Step 3: *Calculate the Lower Control Limit and Upper Control Limit.*

The lower control limit is given by $\mu_{\bar{x}} - 3\sigma_{\bar{x}}$ and the upper control limit by $\mu_{\bar{x}} + 3\sigma_{\bar{x}}$. Here we use 3σ because according to Chebychev's inequality theorem, irrespective of the shape of the distribution, about 89 percent of the values fall within three standard deviations on either side of the sample means and the same according to the normal distribution is 99.7 percent of the values.

Step 4: *Draw the upper control limit, lower control limit and center line in the graph. Center line refers to the mean of the sample distribution.*

Figure 9.1: Control Chart for $\bar{x}$



- If all the points fall within the region, we conclude that the process is in-control.

9.5 $\bar{X}$ Charts when the Mean and the Standard Deviation are Unknown

Step 1: *Calculation of the Grand Mean*

If population mean and the standard deviation are unknown, we can treat sample mean as population mean, but since there are various samples selected, the grand mean $\left(\bar{\bar{X}}\right)$ could best employ the population mean. The grand mean, that is $\bar{\bar{X}}$, is calculated by employing either

$$\bar{\bar{X}} = \frac{\sum X}{n \, x \, k}.$$

Where, ‘n’ denotes the number of samples that are randomly selected on a day and k denotes the number of units included in the sample. Grand means is used to calculate center line.

Step 2: *Computation of Control Limits*

As we discussed earlier, control limit is mean $\pm 3\sigma_{\bar{x}}$. As we have employed the sample mean as a substitute for population mean, the standard deviation cannot be the same as the previous one.

In this case,

Lower Control Limit $= \bar{\bar{X}} - A_2 \bar{R}$

Upper Control Limit $= \bar{\bar{X}} + A_2 \bar{R}.$

In other words,

Lower Control Limit $= \bar{\bar{X}} - \frac{3\bar{R}}{d_2\sqrt{n}}$

Upper Control Limit $= \bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}$

Where, $\bar{R}$ is the average range of sample R and the value of 'd' can be calculated from the table for different values of n.

Step 3: *Drawing Control Chart and Analysis*

Now we can draw the control chart for a given value and determine whether the process is 'in-control' or 'out-of-control'.

Now we look at a few examples.

Example 1

The following data provides the values of sample mean $\bar{X}$, and range R, for samples of size 5 each.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean ($\bar{X}$)	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Range (R)	7	4	8	5	7	4	8	4	7	9

Calculate the value of the lower control limit, upper control limit and central line for the $\bar{X}$ chart. (Use the relevant A_2 factor)

The central line of the $\bar{\bar{X}}$ chart is = Mean of the sample means ($\bar{\bar{X}}$)

Mean of the sample means or grand mean, $\bar{\bar{X}}$ = 106.6 /10 = 10.66 units.

Mean of sample range = $\bar{R} = \sum R / n = 63/10 = 6.3$

Lower Control Limit of $\bar{X}$ chart = $\bar{\bar{X}} - A_2\bar{R}$

$= 10.66 - 0.577 \times 6.3 = 7.0249$

Upper Control Limit of $\bar{X}$ chart = $\bar{\bar{X}} + A_2\bar{R}$

$= 10.66 + 0.577 \times 6.3 = 14.29$

Central line = (14.29 + 7.0249)/2 = 10.657

(The value of A_2 for a sample of size 5 is 0.577 from the control chart factors table.)

Example 2

Hero Bicycle manufactures precision ball bearings for wheel hubs, bottom brackets, headsets, and pedals. Mr. Kumar is responsible for quality control at Sportsline. He has been checking the output of the 5mm bearings used in the front wheel hubs. For each of the last 6 hours, he has sampled 5 bearings, with the following results:

Hour	Bearing Diameter (mm)				
1	5.03	5.06	4.86	4.90	4.95
2	4.97	4.94	5.09	4.78	4.88
3	5.02	4.98	4.94	4.95	4.80
4	4.92	4.93	4.90	4.92	4.96
5	5.01	4.99	4.93	5.06	5.01
6	5.00	4.95	5.10	4.85	4.91

Mr. Kumar wishes to control the variation in the bearing diameter by plotting an $\bar{X}$ chart.

The $\bar{X}$ chart for the ball bearing diameter can be plotted as below:

Hour	Bearing Diameter (mm)					Sample Mean	Sample Range
1	5.03	5.06	4.86	4.90	4.95	4.960	0.20
2	4.97	4.94	5.09	4.78	4.88	4.932	0.31
3	5.02	4.98	4.94	4.95	4.80	4.938	0.22
4	4.92	4.93	4.90	4.92	4.96	4.926	0.06
5	5.01	4.99	4.93	5.06	5.01	5.000	0.13
6	5.00	4.95	5.10	4.85	4.91	4.962	0.25
Total						29.718	1.17

The Grand Mean, $\bar{\bar{X}}$ = 4.953 and the Mean Range $\bar{R}$ = 0.195

The d_2 value for the sample size 5 is 2.326.

The Upper Control Limit = $\bar{\bar{X}} + \frac{3\,\bar{R}}{d_2\sqrt{n}} = 4.953 + \frac{3 \times 0.195}{2.326 \times \sqrt{5}}$

$= 5.0655 \approx 5.066.$

The Lower Control Limit = $\bar{\bar{X}} - \frac{3\,\bar{R}}{d_2\sqrt{n}} = 4.953 - \frac{3 \times 0.195}{2.326 \times \sqrt{5}}$

$= 4.8405 \approx 4.841.$

We observe that none of the sample means are out of the upper as well as lower control limits. Therefore, we can say that the process is in-control.

Check Your Progress - 1

1. In $\bar{x}$ charts, a process is said to be in control, when the points plotted
   a. Are on either side of the center line with small deviations
   b. Follow a random path and are between the control limits
   c. Lie beyond the control limits
   d. Lie near the control limits
   e. Both (a) and (b) above.
2. The grand mean is calculated as
   a. $\frac{\sum x^2}{\sum x}$
   b. $\frac{\sum k x^2}{\sum k x}$
   c. $\frac{\sum x^2}{\sum k x}$
   d. $\frac{\sum \bar{x}}{k}$
   e. $\frac{\sum x^{-2}}{\sum k x}$.
3. Which of the following control charts use mean?
   a. $\bar{X}$
   b. R
   c. P
   d. C
   e. Q
4. Which of the following control charts use variations?
   a. $\bar{X}$
   b. R
   c. P
   d. C
   e. Q
5. What is the alternate name of assignable cause?
   a. Random variation
   b. Small variation
   c. Special variation
   d. Natural difference variation
   e. Chance variation

9.6 R Charts

The R chart is used to monitor changes in variation. We already know that lower the variability in the products manufactured, more the consistency and hence higher the quality. Here range is used as the measure of variation, since it is easier to compute and easier for workers to understand. The upper and lower control limits are three standard errors away from the mean. In the R chart, time variable is represented on the X-axis and the sample range on the Y-axis. The center line is drawn at "$\overline{R}$".

We estimate the standard error by: $\sigma_R = d_3\sigma$

Where,

σ is the population standard deviation,

d_3 is a factor like d_2; also d_3 depends on the size of the sample 'n'.

The lower and the upper control limits are calculated by the equations given below.

$$\text{Lower control limit} = \overline{R} - \frac{3d_3\overline{R}}{d_2} = \overline{R}\left(1 - \frac{3d_3}{d_2}\right)$$

$$\text{Upper control limit} = \overline{R} + \frac{3d_3\overline{R}}{d_2} = \overline{R}\left(1 + \frac{3d_3}{d_2}\right)$$

For easy representation $(1-\left(1-\frac{3d_3}{d_2}\right))$ is denoted as D_3 and $\left(1+\frac{3d_3}{d_2}\right)$ is denoted as D_4.

$$\overline{R} = \frac{\sum_{i=1}^{i=n} R_i}{n} \quad \text{and Range (R)} = L - S$$

Reader should note that "n" that is the sample size is here equal to or lower than 6.

Example 3

The manager of a food chain outlet wants to ensure that the variability in the service time by their bearers is in control. He does a sampling of the average time taken by the bearers in serving the customers in the outlet. He finds the mean of the sample range to be 1.5 minutes. If he has collected 5 samples, each containing 10 observations to conduct the test then what is the upper control limit and lower control limit (in minutes) of the R chart? (Use the relevant D_4 factor).

The Upper Control Limit for the R chart is given as, $UCL = \overline{R} + \frac{3d_3}{d_2}\overline{R}$

$$= \overline{R}\left(1 + \frac{3d_3}{d_2}\right) = \overline{R} \times D_4 = 1.5 \times 1.777 = 2.666 \text{ minutes}$$

As $n \leq 6$, the lower limit will be zero, because the value of D_4 is zero for $n \leq 6$.

9.7 P Charts

P Charts deal with qualitative factors the data possesses and measures the proportion of defects. To use a P Charts be sure that the given data set is in confirmation with the given qualitative factors. In statistical process control, the data set which can take on only two values is called an attribute. The value here may be either success or failure, and either yes or no. The main task is to calculate the proportion of the data points that satisfy or do not satisfy the criteria. To do this we normally use the binomial distribution. Let us say that,

p = Probability of success or proportion of the data points that satisfy a particular characteristic in a given population,

$\bar{p}$ = Proportion of the data points satisfying particular characteristics in a given sample,

q = The probability of failure and the proportion of units that do not show the characteristics under consideration.

The standard deviation is also obtained on similar lines. That is, the mean and the standard deviation of the sampling distribution of the proportion are given by

$$\mu_{\bar{p}} = P \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}$$

Now we plot the P Chart with the fractional values of P and the center line at $\mu_{\bar{p}} = P$.

The control limits are given by

$$\text{Lower control limit} = \mu_{\bar{p}} - 3\sigma_{\bar{p}} = p - 3d\sqrt{\frac{pq}{n}}$$

$$\text{Upper control limit} = \mu_{\bar{p}} + 3\sigma_{\bar{p}} = p + 3d\sqrt{\frac{pq}{n}}$$

$$\text{Central line} = \bar{\bar{p}} = \frac{\Sigma \bar{p}_j}{k} = \frac{p_1 + p_2 + p_3 + \ldots + p_k}{k}$$

Where,

$\bar{p}_j$ is the sample fraction in the jth interval of time

k is the sum of all the samples considered.

Therefore in the equations given above to calculate lower and upper control limits, p should be replaced by $\bar{\bar{p}}$.

Since p cannot be negative and

- if LCL < 0 consider LCD as 0
- if LCL >1 consider LCD as 1.

Example 4

A new shearing machine is set to cut-off a piece of steel from a long bar. For various reasons, the machine at times cuts off a piece that is too long or too short. These unacceptable pieces are automatically dropped in a box and the operator of the shearing machine must count these defectives after every 100 pieces are sheared off.

The record after the first day of operation is:

Number of sheared off pieces	100	100	100	100	100
Number of defectives	5	6	7	4	8

Set out the upper control limit and lower control limit for the P chart.

Given that

Number of sheared off pieces	100	100	100	100	100
Number of defectives	5	6	7	4	8
Proportion of defectives	0.05	0.06	0.07	0.04	0.08

The mean of proportions = 0.30/5 = 0.06

Upper Control Limit for P chart $= \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.06 + 3\sqrt{\frac{0.06 \times 0.94}{100}}$

$= 0.06 + 3 \text{ x } 0.0237 = 0.131$

Lower Control Limit for P chart $= \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.06 - 3\sqrt{\frac{0.06 \times 0.94}{100}}$

$= 0.0\,6 - 3 \text{ x } 0.023 = 0.036$

<u>Check Your Progress - 2</u>

6. The upper control limit for n = 48 and p = 0.23 in p-chart will be

   a. 0.05

   b. 0.06

   c. 0.18

   d. 0.23

   e. 0.41.

7. The manager of a food chain outlet wants to ensure that the variability in the service time by their bearers is in control. He does a sampling of the average time taken by the bearers in serving the customers in the outlet. He finds that the mean of the sample range to be 1.5 minutes. If he has collected five samples each of ten observations to conduct the test then

What is the upper control limit of the R chart? (Given that d_2 =3.078 and d_3 = 0.797 for sample size of 10)

a. 0.335.
b. 1.005.
c. 1.500.
d. 2.005.
e. 2.665.

8. The lower control limit in a R chart if n = 6,

$\overline{R}$ = 5.3, d_2 = 2.534 and d_3 = 0.848 is

a. –10.62
b. 0
c. 10.62
d. 10.76
e. 10.94.

9. Which of the following control charts deal with proportion of defects?

a. $\overline{X}$
b. R
c. P
d. C
e. Q

10. Which distribution is used in P charts?

a. Poison distribution
b. Chi-Square distribution
c. t distribution
d. Binomial distribution
e. F distribution

9.8 Summary

- $\overline{X}$ charts, R charts and P chart are the four techniques that a quality control program uses to check the overall quality of its products.
- The $\overline{X}$ chart monitors changes in the means. It measures the variation of sample means around some acceptable level.
- The R chart is used for monitoring changes in variation.
- P charts deal with the attributes, which are qualitative factors possessed by the data and it measures the proportion of defects.

9.9 Glossary

Control Charts: A chart in which a parameter of considerable importance (such as $\bar{X}$, R, or P) affecting the process in a crucial manner is plotted over a period of time. They help in the identification of the assignable variation in the process and then make required adjustments in the process being monitored.

Outliers: The observations which fall outside the control limits in a control chart.

Out-of-Control Process: A process in which some of the observations fall outside the control limits (that is, outliers). Also when a process shows non-random variation in the absence of any outliers.

P Charts: A control chart used for monitoring the proportion of items in a batch that meets certain specified requirements.

R Chart: A control chart wherein the process variability is monitored.

$\bar{X}$ Chart: A control chart wherein the process mean is monitored.

9.10 Suggested Readings/Reference Material

1. Gupta, S. P. Statistical Methods. 46th Revised ed. New Delhi: Sultan Chand & Sons. 2021.
2. I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management. Pearson Education; Eighth edition, 2017.
3. Gerald Keller. Statistics for Management and Economics. Cengage, 2017.
4. Arora, P. N., and Arora, S. CA Foundation Course Statistics. 6th ed. S Chand Publishing, 2018.
5. Mario F Triola. Elementary Statistics. 13th ed., 2018.
6. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran. Statistics for Business and Economics. 13th Edition, Cengage Learning India Pvt. Ltd., 2019.
7. S D Sharma. Operations Research. Kedar Nath Ram Nath, 2018.
8. Hamdy A. Taha. Operations Research: An Introduction. 10th ed., Pearson, 2016.
9. Malhotra, N. (2012), Marketing Research: An Applied Orientation, 7th ed., Pearson, 2019.
10. Cooper, D.R. and Schindler, P.S. and J. K. Sharma (2018), Business Research Methods, 12th edition, McGraw-Hill Education.

9.11 Self-Assessment Questions

1. Explain the concept of Grand Mean in Statistical Quantity Control.

2. What are different types of control chart? Explain.

3. What is the objective of statistical quality control in an industrial organization?

4. Explain briefly the terms: tolerance limit and control limit as used in statistical quality control.

5. Explain the principle on which a control chart is based.

9.12 Answers to Check Your Progress

1. (b) Follow a random path and are between the control limits

2. (d) $\frac{\sum \bar{x}}{k}$

3. (a) $\bar{\bar{X}}$

4. (b) R

5. (c) Special cause variation. Assignable cause variation is a non-random variation or special cause variation

6. (e) 41

(p+3*sqrt(pq/n))

7. (d) 4.442

8. (b) 0

(For $n \leq 6$, lower is zero)

9. (c) P

10. (d) Binomial distribution

Quantitative Methods

Course Structure

Block	Unit Nos.	Unit Title
I	**Introduction to Statistics and Probability**	
	1.	Arranging Data
	2.	Central Tendency and Dispersion
	3.	Probability
	4.	Probability Distribution and Decision Theory
II	**Statistical Relations and Hypothesis Testing**	
	5.	Statistical Inference and Hypothesis Testing
	6.	Correlation and Linear Regression
III	**Statistical Regression and Quality Control**	
	7.	Multiple Regression
	8.	Time Series Analysis
	9.	Quality Control
IV	**Statistical Distributions, Variations and IT**	
	10.	Chi-Square Test and Analysis of Variance
	11.	Role of IT in Modern Business Enterprise
	12.	Statistical Software Tools
V	**Advanced Statistics**	
	13.	Index Numbers
	14.	Simulation
	15.	Linear Programming
VI	**Business Research**	
	16.	Introduction to Business Research Methods
	17.	Questionnaire Design
	18.	Report Writing