NOTES

extension allows SPSS to run any of the statistics in the free software package R. From version 14 onwards SPSS can be driven externally by a Python or a VB.NET program using supplied 'plug-ins'.

SPSS places constraints on internal file structure, data types, data processing and matching files, which together considerably simplify programming. SPSS datasets have a two-dimensional table structure where the rows typically represent cases (such as individuals or households) and the columns represent measurements (such as age, sex or household income). Only two data types are defined: numeric and text (or "string"). All data processing occurs sequentially case-by-case through the file. Files can be matched one-to-one and one-to-many, but not many-to-many.

The graphical user interface has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text and formulas cannot be stored in these cells. The 'Variable View' displays the metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type and a variety of other characteristics. Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as statistical surveys are more often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS.

SPSS can read and write data from ASCII text files (including hierarchical files), other statistics packages, spreadsheets and databases. SPSS can read and write to external relational database tables via ODBC and SQL.

Statistical output is to a proprietary file format (*.spv file, supporting pivot tables) for which, in addition to the in-package viewer, a stand-alone reader can be downloaded. The proprietary output can be exported to text or Microsoft Word. Alternatively, output can be captured as data (using the OMS command), as text, tab-delimited text, PDF, XLS, HTML, XML, SPSS dataset or a variety of graphic image formats (JPEG, PNG, BMP and EMF).

SPSS server is a version of SPSS with a client/server architecture. It had some features not available in the desktop version, such as scoring functions (Scoring functions are included in the desktop version from version 19).

Versions

Early versions of SPSS were designed for batch processing on mainframes, including; for example, IBM and ICL versions, originally using punched cards for input. A processing run read a command file of SPSS commands and either a raw input file of fixed format data with a single record type, or a 'getfile' of data saved

NOTES

by a previous run. To save precious computer time an 'edit' run could be done to check command syntax without analysing the data. From version 10 (SPSS-X) in 1983, data files could contain multiple record types.

SPSS version 16.0 runs under Windows, Mac OS 10.5 and earlier, and Linux. The graphical user interface is written in Java. The Mac OS version is provided as a universal binary, making it fully compatible with both PowerPC and Intel-based Mac hardware.

Prior to SPSS 16.0, different versions of SPSS were available for Windows, Mac OS X and Unix. The Windows version was updated more frequently, and had more features than the versions for other operating systems.

SPSS version 13.0 for Mac OS X was not compatible with Intel-based Macintosh computers, due to the Rosetta emulation software causing errors in calculations. SPSS 15.0 for Windows needed a downloadable hotfix to be installed in order to be compatible with Windows Vista. The latest version of SPSS is 19.0.

CHECK YOUR PROGRESS

1. What is coding?
2. Who developed SPSS?
3. What are the two types of variables?
4. List a few versions of SPSS used for research.

3.4 PRESENTATION OF DATA: DIAGRAMS AND GRAPHS

In common parlance, we come across the use of charts and graphs to support facts and figures. Graphical presentation of data aids in easy comprehension and is a wonderful visual aid in grasping the trend of data by one look. Graphical representation helps a researcher to demonstrate his point effectively by giving an idea about the shape of the distribution. Some basic methods of presenting data through diagramms are as follows:

1. Bar chart
2. Pie chart
3. Histogram
4. Frequency polygon
5. Ogive

1. Bar chart

A bar chart is a graphical representation that depicts data summarized in a frequency, relative frequency or percentage frequency. A bar chart is prepared by depicting

data that have been classified as frequency on y-axis of the graph. The class intervals are specified on the horizontal axis on x-axis of the graph. The following data can be represented better through a bar chart.

	Series 1	Series 2	Series 3
Category 1	4.3	2.4	2
Category 2	2.5	4.4	2
Category 3	3.5	1.8	3
Category 4	4.5	2.8	5

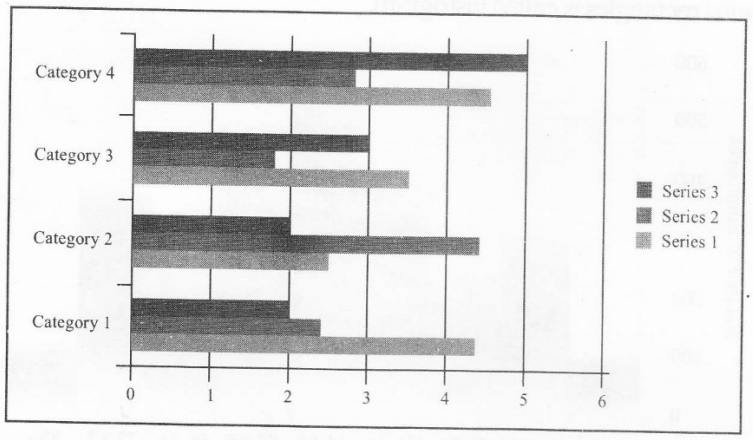

Fig. 3.1 Bar Chart

2. Pie chart

It is a circular depiction of data in which a circle is divided into several sectors or slices, with areas equal to the corresponding component. They are used in a variety of situation to depict budget allocations, market share, etc. It is constructed as follows: Firstly, the proportion of a particular component to the whole is determined. Pie chart is a circular depiction of data, where a circle measures 360^0 totally. Each component proportion is then multiplied by 360 to get the accurate number of degrees to represent each component. Statistical packages like Excel and SPSS can generate pie diagrams once data is fed and therefore it is not necessary for a researcher to calculate relative degrees of each component.

Example: Suppose we have the following data. Resultant pie diagram is given.

	Sales
1st Qtr	8.2
2nd Qtr	3.2
3rd Qtr	1.4
4th Qtr	1.2

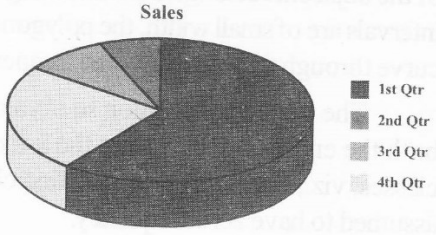

Fig. 3.2 Pie Chart

NOTES

3. Histogram

A histogram can be defined as a set of rectangles, each proportional in width to the range of the values within a class and proportional in height to the class frequencies of the respective class interval. Suppose we have a frequency distribution. For histogram, we first mark, along the x-axis, all the class intervals on a suitable scale. On each class interval, rectangles are erected the heights of which are proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If, however, the classes are of unequal width, then the height of the rectangle will be proportional to the ratio of the frequencies to the width of the class. The resultant diagram of continuous rectangles is called histogram.

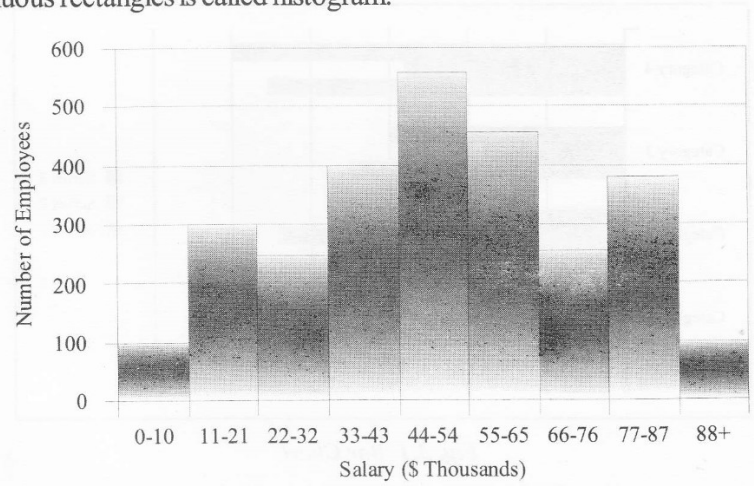

Fig. 3.3 Histogram

4. Frequency polygon

A frequency polygon is a graphical representation of the frequencies in which line segments connecting the dots depict a frequency distribution. For an ungrouped distribution (i.e., without class intervals), the frequency polygon is obtained by plotting points with variate values on the x-axis and corresponding frequencies on the y-axis. Thereafter, plotted points are joined by means of straight lines.

In case we have a grouped frequency distribution with class intervals, the mid-values of class intervals are plotted on x-axis. For equal class intervals, the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width, the polygon is obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

The frequency polygon so obtained is extended to the base line (x-axis) at both the ends so that it meets the x-axis at the mid-points of two hypothetical classes, viz., the class before the first class and the class after the last class, each assumed to have zero frequency.

NOTES

5. Ogive

An ogive is a cumulative frequency curve, or in other words, it is a cumulative frequency polygon. Data values are shown on the horizontal x- axis while cumulative frequencies are shown on the vertical y-axis.

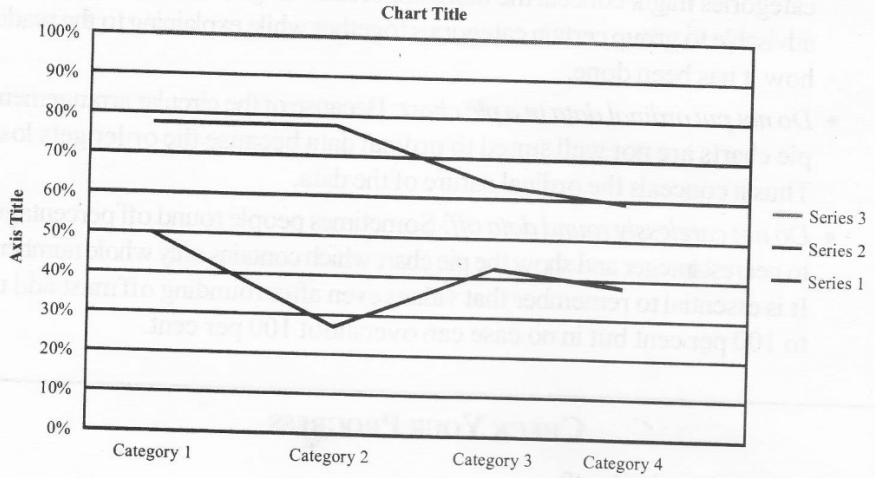

Fig. 3.4 Ogive

Best practices and pitfalls in the use of charts and graphs

Robert E. Stine and Dean Foster (2011:41) outline the following best practices in the use of charts and graphs:

- *The frequencies of a categorical variable may be shown through a bar chart.* Order the categories either alphabetically or by size. The bars can be oriented either horizontally or vertically.
- *The proportions of a categorical variable may be shown by using a pie chart.* Arrange the portions (if you can) to make differences in the sizes more recognizable. A pie chart is a good way to show that one category makes up more than half of the total.
- *Preserve the ordering of an ordinal variable.* Arrange the bars on the basis of the order implied by the categories, not the frequencies.
- *Respect the area principle.* The relative size of a bar or slice should match the count of the associated category in the data relative to the total number of cases.
- *Label your chart to show the categories and indicate whether some have been combined or omitted.* Name the bars in a bar chart and slices in a pie chart.

Robert E. Stine and Dean Foster (2011:42) have outlined certain pitfalls which need to be avoided in the use of charts and graphs:

- *Avoid elaborate plots that may be deceptive:* Instead of 2 D pie chart, modern software may enable one to go in for a three dimensional pie chart

NOTES

to improve its visual appeal. But showing the pie on a slant violates the area principle and makes it harder to compare the shares, the principle feature of data that a pie chart ought to show.

- *Do not show too many categories:* A bar chart or pie chart with too many categories might conceal the more important categories. In such cases, it is advisable to group certain categories together while explaining to the reader how it has been done.
- *Do not put ordinal data in a pie chart:* Because of the circular arrangement, pie charts are not well suited to ordinal data because the order gets lost. Thus it conceals the ordinal nature of the data.
- *Do not carelessly round data off:* Sometimes people round off percentages to nearest integer and show the pie chart which contains only whole numbers. It is essential to remember that values even after rounding off must add up to 100 per cent but in no case can overshoot 100 per cent.

Check Your Progress

5. What is a pie chart?
6. List three best practices for using graphs and charts.

3.5 SUMMARY

- Once the data has been collected, it needs to be carefully sorted and arranged before we can undertake a meaningful analysis. Quantitative as well as qualitative data needs to be sorted, coded or categorized or indexed as the case may be and thus made ready for analysis. Coding refers to the process of conceptualizing research data and classifying them into meaningful and relevant categories for the purpose of data analysis and interpretation. A number is assigned to each category in the form of a code.
  - The codes must be mutually exclusive
  - Coding formats for each question must be comprehensive and
  - Codes must be consistently applied by field investigators or data entry operators
- Statistical tools are helpful in condensing large amounts of data. It also facilitates drawing of useful inferences from data. Quantitative data can be classified using tally marks and presented in the form of a frequency table which shows the variable as well as its frequency.
- Quantitative data can also be arranged using class intervals in a frequency table. It is important to note that the classes should be exhaustive; or in other words, every value should be included in one or the other classes. They should also be mutually exclusive and non-overlapping. The number

of classes should neither be too large nor too small. It could be between 5 and 15 but there are no hard and fast rules.

- SPSS is a comprehensive integrated software package for statistical data analysis. It has the following functions on its menu bar: Data, Transform, Analyze and Graph. These facilitate data analysis and a variety of numerical operations like tables, graphs, correlation, regression analysis, non-parametric tests, comparing means by one-way ANNOVA test and two-way ANNOVA test, etc. SPSS for Windows allows one to store data, perform transformations and analyses, and produce charts and graphs of results. Data are entered using a spreadsheet and results are displayed in a separate output window.
- Graphical presentation of data aids in easy comprehension and is a wonderful visual aid in grasping the trend of data by one look. Graphical representation helps a researcher demonstrates his point very effectively by giving an idea about the shape of the distribution. Some basic methods of presenting data in graphs and charts are as follows:
  - o Bar chart
  - o Pie chart
  - o Histogram
  - o Frequency polygon
  - o Ogive
- A bar chart is prepared by depicting data that have been classified as frequency on y-axis of the graph. The class intervals are specified on the horizontal axis on x-axis of the graph.It is a circular depiction of data in which a circle is divided into several sectors or slices, with areas equal to the corresponding component. They are used in a variety of situation to depict budget allocations, market share, etc.
- For histogram, we first mark off, along the x-axis, all the class intervals on a suitable scale. On each class interval, rectangles are erected whose heights are proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class.
- For an ungrouped distribution, i.e., without class intervals, the frequency polygon is obtained by plotting points with variate values on the x-axis and corresponding frequencies on the y-axis. Thereafter, plotted points are joined by means of straight lines. In case we have a grouped frequency distribution with class intervals, the mid-values of class intervals are plotted on x-axis.
- Ogive is a cumulative frequency curve or a frequency polygon. Data values are shown on the horizontal x-axis while cumulative frequencies are shown on the vertical y-axis.

NOTES

3.6 KEY TERMS

- **Data processing:** The method of sorting, coding and arranging data for undertaking meaningful analysis
- **Coding:** The process of conceptualizing research data and classifying them into meaningful and relevant categories for the purpose of data analysis and interpretation
- **Continuous variables:** Values which are marked on a continuous scale, possibly between appropriate limits
- **Categorical variables:** Those values which may only come from a fixed set of choices

3.7 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Coding refers to the process of conceptualizing research data and classifying them into meaningful and relevant categories for the purpose of data analysis and interpretation. A number is assigned to each category, in the form of a code; for example, if the field relates to caste, Code 1 can be given for general category candidates, Code 2 for persons belonging to Scheduled Castes and Code 3 for Scheduled Tribes and Code 4 for Other Backward Castes. Coding formats may be included in the questionnaire.
2. In 1968, Norman H. Nie, C. Hadlai Hull and Dale H. Brent developed SPSS which is now used widely in academic, business, government and other environments. 'SPSS' stands for 'Statistical Product and Service Solutions'.
3. The two types of variables are *continuous* and *categorical*.
4. SPSS version 16.0 runs under Windows, Mac OS 10.5 and earlier, and Linux. The graphical user interface is written in Java. The Mac OS version is provided as a universal binary, making it fully compatible with both PowerPC and Intel-based Mac hardware. Prior to SPSS 16.0, different versions of SPSS were available for Windows, Mac OS X and Unix. The Windows version was updated more frequently, and had more features, than the versions for other operating systems. SPSS version 13.0 for Mac OS X was not compatible with Intel-based Macintosh computers, due to the Rosetta emulation software causing errors in calculations. SPSS 15.0 for Windows needed a downloadable hotfix to be installed in order to be compatible with Windows Vista. The latest version of SPSS is 19.0.
5. A pie chart is a circular depiction of data in which a circle is divided into several sectors or slices, with areas equal to the corresponding component. They are used in a variety of situation to depict budget allocations, market share, etc. It is constructed as follows: Firstly, the proportion of a particular

component to the whole is determined. Pie chart is a circular depiction of data, where a circle measures 360^0 totally. Each component proportion is then multiplied by 360 to get the correct number of degrees to represent each component. Statistical packages like Excel and SPSS can generate pie diagrams once data is fed and therefore it is not necessary for a researcher to calculate relative degrees of each component.

6. Robert E. Stine and Dean Foster (2011:41) outline the following best practices in the use of charts and graphs:
   - Use a bar chart to show the frequencies of a categorical variable. Order the categories either alphabetically or by size. The bars can be oriented either horizontally or vertically.
   - Use a pie chart to show the proportions of a categorical variable. Arrange the portions (if you can) to make differences in the sizes more recognizable. A pie chart is a good way to show that one category makes up more than half of the total.
   - Preserve the ordering of an ordinal variable. Arrange the bars on the basis of the order implied by the categories, not the frequencies.

3.8 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the utility of data processing?
2. Write a short note on SPSS.
3. Why is graphical presentation of data required?
4. Discuss histograms and frequency polygons as tools for data presentation.
5. According to Stine and Foster, what are the pitfalls of using charts and graphs?

Long-Answer Questions

1. Write a note on data processing.
2. Discuss the use of SPSS in data research.
3. Discuss various methods of presenting data graphically.

3.9 FURTHER READING/REFERENCES

Babbie, Earl. 2009. *Research Methods in Sociology*. New Delhi: Cengage Learning.

Gupta, S.C. and V.K. Kapoor. 2000. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand.

Singh, Kultar. 2007. *Quantitative Social Research Methods*. New Delhi: Sage Publications.

NOTES

References

Kultar Singh. 2007. *Quantitative Social Research Methods*. New Delhi: Sage, pp. 81–83.

Singh, Kultar. 2007. *Quantitative Social Research Methods*. New Delhi: Sage.

Bajpai, Naval. 2010. *Business Statistics*. New Delhi: Pearson Education.

Gupta, S.C. and V.K. Kapoor. 2002. *Fundamentals of Mathematical Statistics*. New Delhi: Sultan Chand.

Stine, Robert and Dean Foster. 2011. *Statistics for Business Decision-Making and Analysis*. New Jersey: Pearson Higher Education.

Learning Resources Help Guide, *What is SPSS*? University of Lincoln, 2002.

http://www.nyu.edu/its

UNIT 4 APPLICATION OF STATISTICAL TESTS

NOTES

Structure

4.0 INTRODUCTION

A test that offers a mechanism to make quantitative decisions about one or more than one processes is known as a statistical test. This test intends to check for the sufficiency of evidence to reject an inference or hypothesis about the process. This inference is known as null hypothesis. In case the null hypothesis is required to be believed as true, it should not be rejected. If considered true, a null hypothesis is rejected, then the result may be disappointing. The result in such a case would probably indicate the absence of enough data to prove the required theory.

4.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss the use of mean, median and mode and their limitations
- Explain the meaning and use of the measures of variation or dispersion and in particular, quartile deviation, mean deviation and standard deviation
- Describe the meaning and interpretation of Karl Pearson's coefficient of correlation, as well as compute it

- Describe the meaning and interpretation of Spearman's rank correlation, as well as compute it
- Define and interpret application of the Chi-square test

NOTES

4.2 MEASURES OF CENTRAL TENDENCY

Quantitative data show a tendency to concentrate on certain values, usually somewhere in the centre of the distribution. Measures of this tendency are called *measures of central tendency or averages.*

Arithmetic mean (or simply mean), median and mode are some measures of central tendency. They are statistical values which help us to comprehend a mass of data easily.

4.2.1 Arithmetic Mean

Arithmetic mean of a set of observations is their total sum divided by the number of observations.

Example 1:

For instance, the weights of 5 children in a class are 30, 35, 44, 28 and 40 kg respectively. The arithmetic mean or average weight of these children is (30 + 35 + 44 + 28 + 40)/5 = 177/5 = 35.4

Example 2:

In case we have the following frequency distribution:

Variable (X):	1	2	3	4	5
Frequency (f):	8	6	2	7	8

The arithmetic mean is obtained as follows:

$$\frac{\Sigma xf}{\Sigma f} = [x_1f_1 + x_2f_2 + \ldots + x_nf_n]/[f_1 + f_2 + \ldots f_n] = (1 \times 8 + 2 \times 6 + 3 \times 2 + 4 \times 7 + 5 \times 8)/(8 + 6 + 2 + 7 + 8)$$

$$= 94/31 = 3.03$$

The symbol 'Σ' represents sum of all units.

Example 3:

In case we have a frequency distribution with a class interval, we identify the mid-value of each class interval which in turn is multiplied with the frequency of class interval before calculating the average.

For example, consider the following frequency distribution:

Marks Interval	No. of Students	Mid-Point of Class	
	(f)	X	f× X
0–10	13	5	65
10–20	2	15	30

20–30	5	25	125
30–40	6	35	210
40–50	8	45	360
50–60	4	55	220
60–70	2	65	130
70–80	1	75	75
80–90	1	85	85
Total	42		1300

Arithmetic mean of the above frequency distribution is 1300/42 = 30.95

The above average at once gives an idea as to how poorly the class has fared in the examination.

Merits of mean

The mean is well defined. It is easy to understand and calculate. It is also based on all observations. It is amenable to further mathematical treatment, i.e., in case there are two separate series and if their means have been computed, the mean of the composite data (after combining the two series) can easily be obtained. Compared to other measures of central tendency, mean is least affected by fluctuations of sampling. In other words, it is a stable average.

Demerits of mean

(i) One cannot find out the mean by mere inspection of data or by looking at a graph, as is possible with other measures of central tendency and in cases with small data.

(ii) In case we are talking about qualitative characteristics like beauty, pride, honesty, etc., it is not possible to determine mean, which is possible only in case of quantitative data.

(iii) If there are one or two extreme values in a distribution, they disproportionately influence mean, which will not truly represent other items in the distribution.

(iv) Arithmetic mean may lead to inaccurate conclusions if the details from which it is computed are not given.

For example, compare growth rates of the following two countries:

Year:	2005	2006	2007	2008	2009	Average growth over 5 years
Country A:	10%	8%	6%	4%	2%	6%
Country B:	2%	4%	6%	8%	10%	6%

Though both countries have registered an average growth rate of 6 per cent during the last 5 years, if we look at the mean alone, it can lead to misleading conclusions. As one can see easily, country B is on the upswing and has registered a steady growth, while country A is on the downhill as far as growth is concerned. There is a decreasing trend in growth in case of country A, while in the case of country B there is an increasing trend.

NOTES

(v) If we have a frequency distribution with one or both extreme class intervals are open, i.e., < 10 marks or > 80 marks, average cannot be computed. In such cases, it is not possible to compute mid value of the class interval.

(vi) In extremely asymmetrical (or skewed distribution), the arithmetic mean is not a suitable measure.

4.2.2 Median

Median refers to middle value in a series of observations and is thus a 'positional' average. In case we have a raw data, it is arranged either in ascending order or descending order and then median is determined.

For instance, if the heights of 5 students in a group are 120 cm, 134 cm, 135 cm, 138 cm, 140 cm, then median height of this group is 135 cm for it is the middle value. It is important to note that median divides the entire series into two equal parts. For instance, in the above series, there are equal number of observations above median value (2) and equal number of observations below it (2).

In the above series, there were odd number of observations (5) and hence it was easy to locate median or the middle value. What if there is even number of items in a given series, say 6 or 8? In such an event, though any value between middle two values can be regarded as median yet, by convention, the arithmetic mean of middle two values is taken as median. For example, the numbers of cinema halls in six cities are as follows:

20, 24, 28, 30, 33, 35

Here, there is no single middle value but two of them, i.e., 28 and 30. They divide entire series into two equal parts. Median of the above series is (28 + 30)/ 2 = 29.

Example 4: Consider the following distribution of families in a street arranged by number of children in each family.

No. of Children	No. of Families	Cumulative Frequency
X	f	c.f.
1	4	4
2	5	9
3	6	15
4	5	20
5	3	23
6	1	24
7	1	25
Total	N=25	

The above is an example of a discrete frequency distribution. In such cases, the cumulative frequency is first computed. In other words, number of families with 3 or less than 3 children is obtained by adding frequencies of those having one, two and three children, i.e., 4+ 5 + 6 = 15. After tabulating the cumulative frequencies given in the last column, the following rule is adopted for determining median:

1. Determine N/2, where N stands for a sum of total frequencies or $f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7$
2. What is the cumulative frequency just greater than N/2 ? The corresponding x value is median.

Thus using the above rule, N/2 is 12.5 and cumulative just greater than N/2 is 15. The corresponding x value is 3. In other words, median for the above frequency distribution is 3.

Besides the above discrete distribution, there could be continuous distribution with class intervals and frequencies. In such cases, median class is first determined by finding the class interval corresponding to 'f' which is just greater than N/2 and thereafter median is computed by the following formula:

$$\text{Median} = l + \frac{h}{f}[N/2 - C]$$

In the above formula, 'l' stands for lower limit of median class; 'f' for frequency of median class; 'h' is the magnitude of the median class and 'c' is the cumulative frequency of the class preceding the median class.

Merits of median

It is well-defined, easy to understand and is easy to calculate. In case the number of items is not too many, one can even locate it just by looking at the end and finding the middle value. Unlike mean, it is not at all affected by extreme values and can even be calculated in cases of distributions with open-ended classes. In case of qualitative data which cannot be measured quantitatively but can be arranged in ascending or descending order of magnitude, median can be used to find out the average value, viz., average intelligence or honesty. It is also used for determining the typical value in problems concerning wages, distribution of wealth, etc.

Demerits of median

In case of even number of observations, median cannot be determined exactly unlike mean. In such cases, it is estimated by taking mean of two middle terms. Further, it is not based on all observations and is therefore termed as *insensitive*.

Consider the following five observations:

10, 15, 30, 40, 45

The median is obviously the middle value, i.e., 30. If one replaces first two values with any value less than 30 or last two observations with any value above 30, the value of median is not affected and still remains 30. Unlike mean, it is affected much by fluctuations of sampling.

4.2.3 Mode

Mode refers to the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely. It is that variable which

NOTES

is predominant in the series. For instance, we keep hearing examples of mode in common parlance; like the average family size in Uttar Pradesh is 4 or an average American's height is about 5 feet 10 inches. In all such cases, the value given refers to mode.

Example 5:

For instance, consider the following frequency distribution:

Variable X:	1	2	3	4	5	6
Frequency F:	4	5	7	15	6	5

In the above series, maximum frequency is 15 and value of variable 'x' corresponding to that frequency is '4'. Mode for the above distribution, therefore, is 4.

However, computation of mode presents certain difficulties in the following cases where:

(i) If the maximum frequency is repeated.

(ii) If the maximum frequency occurs in the very beginning or at the end of the distribution

(iii) If there are irregularities in the distribution.

In such cases, mode is determined by the method of grouping.

If there is a continuous frequency distribution, mode is determined by using

$$\text{Mode} = 1 + \frac{h[f_1 - f_0]}{[2f_1 - f_0 - f_2]}$$

where the 'l' stands for lower limit, 'h' the magnitude and f_1 the frequency of the modal class, and f_0 and f_2 are frequencies of the classes preceding and succeeding the modal class respectively.

Merits of mode

It is easy to understand and easy to calculate. In some cases, it can be located by inspection. Mode is not affected by extreme values. Open-ended classes do not pose any problems in the determination of mode. Mode is often the most appropriate one to find ideal size and hence used in business forecasting to make decisions like manufacture of ready-made shirts or shoes, etc.

Demerits of mode

It is not well-defined. In some cases, there can be two or even more modes for the same distribution. It is not based on all observations. Unlike mean, mode is affected by fluctuations of sampling.

Check Your Progress

1. Find mean and median of the following series of values:
   10, 12, 14, 18, 30
2. Find mean, median and mode of the following discrete frequency distribution:

X:	1	2	3	4	5	6	7
F:	4	6	8	12	11	7	3

3. Compute the mean, median and mode in the case of following continuous frequency distribution of the marks of students, in the paper of research methodology:

Class Interval	No. of Students (f)
0–10	4
10–20	5
20–30	5
30–40	6
40–50	10
50–60	15
60–70	8
70–80	7

4.2.4 Measures of Dispersion

Before we take up quartile deviation, it is important to understand the concept of 'partition values'. Often, we hear that a person has scored 92 percentile in TOEFL exam. By that we mean, of all the persons who took that examination on a particular date, 92 per cent of students have secured less than him/ her and that only 8 per cent managed to secure a higher mark than him. Partition values are the values which divide the series into a number of equal parts. In case of percentiles, they divide series into 100 equal parts while deciles divide a series into ten equal parts.

The three points which divide the series into four equal parts are called quartiles. The first quartile or Q_1 is that value which exceeds 25 per cent of the observations and is exceeded by 75 per cent of the observations. Opposite is true for third quartile or Q_3. The second quartile, Q_2, which divides the series into two equal parts and is in the middle is nothing but the median.

Example 6:

In the case of following discrete frequency distribution, quartiles are calculated as follows:

Variable X:	1	2	3	4	5	6
Frequency F:	6	12	13	16	12	11

NOTES

In the above distribution, the less than cumulative frequency is computed as follows:

Cumulative frequency: 6 18 31 47 59 70

N = Sum of total frequencies = $\Sigma f_i = 70$

Here, N/2 = 70/2 = 35. Cumulative frequency (c.f.), just greater than 35 is 47 and corresponding x is 4, which is median.

First quartile Q_1: Here N/4 = 70/4 = 17.5 and c.f. just greater than 17.5 is 18. Thus $Q_1 = 2$.

Third quartile Q_3. Here 3N/4 = 52.5 and c.f. just greater than 52.5 is 59. Hence Q_3=5.

In the case of a continuous frequency distribution with class intervals, using the same analogy adopted for median, first quartile is computed by the formula:

$$Q_1 = l + \frac{h}{f}\left[\frac{N}{4} - C\right] \qquad \text{while } Q_3 = l + \frac{h}{f}\left[\frac{3N}{4} - C\right]$$

where 'l' stands for lower limit of class interval containing first quartile Q_1; 'f' for frequency of that class; 'h' is the magnitude of that class and 'c' is the cumulative frequency of the class preceding that class.

In case of formula for third quartile Q_3 the class interval is the one containing Q_3.

4.2.5 Measures of Variation or Depression

While measures of central tendency give us a clue about the concentration of observations about the central part of the distribution, they are not sufficient in themselves as can be seen from the following two series:

Series 1: 30 40 50 60 70

Series 2: 46 48 50 52 54

In both cases, mean and median are 50 but that is not all. It does not give us an idea about the spread of the series or its scattering. It is therefore important that measures of central tendency or averages are supplemented by measures of dispersion. Dispersion means scattering. In series 2, there is less variation around mean (or homogeneous) as compared to series 1 (marked by heterogeneity). Measures of variation or dispersion seek to capture the degree to which numerical data tend to spread about an average value.

Range

The simplest measure of dispersion is the range which can be defined as the difference between the highest and lowest value in a series. It gives us an idea of dispersion.

Example 7: The marks secured by 6 students who opted for French elective in a class are as follows:

25, 40, 55, 60, 80, 60

NOTES

In this case, maximum mark is 80 while lowest mark is 25 and hence range = 80–25 = 55.

As range takes into account only two extreme values in a distribution and ignores the rest, it is therefore not an accurate measure of variation. There are better measures which are given hereinunder:

4.2.6 Quartile Deviation

Quartile deviation is defined as

$$Q = [Q_3 - Q_1]/2$$

Q_1 stands for first quartile while Q_3 stands for third quartile of the distribution. It is a better measure than range as it makes use of 50 per cent of values in a given distribution.

4.2.7 Mean Deviation

Suppose we have the following distribution

Variable X : x_1 x_2 x_3.....x_n

Frequency 'f': f_1 f_2 f_3.....f_n

Then mean deviation from the average A (usually mean, median or mode) is given by the following formula:

Mean deviation from average $A = \frac{1}{N}\sum f_i |x_i - A|$

Where the sum of all frequencies is N.

$|x_i - A|$ represents modulus or absolute value of the deviation (x_i–A), where the negative sign is ignored. Mean deviation is a better measure of dispersion than range or quartile deviation as it is based on all observations in a distribution. However, by ignoring the signs of the deviations (x_i–A) does not make it amenable to further mathematical treatment.

4.2.8 Standard Deviation

Standard deviation is defined as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_i f_I (X_i - \bar{x})^2}$$

Where $\bar{x}$ stands for the arithmetic mean of the distribution and $\Sigma f_i = N$.

Standard deviation is denoted by Greek letter, sigma (σ). In practice, one can infer the following from the standard deviation. The greater the value of standard deviation, the greater is the magnitude of the values from their mean and in such cases one can conclude that there is more heterogeneity. On the other hand, smaller the value of standard deviation, greater is the uniformity in the observations.

Thus, it overcomes the main drawback encountered in mean deviation (of ignoring the positive and negative signs) by going in for the positive square root of

NOTES

the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It enables further mathematical treatment. In other words, if we have two series with different sizes, means and standard deviations, then the standard deviation of the combined series of data can be easily computed. Standard deviation is affected least by fluctuations of sampling. It gives greater weight to extreme values. Keeping in view its merits, standard deviation finds many applications in statistics and is also the best and most powerful measure of dispersion.

The square of standard deviation is called the *variance* and is given by:

$$\sigma^2 = \frac{1}{N}\sum_i f_i(x_1 - \bar{x})^2 = \frac{1}{N}\sum_i f_i(x_1)^2 - \left[\frac{1}{N}\sum_i f_i x_1\right]^2$$

Example 8:

Consider the following data in relation to exports of Cadbury units in India in 2010.

Unit:	1	2	3	4	5	6	7
Exports (in crores):	120	30	40	100	22	77	45

Compute quartile deviation, mean deviation, standard deviation.

The less than cumulative frequency is as follows:

c.f. = 120 150 190 290 312 389 434

Here N = total frequency = 434 and hence N/4 = 108.5 and the cumulative frequency just higher than 108.5 is 150 and the corresponding value of variable 'x' is 2. Hence, first quartile is 2.

In the same way, 3N/4 = 325.5 and the cumulative frequency just greater than 3N/4 is 389 and the corresponding value of the variable 'x' is 6. Hence, third quartile is 6.

Quartile deviation is $[Q_3 - Q_1]/2 = [6 - 2]/2 = 2$.

Mean deviation from mean 'A' $= \frac{1}{N}\sum_i f_i |x_1 - A|$

Mean for the above series is first computed as follows:

Mean = A = $\frac{\Sigma xf}{\Sigma f}$ = $[x_1 f_1 + x_2 f_2 + \ldots + x_n f_n]/[f_1 + f_2 + \ldots f_n] = 4$ [rounded off]

Unit 'x':	1	2	3	4	5	6	7
Exports [in crores 'f']:	120	30	40	100	22	77	45
$\lvert x_1 - A \rvert$	3	2	1	0	1	2	3
$f_i \lvert x_1 - A \rvert$	360	60	40	0	22	154	135

Mean deviation from mean 'A' $= \frac{1}{N}\sum_i f_i |x_1 - A| = 1.78$

NOTES

In order to compute standard deviation, we can first compute variance and then compute its square root which is nothing but standard deviation.

Variance is given by the following formula:

$$\sigma^2 = \frac{1}{N}\sum_i f_i(x_1 - \bar{x})^2 = \frac{1}{N}\sum_i f_i(x_1)^2 - \left[\frac{1}{N}\sum_i f_i x_1\right]^2$$

Unit 'x':	1	2	3	4	5	6	7
Exports [in crores 'f']:	120	30	40	100	22	77	45
fx_i	120	60	120	400	110	462	315
fx_i^2	120	120	360	1600	550	2772	2205

$= [434 - 13] = 421$

Hence standard deviation is $\sigma = \sqrt{\sigma^2} = 20.52$ (rounded off two decimals)

Check Your Progress

4. Calculate (a) quartile deviation (b) mean deviation from mean and (c) standard deviation for the following data

Age 'x':	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No members In a block 'f'	3	5	6	10	7	8	5

4.3 CORRELATION

Suppose we have a bivariate distribution, i.e., a distribution involving two variables 'x' and 'y'. One would like to know whether there is any relation or association between these two variables. Are movements in one variable accompanied by similar movements in the other variable? Measures of association are statistics that measure the strength of a relationship between two or more variables. Correlation is a quantitative measure of the relationship between two variables.

If changes in one variable affect a change in the other variable, then the variables are said to be correlated. 'If the two variables deviate in the same direction, i.e., if the increase (or decrease) in one results in a corresponding increase (decrease) in the other, correlation is said to be direct or positive. But if they constantly deviate in the opposite directions, i.e., an increase (or decrease) in one results in a corresponding decrease (or increase) in the other, correlation is said to be diverse or negative.'

For example, female literacy is negatively correlated to infant mortality and maternal mortality. As female literacy goes up, World Bank studies have shown that they have resulted in lower infant and maternal mortality rates. On the other hand, there is a positive correlation between female literacy and female work

NOTES

participation rates. As female literacy rates go up, so will their work participation rates, thereby indicating a positive or direct correlation between these two variables. If onion prices or for that matter prices of any commodity go up by a big margin, their consumption is likely to come down, thereby indicating a negative correlation between the price of a commodity and its demand. As incomes of a people go up, so will their expenditures, which is a case of positive correlation. In the same way, there exists positive correlation between heights and weights of students.

Interpreting correlation coefficient values

A correlation coefficient indicates both the type of correlation as well as the strength of the relationship. The coefficient value illustrates the strength whereas the sign tells us whether variables change in the same direction or in opposite directions. A positive correlation indicates that as one variable increases, the other variable also increases in a similar way. On the other hand, a negative correlation which is preceded by – sign points out there is an inverse relationship between the two variables, i.e., an increase in one variable is associated with the decrease in the other variable.

Correlation coefficient value lies between –1 and 1. Correlation is regarded as perfect if the deviation in one variable is followed by a corresponding and proportionate deviation in the other. If correlation coefficient 'r' = + 1, the correlation is perfect and positive and if r = –1, correlation is perfect and negative.

A zero correlation indicates that there is no correlation between the variables. In other words, when there is zero correlation, it implies that there is no relationship between the two variables and any change in one variable is not associated with change in the other variable.

In practice, the following points may be kept in view. Correlation is very low if the Pearson Correlation Coefficient has a value under 0.20 and as low if the value ranges between 0.21 and 0.40. Any correlation coefficient value of above 0.70 is considered high.

4.3.1 Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation is a quantitative measure of the degree of relationship between two variables. If the variables are 'x' and 'y', it is given by the following formula:

$$\rho x, y = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

The numerator stands for covariance between X and Y while denominator stands for the product of standard deviation of x and y. For practical purposes, it is given by the following formula:

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

There are a number of assumptions underlying the use of Karl Pearson's correlation coefficient. Firstly, the variables 'x' and 'y' are linearly related. The forces operating on each of the variable series are not independent of each other but are related in causal fashion. There should be cause and effect relationship between different forces operating on the items of the two variable series. They must also be common to both the series. 'If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation between:

(a) The series of heights and incomes of individuals over a period of time,

(b) The series of marriage rate and the rate of agricultural production in a country over a period of time

(c) The series relating to the size of the shoe and intelligence of a group of individuals

Should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other. However, if in any of the above cases the value of correlation coefficient 'r' for a given set of data is not zero, then such correlation is termed as chance correlation or spurious or nonsense correlation.'

Example 9:

Given below are details of sales revenue and advertisement expenses of a company (in lakhs) in last 6 years. Examine whether there is any correlation between advertisement and sales revenue.

Year	2005	2006	2007	2008	2009	2010
Ad. Expenses	3	4	6	10	15	20
Sales revenue	30	34	40	50	70	90

Calculation of correlation coefficient:

	X	y	x^2	y^2	xy
	3	30	9	900	90
	4	34	16	1156	136
	6	40	36	1600	240
	10	50	100	2500	500
	15	70	225	4900	1050
	20	90	400	8100	1800
Total	**58**	**314**	**786**	**19156**	**3816**

$$r = 6(3816) - (58)(314) / \sqrt{\left(6(786) - (58)^2\right)\left(\sqrt{6(19156)} - (314)^2\right)}$$

$$= \frac{22896 - 18212}{\left[\sqrt{(4716 - 3364)\sqrt{114936 - 98596}}\right]}$$

NOTES

$= 4684/36.77\,(127.83)$

$= 0.9965$

In other words, there is a high degree of positive correlation between advertisement expenses and sales revenue as the value of correlation coefficient 'r' is close to +1.

4.3.2 Spearman's Rank Correlation

Let us take two characteristics A and B (beauty and intelligence) which are not quantifiable but a group of individuals could be ranked in terms of their possession of above characteristics. In such a case, is it possible to find out whether there is any correlation between A and B? Suppose we have ranks of n individuals x_1, $x_2,\ldots,x_n$ and $y_1, y_2\ldots,y_n$, based on two characteristics A and B, then Pearson's coefficient of correlation between the ranks x_i and y_i is called the rank correlation coefficient between A and B for that group of individuals. This is given by the formula:

$$\Sigma = 1 - \sum_{i}^{n} \frac{d_i^2}{2n\sigma} X^2 = 1 - \frac{6\sum_i^n d_i^2}{n\left(n^2-1\right)}$$

$$d_i = (x_i - \bar{x}) - (y_i - \bar{y})$$

Spearman's formula is the only ideal formula for determining correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially as per their ranks. In case of extreme observations too, Spearman's rank correlation coefficient is preferred to that of Pearson's formula.

However, in practice, there could be cases in which some of the individuals receive the same rank In a ranking of merit (tied ranks). If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B or if there is more than one item with the same value in the series, there is a slight change is required in the above formula.

Example 10:

The ranks of 6 students in Sociology and Statistics are as follows: (1,6), (2,3), (3,4), (4,5), (5,2), (6,1). Calculate the rank correlation coefficient between a student's proficiency in Sociology and Statistics.

Rank in Sociology x	Rank in Statistics y	d = x – y	d^2
1	6	–5	25
2	3	–1	1
3	4	–1	1
4	5	–1	1
5	2	3	9
6	1	5	25
Total			**62**

Rank correlation is given by

$1 - 6\,(62)/\,6(35) = -0.77$

In other words, there exists negative correlation between ranks obtained by students in Sociology and Statistics.

'Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If the assumption is violated, then we need a measure which is distribution free (non parametric). Spearman's correlation is such a measure (i.e., distribution-free) since no strict assumptions are made about the form of the population from which sample observations are drawn.'

Check Your Progress

5. When is correlation said to be diverse or negative?
6. Determine Karl Pearson's Coefficient of Correlation for the following data and interpret the result

X:	4	6	9	12	18	22	33	104
Y:	21	18	15	13	11	10	6	94

7. When is Spearman's rank correlation coefficient preferred to that of Pearson's formula?

4.4 CHI-SQUARE TEST

Let us assume that we have two groups, viz., students studying in government schools and those studying in private schools. Obviously we wish to know whether there exist any significant differences between these two groups in terms of their educational attainment. The two groups could even be a control group and an experimental group. Statistics enables us to interpret whether any observed differences between these groups are significant, or otherwise. This gives rise to two issues: What is the probability that a difference of a particular size could have occurred by chance? When is a particular difference termed as significant? The field of statistics enables us to find out when differences can be termed 'statistically significant'.

Often answers to the above questions are related to the size of the differences between the groups being compared, the variation of the measure (scores) within each group (i.e., variance or standard deviation) and the size of the difference between the groups (i.e., difference between the means). In the field of testing of hypothesis, the null hypothesis normally takes the form that the groups are alike and that differences observed occurred purely by chance. If the analyst concludes that the groups come from different populations, which means rejecting the null hypothesis, he or she wishes to be sure that there is only a small probability that he