UNIT 5  BASIC MEASURES OF DATA MANAGEMENT AND DATA ANALYSIS

Structure

5.1  Introduction
5.2  Objectives
5.3  Data Management and Analysis
5.4  Frequency Distribution and Percentages
5.5  Measures of Central Tendency
5.6  Role of Statistics in Research
5.7  Let Us Sum Up
5.8  Unit End Questions
5.9  Glossary
5.10 References
5.11 Suggested Readings

5.1  INTRODUCTION

In Units 3 and 4 of this Block, we have studied how survey is used in quantitative research and how various macro sources of data are useful in quantitative research. Both quantitative and qualitative data will need to be sorted, coded and categorized so that the collected data is ready for analysis. In this unit we will understand how the quantitative data that has been collected in a research study must be managed and analysed. Data processing and analysis is necessary to present them in a way that can be easily understood by other researchers and readers.

The presentation of data using tables and statistical measures aid in understanding the main findings of the study. Data analysis is useful to present patterns of relationship that exist between groups selected for study. The purpose of data management and data analysis is to summarise the collected data and organise them in a logical manner to answer the research questions raised in a study. In this unit, you will learn the process of data management and analysis and the role of statistics in research. Data which have been gathered from macro databases or through field studies can be processed for analysis by following the steps presented in this unit. You will also study data analysis. The unit ends will a discussion on role of statistics in research. Let us look at the objectives of this unit.
5.2 OBJECTIVES

After studying this unit, you should be able to:
• Describe the management and processing of data;
• Explain the procedure of classification of data;
• Learn tabulation of data;
• Comprehend the measures of central tendency; and
• Explain the role of statistics in research.

5.3 DATA MANAGEMENT AND ANALYSIS

The raw data that is gathered through survey or from macro data bases is processed and managed so that it is suitable for analysis. Based on the objectives of the research study, researchers have to format data files. Data management involves several tasks performed by researchers such as sorting, appending, merging and collapsing data so that the data are easier to handle and use.

Data are sorted by putting observations in a specific order. For example, data can be sorted according to a month or year. If researchers need more than one data file to perform data analysis, then both data files are combined into a single file. This is done by appending the observations of the second file to the first file. If the researcher finds that two files have the same observations, but under different variables, then the files are merged for analysis.

The process of data management involves editing, coding, classification and tabulation of the collected data. Let us read about each of these in detail.

• **Editing**: Editing of data is a process of examining the collected raw gathered in surveys to detect any errors, omissions and contradictions. Editing of data is also referred to as cleaning of data. For example, if the study has school girls as the respondents, we would expect the ages in a specific range. If the data by error shows age of a respondent that is not compatible to the range then it must be checked. Cleaning is also useful in checking inconsistency in data where a response is not likely or impossible. For instance, illiterate persons would not state that they read newspapers. Missing values or responses not given are also detected during data cleaning.

Editing or cleaning of data ensures that the data are as accurate as possible, consistent with other data gathered and entered in a uniform manner to facilitate the process of coding. Such data are corrected where possible and managed for coding. For instance, if the respondent
in a study of teenagers gives an age that does not fall into the specific age range such as thirteen to nineteen, then it must be verified and a new age entered for coding.

- **Coding:** Coding of data is defined as the process of conceptualizing research and classifying them into meaningful and relevant categories. Coding is done to facilitate the purpose of data analysis and interpretation. It is the process of assigning numerals or other symbols to the responses to questions that can be grouped into specific categories or classes. All the data must fall into a specific category. For example, if respondents must be classified according to their gender, all males are coded as 1 and all females are coded as 2. So the categories 1 and 2 are mutually exclusive which means that data can be coded as either 1 or 2. The coding of data can be done manually by transferring data into a coding sheet. But, large amount of data can be coded using computer facility into an excel sheet or spread sheets. The process of coding is done by a coder or the researcher in smaller studies and by a team of coders in a relatively large research study.

- **Classification:** Classification is the next step in management of data. The data that has been coded is arranged into specific groups or classes based on some common characteristic. Data are classified based on attributes or class intervals. For example, women are classified as married or unmarried based on an attribute, marital status. Women can be classified according to their age and income based on class intervals. For example, women who have an income between Rs. 1000 to Rs. 5000 are classified into one group while those with an income between Rs. 5001 and 7000 belong to another group.

- **Tabulation:** When the data have been classified into specific groups they are arranged in a logical order into tables. This process of arranging data into tables is called tabulation. Data are displayed in the compact form of tables for readers to easily understand the results of the study. Tabulation reduces the need for lengthy description and explanation of data. It enables easy comparison of data and further statistical analysis. The process of tabulation can be done manually by hand or with the help of computer. Hand tabulation can be carried out for smaller studies. But in larger studies which result in relatively greater amounts of quantitative data, it is only feasible when computers are used for tabulation.

There are a few guidelines to be followed while constructing tables:

- Tables must have a simple, brief and clear title for the reader to understand the type of data being presented in it.
- Titles of tables must be placed above the body of the table.
- Tables must be numbered consecutively for easy search.
- References or sources of data and brief notes of explanation can be provided below the table.
- The data on a table can be presented in rows and columns for easy reading. Rows and columns must also be given brief titles.
- The totals of rows must be placed in the extreme right column and the totals of columns must be placed at the bottom.

We have learnt that coded data are classified according to attributes or class intervals. The classification of data is done depending on the type of data that is being presented in the table. The classification of data according to attributes is done on the basis of some common characteristic which is either qualitative or quantitative. (We have studied qualitative and quantitative variables in Unit 1 of this Block)

The classification based on such descriptive or numerical attributes is called classification according to attributes. The analysis of the data classified according to attributes is called as the statistics of attributes. Data analysis obtained through the statistics of attributes is widely used in social sciences like Psychology, Sociology, Political Science, Media Studies, and Gender Studies.

The classification of data according to attributes can be simple where only one attribute is considered and the data divided into two groups. For example, the respondents in a study are classified as either male or female. There is also a complex classification of data according to attributes in which the data are classified into a number of groups and classes based on the presence of two or more attributes. For example, data can be classified according to sex with different categories of income.

Take up the following exercise to assess your understanding of the sections that you have just finished reading.

**Check Your Progress:**

i) What is understood by editing, coding, tabulation and classification of data?
2) Correct the following statements.
   a) The categories for coding of data are inclusive.
   b) This process of arranging data into rows is called tabulation.
   c) Researches are classified based on attributes or class intervals.
   d) Age, height, weight, income are qualitative data.
   e) Tabulation induces the need for lengthy explanation of data

In the following section we will read about statistical tools that can be applied to large quantity of data collected from the field.

5.4 FREQUENCY DISTRIBUTION AND PERCENTAGES

Frequency distribution is an arrangement in which the frequencies or percentages of the measured variable are shown. It is the most important method of describing quantitative data. Quantitative data that has been gathered in a research study are arranged by the researcher in a logical and accurate order. Once the data are arranged they begin to show a pattern which can be described by the researcher.

A typical frequency distribution table consists of two columns. This can be observed in table 5.1. The first column on the left contains all the values of the variable that is being studied. The second column on the right shows the number of times each value occurs in the data set. The sum of the frequency column is the number (N) of persons or items that make up the distribution.

A frequency distribution can also be constructed using grouped intervals. Data are divided into a number of groups or classes which are called class intervals. Each group or class interval has an upper limit and a lower limit which are called as class limits. The difference between the two class limits is known as class magnitude.

The number of data items that fall into a class is called frequency of that class. The frequency of a class is calculated by hand or using a tally sheet. All the class intervals are written on a tally sheet and a small stroke is made for each item in the corresponding class. For every four lines in a class interval, the fifth line is indicated by a cross stroke (||||) in the tally sheet (see table 5.1).
Table 5.1: Tally Sheet showing the number of women placed in different income groups

<table>
<thead>
<tr>
<th>Income Groups (Income in Rupees)</th>
<th>Tally Marks</th>
<th>Number of Women (frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 1000</td>
<td>I I I I I I</td>
<td>16</td>
</tr>
<tr>
<td>1001-2000</td>
<td>I I I I I I I</td>
<td>8</td>
</tr>
<tr>
<td>2001-3000</td>
<td>I I I I I I</td>
<td>11</td>
</tr>
<tr>
<td>3001-4000</td>
<td>I I I I I</td>
<td>7</td>
</tr>
<tr>
<td>4001-5000</td>
<td>I I I I</td>
<td>6</td>
</tr>
<tr>
<td>5001 and above</td>
<td>I I</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

All the groups or classes with their respective frequencies which are placed in a table are referred to as a frequency distribution. The data now begin to show a pattern which can be interpreted in data analysis. Frequency of data can also be expressed in terms of percentages. Percentage of a response can be calculated by dividing the frequency of the response by the total number of responses in the distribution.

Researchers can display the frequency data by putting it either in tables or graphs. Graphs consist of two perpendicular lines, the x-axis (horizontal) and the y-axis (vertical). The scores are presented along the x-axis and the corresponding frequency along the y-axis. A common form of graph is the histogram or bar chart. The frequencies are represented by vertical bars in a histogram (see figure 5.1).

In figure 5.1 the histogram represents the case of six women with daily income in rupees. For example, the first woman earns Rs. 100 daily, second woman earns Rs. 200, third woman earns Rs. 300, fourth woman earns Rs. 400, the fifth woman earns Rs. 500 and the sixth woman earns Rs. 350 per day.
A **pie chart** is a circle graph divided into various segments wherein each segment displays the relevant information proportional to its size. Figure 5.2 shows the graphical representation of the number of women and their education levels in a pie chart. For example, 100 women are illiterate (category 1), 200 women have primary education (category 2), 300 women have completed high school (category 3), 400 women have completed matriculation (category 4), 500 women have intermediate college education (category 5) and 350 women are graduates (category 6).

![Pie Chart](image)

**Figure 5.2: Graphical Representation-Pie Chart**

In the case of data generated by large surveys, frequency distribution and percentages are calculated using computers. In smaller studies, the tally sheet can be used and frequency distribution done by hand.

Let us now read about another important statistical application on large data.

### 5.5 MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called **measures of central location**. They are also called as summary statistics. Measures of central tendency are also referred to as **statistical averages**. Measures of central tendency provide us the typical score that is characteristic of an entire distribution. They tell us the score around which most of the items have a tendency to cluster. They are considered the most representative score for the entire data. The most widely used measures of central tendency which are discussed here are mode, median and mean.
The mode is defined as the score or value that occurs most frequently or commonly in a distribution. For example, if the scores in a given distribution are:

26 28 30 32 35 36 36 36 36 37 38 38 39 41 42

The mode in the above distribution would be 36 because the score 36 occurs four times, more than any other score. The mode is determined by finding an attribute that is most observed. It can be applied to both quantitative and qualitative data. Mode is easy to determine but it pays attention to only one score. This can hide more important facts about the data. For the data in table 5.2 the mode is 45 as it is most widely represented value (14).

Table 5.2: Mode as a Measure of Central Tendency

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>14</td>
</tr>
<tr>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>43</td>
<td>10</td>
</tr>
<tr>
<td>42</td>
<td>10</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
</tr>
<tr>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>10</td>
</tr>
</tbody>
</table>

Let us now read about how to find out mode of a grouped data. (IGNOU, 2005, MSO 002, Book -2, pp. 53-54)

1) Identify the class with maximum frequency. This is the ‘modal class’.

2) Apply the formula: 
   \[ \text{Mode} = L + \left[\frac{(f_m - f_l)}{(f_m - f_l) + (f_m - f_2)}\right] \times i \]
   
Or
Quantitative Research

Mode = L + [(fm - f1)/(2fm - f1 - f2)]i

Where, L = the lower limit of the modal interval, fm = frequency of the modal class, f1 = frequency of the class preceding modal class, f2 = frequency of the class succeeding modal class, i = the size of the class interval (i = upper limit of the class - lower limit of the class).

For example: Find the Modal Income on the Basis of the following data

<table>
<thead>
<tr>
<th>Income (in thousand)</th>
<th>5-10</th>
<th>10-15</th>
<th>15-20</th>
<th>20-25</th>
<th>25-30</th>
<th>30-35</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Households</td>
<td>8</td>
<td>16</td>
<td>29</td>
<td>22</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

Substituting the values in the formula, we see

Mode = L + [(fm - f1)/(2fm - f1 - f2)]i

Mode = 15 + [(29-16)/(2*29-16-22)]*5
= 15 + (13/21)*5
= 15 + 3.33
= 17.33

Thus, the modal income is 18.33 thousands.

Box No.5.1

The mode is not affected by extreme values in a distribution. It gives an idea of the score to indicate where the observed data are concentrated in a frequency distribution. The mode does not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set.

Let us look at another measure of central tendency that is, median.

- The median, second measure of central tendency, is the midpoint of a distribution. In a frequency distribution, half of the scores lie above the median and half of the scores lie below it. In order to determine the median, the scores must be arranged from the smallest to the largest value to locate the midpoint. In other words the scores
must be arranged in the ascending or descending order. For example, the scores in a distribution are:

\[ 4 \ 5 \ 6 \ 8 \ 10 \ 12 \ 14 \ 17 \ 18 \ 20 \ 22 \ 23 \ 24 \]

The median score in the above distribution is 14, since there are six scores below this number and six above it.

We can also write the median using the formula given below:

\[ \text{Median} (M) = \text{Value of } \frac{[n+1] \text{th item}}{2} \]

Now consider these scores:

\[ 2 \ 3 \ 5 \ 7 \ 9 \ 12 \ 13 \ 15 \ 17 \ 18 \ 19 \ 21 \]

No score neatly divides this distribution. To determine the median, the two middle scores must be added and divided by 2.

\[ \text{Median} = \frac{12+13}{2} = 12.5 \]

Calculating median from grouped data (Adopted from IGNOU, 2008, MSO 002, Book -2, p.51)

Step 1: Divide the total number of items (\(N\) or \(\sum F_i\)) by two.

Step 2: Start at the low end of the frequency distribution and sum the scores in each interval until the interval containing the median is reached (C. F.).

Step 3: Subtract the sum obtained in step two from the number necessary (calculated at step 1) to reach the median (\(N/2 - \text{C. F.}\)).

Step 4: Now calculate the proportion of the median interval that must be added to its lower limit in order to reach the median score. This is done by dividing the number obtained in step 3 by the number of scores (\(f\)) in the median interval and then multiplying by the size of the class interval (\(i\)), i.e.

\[ \left(\frac{N}{2} - \text{C. F.}\right) / f \times i \]

Step 5: Finally, add the number obtained in step 4 above to the exact lower limit of the median interval.

\[ \text{Median} = L + \left(\frac{N}{2} - \text{C. F.}\right) / f \times i \]

Where, \(L\) = the lower limit of the median interval,
\(N\) = the total number of scores;
\(C. F.\) = the sum of the scores in the intervals below the median interval,
\(f\) = the number of scores in the median interval;
\(i\) = the size of the class interval.
Given below is an example for finding the median for a grouped data.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-21</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21-24</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>24-27</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>27-30</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>30-33</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>33-36</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>36-39</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>39-42</td>
<td>6</td>
<td>41</td>
</tr>
<tr>
<td>42-45</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>45-48</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>48-51</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>ΣFi = 50</td>
<td></td>
</tr>
</tbody>
</table>

Median = L + [(N / 2 - C.F.) / f] * i

N or ΣFi / 2 = 50 / 2 = 25

Lower limit of the median class (L) = 33

Cumulative frequency of the class preceding the median class (C.F.) = 19
Cumulative frequency of the class preceding the median class (C.F.) = 19

Frequency of the median class (f) = 8
Size of the class interval = 3

Median = 33 + [(25-19) / 8] * 3 = 33 + 2.25 = 35.25
The third type of central tendency measure is the mean. It is also called as arithmetic average and the commonly used measure of central tendency. The mean is defined as the value which is obtained by adding all the observations and dividing the sum by the number of the observations. In other words the mean is calculated by dividing the total of the values of given items in a frequency series by the total number of items. It is the average of a set of scores. It can be used with both discrete and continuous data, although its use is most often with continuous data (see types of variables in Unit 1).

We can work out the mean using the formula given below:
The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have \( n \) values in a data set and they have values \( x_1, x_2, \ldots, x_n \), then the sample mean, usually denoted by \( \bar{x} \) (pronounced x bar), is:

\[
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}
\]

This formula is usually written in a slightly different manner using the Greek capital letter, \( \sum \), pronounced “sigma”, which means “sum of…”:

\[
\bar{x} = \frac{\sum x}{n}
\]

Calculating Mean of a grouped data (IGNOU, 2005, MSO 002, Book -2, p.49)
Mean of a grouped data can be calculated by using the following formula

\[
M = \frac{\sum F_i \times X_i}{\sum F_i}
\]

Where, \( M \) is the mean, \( X_i \) are the midpoint of class intervals, \( F_i \) are the number of cases in various intervals, \( \sum F_i \) is the total number of scores or sum of frequencies of various intervals. Read the example given below in the box to understand how to calculate Mean of a grouped data.

Following is the frequency (8, 9, 12, 9, 7, and 5) of households in a community owning numbers of chickens, arranged in six groups (1-3, 4-6, 7-9, 10-12, 13-16 and 16-18).

<table>
<thead>
<tr>
<th>Number of Chickens</th>
<th>Mid- Point of the Interval ((X_i))</th>
<th>Frequency: Number of Households ((F_i))</th>
<th>(F_i \times X_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>2</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>4-6</td>
<td>5</td>
<td>9</td>
<td>45</td>
</tr>
<tr>
<td>7-9</td>
<td>8</td>
<td>12</td>
<td>96</td>
</tr>
<tr>
<td>10-12</td>
<td>11</td>
<td>9</td>
<td>99</td>
</tr>
<tr>
<td>13-15</td>
<td>14</td>
<td>7</td>
<td>98</td>
</tr>
<tr>
<td>16-18</td>
<td>17</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>376</td>
</tr>
</tbody>
</table>
A short method of calculating mean from grouped data

\[
\text{Mean (M)} = \text{AM} + \left(\frac{\sum F_i \times D_i}{\sum F_i}\right)i
\]

Also,

\[
D_i = \frac{(\text{AM} - X_i)}{i}
\]

Where, M is the mean, AM = Assumed mean, \(X_i\) are the midpoint of class intervals, \(F_i\) are the number of cases in various intervals, \(\sum\) is the symbol of sum total, \(D_i\) are the deviations of the midpoints of the various classes from the midpoints of the class having the assumed mean divided by the size of the class interval (equation 4) and i is the size of the class intervals.

You will notice that the mean is not often one of the actual values that you have observed in your data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. Unlike the mode and the median, the mean takes into account all the values in a frequency distribution. The calculation of the mean includes the effect of extreme values in a distribution. The mean may be considered as the score that could be assigned to each individual if the total is evenly distributed among all members of the sample. The data used to calculate the mean should be at the interval or ratio level. You have already studied interval and ratio data of measurement in Unit 1 of this block.

Box No. 5.2

The median is often used to summarize the location of a distribution. It is considered the most appropriate measure of central tendency in the case of qualitative data. Where there are extremely large or small observations in the case of a distribution with quantitative data, the median provides a better measure of location than the mean. We usually prefer the median over the mean (or mode) when our data is skewed (i.e., the frequency distribution for our data is skewed).

If we consider the normal distribution, then the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data as the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values.
We often test whether our data is normally distributed as this is a common assumption underlying many statistical tests. An example of a normally distributed set of data is presented below:

![Histogram of normally distributed data](image)

**Figure 5.4:**

When you have a normally distributed sample you can legitimately use both the mean or the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of central tendency as it is the measure that includes all the values in the data set for its calculation, and any change in any of the scores will affect the value of the mean. This is not the case with the median or mode.

However, when our data is skewed, for example, as with the right-skewed data set as figure 5.5:

We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean.
When a researcher has to decide which of the three measures of central tendency is useful to describe a set of data, two main factors must be considered.

- First of all, the level of measurement used may determine the measure of central tendency to be chosen. We have studied the levels of measurement in Unit 1. If the data are at a nominal level only, mode is meaningful. If the data are at the ordinal level, either mode or median may be used. All the three measures of central tendency are appropriate for interval and ratio data.
- The second factor that decides the choice of the measure of central tendency is the purpose it or which

To summarize, we can learn when to use the mean, median and mode using the following table to know what the best measure of central tendency is with respect to the different types of variable.

<table>
<thead>
<tr>
<th>Type of Variable</th>
<th>Best measure of central tendency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Mode</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Median</td>
</tr>
<tr>
<td>Interval/Ratio (not skewed)</td>
<td>Mean</td>
</tr>
<tr>
<td>Interval/Ratio (skewed)</td>
<td>Median</td>
</tr>
</tbody>
</table>
5.6 ROLE OF STATISTICS IN RESEARCH

The role of statistics in research lies in facilitating the process of data management and data analysis. Statistics helps to design quantitative research, analysing data and drawing conclusions. In other words, statistics provides mathematical models for organizing, summarizing and analysing data that have been collected and measured in a research study. It is important to understand statistics to classify and tabulate data. The knowledge of statistics is useful to summarise classified data. We can also develop indices to measure a construct. We have studied concepts and constructs in Unit 1 of this Block.

There are two types of research statistics: descriptive and inferential statistics.

- Descriptive statistics are intended to reduce the data sets for easier interpretation. For example, in a study of 100 women to find out ownership of property among them, it would be difficult to tell how many of them own property just by looking at all the data sheets. We can draw a conclusion only if data are organized to give results. Data collected during a research study have little meaning or usefulness until they are displayed or summarized using descriptive statistics. The two commonly used methods of data management are data distribution and summary statistics. We have studied data distribution using frequency distribution and summary statistics such as measures of central tendency.

- Researchers are interested in results that can be stated or generalized for the population from which the sample was drawn. The statistics used to determine the degree to which the results can be generalized from the sample to the population is called inferential statistics. For instance, in a study of girls dropping out from school and early marriage, the relationship between these two factors could be tested using methods of inferential statistics. The results could then be generalized to the population of girls in that area.

It is difficult to perform data analysis manually in relatively large studies. Manual data analysis can become a tiring and cumbersome process often leading to errors in calculations. Huge amounts of data are analysed using computers and statistical analysis software. There are several software packages such as Statistical Package for Social Sciences (SPSS), Minitab, Statistical Analysis System (SAS), Stata and S-Plus which are available for statistical analysis. Minitab is an easy-to-use package, used interactively on many computer systems to provide data manipulation and statistical analysis for small to medium sized datasets. Data are held in a temporary worksheet of rows and columns. SAS (Statistical Analysis System) is a general purpose
data analysis system for managing data and analyzing it statistically. Stata is a complete, integrated statistical package that provides everything needed for data analysis, data management, and graphics. S-Plus provides functionality in a wide variety of areas, including statistical tests, quality control, and graphics. Microsoft Excel can also be used to perform most of the statistical analysis except in the case of some advanced statistical models and tests.

The most popular quantitative data analysis software widely used in social science research using computers is the Statistical Package for Social Sciences commonly referred to as SPSS (You have read about SPSS in Block 2 Unit 4). SPSS is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions. It is a comprehensive statistical package for generating distribution charts and tables. SPSS can perform simple to complex data analysis. We can use SPSS in calculating descriptive statistics and presenting a display of the data in the form of tables, charts and graphs. SPSS is also capable of inferential statistical analysis. It provides a window user interface which makes the software user-friendly for many beginners.

But knowledge of the software alone is not enough, a sound understanding of basic statistical concepts and techniques of analysis is needed. Nearly all these packages are menu driven which means that the instructions for operating the programme are shown on the screen. The user can easily follow the set of instructions to carry out data analysis. You will study data analysis techniques in detail in Unit 2 of Block 5.

**Check Your Progress**

1) How is a frequency distribution drawn in a research study?
ii) What are the different measures of central tendency?

iii) Correct the following statements.

   a) Frequency is divided into a number of groups or classes called class intervals.

   b) The difference between the two class limits is known as class frequency.

   c) The mean is the midpoint of a distribution.

   d) Mode is the score or value that is unique in a distribution.

   e) There are two types of researches: descriptive and inferential

Revisit what you have learnt in this unit by going through the following section

5.7 LET US SUM UP

Data which have been gathered from research studies can be processed for analysis by following the steps of data management. The main purpose of data management and data analysis is to summarise the collected data and organise them in a logical manner to answer the research questions raised in a study. The process of data management involves editing, coding, classification and tabulation of the collected data. Data must be sorted, coded and categorized so that it is ready for analysis.

There are two types of research statistics: descriptive and inferential statistics. Descriptive statistics are intended to reduce the data sets for easier interpretation. Descriptive statistics include measures of central tendency. The most widely used measures of central tendency are mode, median and mean. These measures provide us the typical score that is characteristic of an entire distribution. Inferential statistics allow a researcher to generalize sample results to a population.
Data analysis in relatively large studies is usually done using software packages. There are several software packages such as SPSS, Minitab, SAS, Stata, S-Plus and Microsoft Excel are available for statistical analysis. SPSS is a popular and widely used package for statistical analysis. It is useful for generating distribution charts and tables. SPSS can perform simple to complex data analysis.

A sound understanding of basic statistics is needed along with the knowledge of using statistical packages for data analysis. There has little attempt in this unit to explain the mathematical principles and formulas of measures of central tendency. The emphasis has been on providing an understanding of the reason behind using these statistics for data analysis.

5.8 UNIT END QUESTIONS

1) Discuss the process of data management and the importance of the various processes.

2) Why is tabulation needed in data management? What are the characteristics of a good table?

3) Discuss the process of creating frequency distribution while classifying data.

4) Describe the various measures of central tendency.

5) Explain the role of statistics in research.

6) Describe the various processes involved in data management and data analysis in a research study.

5.9 GLOSSARY

Descriptive statistics : Statistical methods and techniques designed to reduce data sets to allow for easy interpretation.

Frequency distribution : A collection of scores or measurements ordered according to their magnitude and their respective frequencies.

Inferential statistics : Statistical methods and techniques that allow a researcher to generalize sample results to a population.

Mean : The arithmetic average of a set of scores.

Median : The midpoint of a distribution of scores.

Mode : The score that occurs most often in a frequency distribution.
5.10 REFERENCES


New Delhi: IGNOU.


5.11 SUGGESTED READINGS

