Chapter 1
Data Management

Learning Objectives

After completing this chapter, you should be able to do the following:

- Explain different types of data generated in management research.
- Know the characteristics of variables.
- Learn to remove the outliers from the data by understanding different data cleaning methods before using in SPSS.
- Understand the difference between primary and secondary data.
- Know the formats used in this book for using different commands, subcommands, and options used in SPSS.
- Learn to install SPSS package for data analysis.
- Understand the procedure of importing data in other formats into SPSS.
- Prepare the data file for analysis in SPSS.

Introduction

In today's world of information technology, enormous data is generated in every organization. These data can help in strategic decision-making process. It is therefore important to store such data in a warehouse so that effective mining can be done later for getting answers to many of the management issues. Data warehousing and data mining are therefore two important disciplines in the present-day scenario. Research in any discipline is carried out in order to minimize inputs and effectively utilizing the human resources, production techniques, governing principles, marketing policies, and advertisement campaigns to maximize outputs in the form of productivity. To be more specific, one may be interested to identify new forms of resources, devise organizational systems and practices to motivate culturally diverse set of individuals, and evaluate the existing organizations so as to make them more productive to the new demands on them. Besides, there may be any number of other issues like

J.P. Verma, *Data Analysis in Management with SPSS Software*,
DOI 10.1007/978-81-322-0786-3_1, © Springer India 2013

effective leadership, skill improvement, risk management, customer relationships, and guiding the evolution of technology, etc., where the researcher can make an effective contribution.

A researcher may use varieties of data analysis techniques in solving their research problems like: How to *motivate* people for work? How to make a television or FM channel more popular? How to enhance the *productivity* at work? Which strategy becomes more efficient? How organizational structure promotes *innovation*? How to *measure* training effectiveness? Due to cutthroat competition, the research issues have grown in number, scope, and complexity over the years. Due to availability of computer software for advanced data analysis, researcher has become more eager to solve many of these complex issues.

The purpose of data analysis is to study the characteristics of sample data for approximating it to the population characteristics. Drawing conclusion about the population on the basis of sample would be valid only if the sample is true representative of the population. This can be ensured by using the proper sampling technique. However, large sample need not necessarily improves the efficiency in findings. It is not the quantity but the quality of the sample that matters.

Data generated in management research may be analyzed by using different kinds of statistical techniques. These techniques differ as per the nature of the study which can be classified into any of the five categories; descriptive study, analytical study, inductive study, inferential study and applied study. Choosing statistical technique in data analysis depends upon nature of the problem. It is therefore important to know the situation under which these techniques are used.

Descriptive study is used if an organization or a group of objects needs to be studied about its different characteristics. In such studies, we usually tabulate and compile the data in a meaningful manner so that the statistics like mean, variance, standard error, coefficient of variance, range, skewness, kurtosis, percentiles, etc., can be computed in different groups.

Analytical studies are used for studying the functional relationships among variables. Statistics like product moment correlation, partial and multiple correlations are used in such study. Consider a study where it is required to explore the parameters on which the sale depends. One may like to find correlation between sales data and independent variables like incentives, salesman's IQ, number of marketing hours, and advertisement campaigns. Here, correlation between the sales data and other parameters may be investigated for their significance. Thus, in all those situations where relationships are investigated between the performance parameter and other independent parameters, analytical studies are used.

Inductive studies are those studies which are used to estimate some phenomenon of an individual or of an object on the basis of the sample data. Here, the phenomenon which we estimate does not exist at the time of estimation. One may estimate company's performance in the next 3 years on the basis of some of its present parameters like EPS, P/E ratio, cash reserves, demands, and production capacity.

In inferential study, inferences are drawn about the population parameters on the basis of sample data. Regression analysis is being used in such studies. The difference

between inferential and inductive studies is that the phenomenon which we infer on the basis of the sample exists in the inferential studies, whereas it is yet to occur in the inductive studies. Thus, assessing satisfaction level in an organization on the basis of a sample of employees may be the problem of inferential statistics.

Finally, applied studies refers to those studies which are used in solving the problems of real life. The statistical methods such as times series analysis, index numbers, quality control, and sample survey are included in this class of analysis.

Types of Data

Depending upon the data types, two broad categories of statistical techniques are used for data analysis. For instance, parametric tests are used if the data are metric, whereas in case of nonmetric data, nonparametric tests are used. It is therefore important to know in advance the types of data which are generated in management research.

Data can be classified in two categories, that is, metric and nonmetric. Metric and nonmetric data are also known as quantitative and qualitative data, respectively. Metric data is analyzed using parametric tests such as t, F, Z, and correlation coefficient, whereas nonparametric tests such as sign test, median test, chi-square test, Mann-Whitney test, and Kruskal-Wallis test are used in analyzing nonmetric data.

Certain assumptions about the data and form of the distribution need to be satisfied in using parametric tests. Parametric tests are more powerful in comparison to that of nonparametric tests, provided required assumptions are satisfied. On the other hand, nonparametric tests are more flexible and easy to use. Very few assumptions need to be satisfied before using these tests. Nonparametric tests are also known as distribution-free tests.

Let us understand the characteristics of different types of metric and nonmetric data generated in research. Metric data is further classified into interval and ratio data. On the other hand, nonmetric data is classified into nominal and ordinal. The details of these four types of data are discussed below under two broad categories, namely, metric data and nonmetric data, and are shown in Fig. 1.1.

Metric Data

Data is said to be metric if it is measured at least on interval scale. Metric data are always associated with a scale measure, and, therefore, it is also known as scale data or quantitative data. Metric data can be measured on two different types of scale, that is, *interval* and *ratio*.

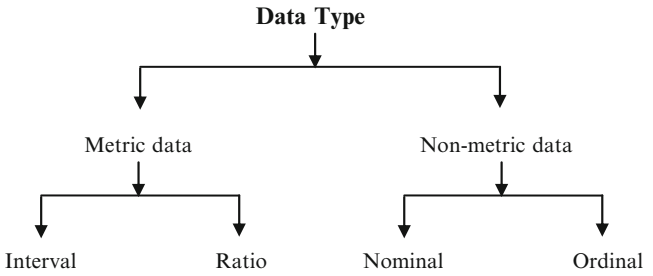


Fig. 1.1 Types of data and their classification

Interval Data

The interval data is measured along a scale where each position is equidistant from one another. In this scale, the distance between two pairs is equivalent in some way. In interval data, doubling principle breaks down as there is no zero on the scale. For instance, the 6 marks given to an individual on the basis of his IQ do not explain that his nature is twice as good as the person with 3 marks. Thus, interval variables measured on an interval scale have values in which differences are uniform and meaningful, but ratios are not. Interval data may be obtained if the parameters of job satisfaction or level of frustration is rated on scale 1–10.

Ratio Data

The data on ratio scale has a meaningful zero value and has an equidistant measure (i.e., the difference between 30 and 40 is the same as the difference between 60 and 70). For example, 60 marks obtained on a test are twice of 30. This is so because zero can be measured on ratio scale. Ratio data can be multiplied and divided because of an equidistant measure and doubling principle. Observations that we measure or count are usually ratio data. Examples of ratio data are height, weight, sales data, stock price, advance tax, etc.

Nonmetric Data

Nonmetric data is a categorical measurement and is expressed not in terms of numbers but rather by means of a natural language description. It is often known as "categorical" data. Examples of such data are like employee's category = "executive," department = "production," etc. These data can be measured on two different scales, that is, nominal and ordinal.

Nominal Data

Nominal data is a categorical variable. These variables result from a selection in categories. Examples might be employee's status, industry types, subject specialization, race, etc. Data obtained on nominal scale is in terms of frequency. In SPSS,[1] nominal data is represented as "nominal."

Ordinal Data

Variables on the ordinal scale are also known as categorical variables, but here the categories are ordered. The order of items is often defined by assigning numbers to them to show their relative position. Categorical variables that assess performance (good, average, poor, etc.) are ordinal variables. Similarly, attitudes (strongly agree, agree, undecided, disagree, and strongly disagree) are also ordinal variables. On the basis of the order of an ordinal variable, we may not know which value is the best or worst on the measured phenomenon. Moreover, the distance between ordered categories is also not measureable. No arithmetic can be done with the ordinal data as they show sequence only. Data obtained on ordinal scale is in terms of ranks. Ordinal data is denoted as "ordinal" in SPSS.

Important Definitions

Variable

A variable is a phenomenon that changes from time to time, place to place, and individual to individual. Examples of variable are salary, scores in CAT examination, height, weight, etc. The variables can further be divided into discrete and continuous. *Discrete variables* are those variables which can assume value from a limited set of numbers. Examples of such variables are number of persons in a department, number of retail outlets, number of bolts in a box, etc. On the other hand, *continuous variables* can be defined as those variables that can take any value within a range. Examples of such variables are height, weight, distance, etc.

[1] SPSS, Inc. is an IBM company which was acquired by IBM in October, 2009.

Attribute

An attribute can be defined as a qualitative characteristic that takes sub-values of a variable, such as "male" and "female," "student" and "teacher," and married and unmarried.

Mutually Exclusive Attributes

Attributes are said to be mutually exclusive if they cannot occur at the same time. Thus, in a survey, a person can choose only one option from a list of alternatives (as opposed to selecting as many that might apply). Similarly in choosing the gender, one can either choose male or female.

Independent Variable

Any variable that can be manipulated by the researcher is known as independent variable. In planning a research experiment, to see the effect of low, medium, and high advertisement cost on sales performance, advertisement cost is an independent variable as the researcher can manipulate it.

Dependent Variable

A variable is said to be dependent if it changes as a result of change in the independent variable. In investigating the impact on sales performance by the change in advertisement cost, the sales performance is a dependent variable, whereas advertisement cost is an independent variable. In fact, a variable may be a dependent variable in one study and independent variable in some other study.

Extraneous Variable

Any additional variable that may provide alternative explanations or cast doubt on conclusions in an experimental study is known as extraneous variable. If the effect of three different teaching methods on the student's performance is to be compared, then the IQ of the students may be termed as extraneous variable as it might affect the learning efficiency during experimentation if the IQs are not same in all the groups.

The Sources of Research Data

In designing a research experiment, one needs to specify the kind of data required and how to obtain it. The researcher may obtain the data from the reliable source if it is available. But if the required data is not available from any source, it may be collected by the researcher themselves. Several agencies collect data for some specified purposes and make them available for the other researchers to draw other meaningful conclusions as per their plan of study. Even some of the commercial agencies provide the real-time data to the users with cost. The data so obtained from other sources are referred as *secondary data,* whereas the data collected by the researchers themselves are known as *primary data*. We shall now discuss other features of these data in the following sections:

Primary Data

The data obtained during study by the researchers themselves or with the help of their colleagues, subordinates, or field investigators are known as primary data. The primary data is obtained by the researcher in a situation where relevant data is not available from the reliable sources or such data do not exist with any of the agency or if the study is an experimental study where specific treatments are required to be given in the experiment. The primary data is much more reliable because of the fact that the investigator himself is involved in data collection and hence can ensure the correctness of the data. Different methods can be used to collect the primary data by the researcher. These methods are explained below:

By Observation

The data in this method is obtained by observation. One can ensure the quality of data as the investigator himself observes real situation and records the data. For example, to assess the quality of any product, one can see as to how the articles are prepared by the particular process. In an experimental study, the performance of the subjects, their behavior, and other temperaments can be noted after they have undergone a treatment. The drawback of this method is that sometimes it becomes very frustrating for the investigator to be present all the time for collecting the data. Further, if an experiment involves the human being, then the subjects may become conscious in the presence of an investigator, due to which performance may be affected which will ultimately result in inaccurate data.

Through Surveys

This is the most widely used method of data collection in the area of management, psychology, market research, and other behavioral studies. The researcher must try to motivate respondents by explaining them the purpose of the survey and impact of their responses on the results of the study. The questionnaire must be short and must hide the identity of respondents. Further, the respondent may be provided reasonable incentives as per the availability of the budget. For instance, a pen, a pencil, or a notepad with print statements like "With best Compliments from. . ." or "With Thanks. . .," "Go Green," "Save Environment" may be provided before seeking their opinion on the questionnaire. You can print your organization name or your name as well if you are an independent researcher. The first two slogans may promote your company as well, whereas the other two convey the social message to the respondents. The investigator must ensure the authenticity of the collected data by means of cross-checking some of the sampled information.

From Interviews

The data collected through the direct interview allows the investigator to go for in-depth questioning and follow-up questions. The method is slow and costly and forces an individual to be away from the job during the time of interview. During the interview, the respondent may provide the wrong information if certain sensitive issues are touched upon, and the respondent may like to avoid it on the premise that it might suffer their reputation. For instance, if the respondent's salary is very low and the questions are asked about his salary, it is more likely that you end up with the wrong information. Similarly in asking the question, as to how much you invest on sports for your children in a year, you might get wrong information due to the false ego of respondent.

Through Logs

The data obtained through the logs maintained by the organizations may be used as primary data. Fault logs, error logs, complaint logs, and transaction logs may be used to extract the required data for the study. Such data provide valuable findings about system performance over time under different conditions if used well, as they are empirical data and obtained from the objective data sources.

Primary data can be considered to be reliable because you know how it was collected and what was done to it. It is something like cooking yourself. You know what went into it.

Secondary Data

Instead of data obtained by the investigator himself if it is obtained from some other sources, it is termed as secondary data. Usually, companies collect the data for some specific purpose, and after that, they publish it for the use of the researchers to draw some meaningful conclusions as per their requirements. Many government agencies allow their real-time data to the researchers for using in their research study. For instance, census data collected by the National Sample Surveys Organization may be used by the researchers for getting several demographic and socio-economic information. Government departments and universities maintain their open-source data and allow the researchers to use it. Nowadays, many commercial agencies collect the data in different fields and make it available to the researchers with nominal cost.

The secondary data may be obtained from many sources; some of them are listed below:

- Government ministries through national informatics center
- Government departments
- Universities
- Thesis and research reports
- Open-source data
- Commercial organization

Care must be taken to ensure the reliability of the agency from which the data is obtained. One must ensure to take an approval of the concerned department, agency, organization, universities, or individuals for using their data. Due acknowledgment must be shown in their research report for using their data. Further, data source must be mentioned while using the data obtained from secondary sources.

In making comparison between primary and secondary data, one may conclude that primary data is expensive and difficult to acquire, but it is more reliable. Secondary data is cheap and easy to collect but must be used with caution.

Data Cleaning

Before preparing the data file for analysis, it is important to organize the data on paper first. There are more chances that the data set may contain error or outlier. And if it is so, the results obtained may be erroneous. Analysts tend to waste lot of time in drawing the valid conclusions if data is erroneous. Thus, it is utmost important that the data must be cleaned before analysis. If data is clean, the analysis is straightforward and valid conclusions may be drawn.

In data cleaning, the invalid data is detected first and then it is corrected. Some of the common sources of errors are as follows:

- Typing errors in data entry
- Not applicable option or blank options are coded as "0"
- Data for one variable column is entered under the adjacent column
- Coding errors
- Data collection errors

Detection of Errors

The wrongly fed data can be detected by means of descriptive statistics computed by SPSS. Following approaches may be useful in this regard.

Using Minimum and Maximum Scores

By looking to the minimum and maximum scores of each variable in descriptive statistics, one can identify the error, if any, by knowing the acceptable limits of minimum and maximum scores of each variable. For instance, if the maximum score for the variable showing percentage of marks is 650, one must think of some typographical error while feeding the data as percentage of marks cannot be more than 100%.

Using Frequencies

Frequencies of each score obtained in descriptive statistics may be used to identify the "dirty" data among the entered variables. For instance, most of the biometric data are normally distributed, and, therefore, if any variable shows large frequency for any values, it must be checked for any systematic error.

Using Mean and Standard Deviation

Normally, the value of standard deviation is less than the mean except in case of certain distribution like negative binomial. Thus, if the standard deviation for any of the variables like age, height, or IQ is more than their mean, it can only be if some of the values of these variables are outliers. Such entries can easily be identified and removed.

Logic Checks

Errors in data may also be detected by observing as to whether the responses are logical or not? For example, one would expect to see 100% of responses, not 110%. Similarly, if a question is asked to a female employee as to whether they have

availed maternity leave so far or not and if the reply is marked "yes" but you notice that the respondent is coded as male, such logical errors can be spotted out by looking to the values of the categorical variable. Logical approach should be used judiciously to avoid the embarrassing situation in reporting the finding like 10% of the men in the sample had availed the maternity leave during the last 10 years.

Typographical Conventions Used in This Book

Throughout the book, certain convention has been followed in writing commands by means of symbol, bold words, italic words, and words in quotes to signify the special meaning. Readers should note these conventions for easy understanding of commands used in different chapters of this book.

Start ⇒ All Programs	Denotes a menu command, which means choosing the command **All Program** from the **Start** menu. Similarly **Analyze ⇒ Correlate ⇒ Partial** means open the **Analyze** menu, then open the **Correlate** submenu, and finally choose **Partial**.
Regression	Any word written in bold refers to the main command of any window in the SPSS package.
Prod_Data	Any word or combination of words written in italics form during explaining SPSS is referred as variable.
"Name"	Any word or combination of words written in quotes refers to the subcommand.
'Scale'	Any word written in single quote refers to one of the option under subcommand.
Continue	This refers to the end of selection of commands in a window and will take you to the next level of options in any computation.
OK	This refers to the end of selecting all the options required for any particular analysis. After pressing the OK invariably, SPSS will lead you to the output window.

How to Start SPSS

This book has been written by referring to the IBM SPSS Statistics 20.0 version; however, in all the previous versions of SPSS, procedure of computing is more or less similar.

The SPSS needs to be activated on your computer before entering the data. This can be done by clicking the left button of the mouse on SPSS tag by going through the SPSS directory in the **Start** and **All Programs** option (if the SPSS directory has been created in the Programs file). Using the following command sequence, SPSS can be activated on your computer system:

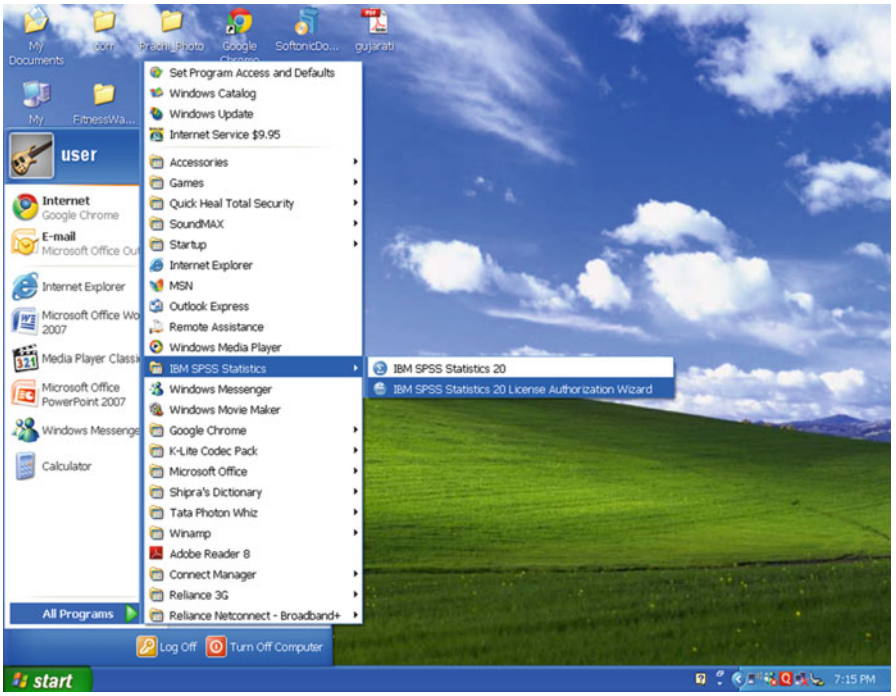


Fig. 1.2 Commands for starting SPSS on your computer

Start → All Programs → IBM SPSS Statistics → IBM SPSS Statistics 19

If you use the above-mentioned command sequence, the screen shall look like Fig. 1.2.

After clicking the tag SPSS, you will get the following screen to prepare the data file or open the existing data file.

If you are entering the data for new problem and the file is to be created for the first time, check the following option in the above-mentioned window:

And if the existing file is to be opened or edited, select the following option in the window:

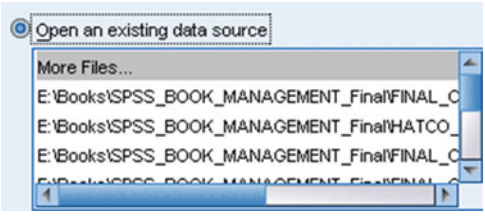


Table 1.1 FDI inflows and trade (in percent) in different states

S.N.	FDI	Exports inflows	Imports	Trade
1	4.92	4.03	3.12	3.49
2	0.07	4.03	3.12	3.49
3	0.00	1.11	2.69	2.04
4	5.13	17.11	27.24	23.07
5	11.14	13.43	11.24	12.14
6	0.48	1.14	3.41	2.47
7	0.30	2.18	1.60	1.84
8	29.34	20.56	18.68	19.45
9	0.57	1.84	1.16	1.44
10	0.03	1.90	1.03	1.39
11	8.63	5.24	9.24	7.59
12	0.00	3.88	6.51	5.43
13	2.20	7.66	1.57	4.08
14	2.37	4.04	4.76	4.46
15	34.01	14.53	3.35	7.95
16	0.81	1.00	1.03	1.02

Click **OK** to get the screen to define the variables in the **Variable View**. Details of preparing data file are shown below.

Preparing Data File

The procedure of preparing the data file shall be explained by means of the data shown in Table 1.1.

In SPSS, before entering data, all the variables need to be defined in the **Variable View**. Once **Type in data** option is selected in the screen shown in Fig. 1.3, click the **Variable View**. This will allow you to define all the variables in the SPSS. The blank screen shall look like Fig. 1.4.

Now you are ready for defining the variables row wise.

Defining Variables and Their Properties Under Different Columns

Column 1: In first column, short name of the variables are defined. The variable name should essentially start with an alphabet and may use underscore and numerals in between, without any gap. There should be no space between any two characters of the variable name. Further, variable name should not be started with numerals or any special character.

Column 2: Under the column heading "Type," format of the variable (numeric or nonnumeric) and the number of digits before and after decimal are defined. This can be done by double-clicking the concerned cell. The screen shall look like Fig. 1.5.

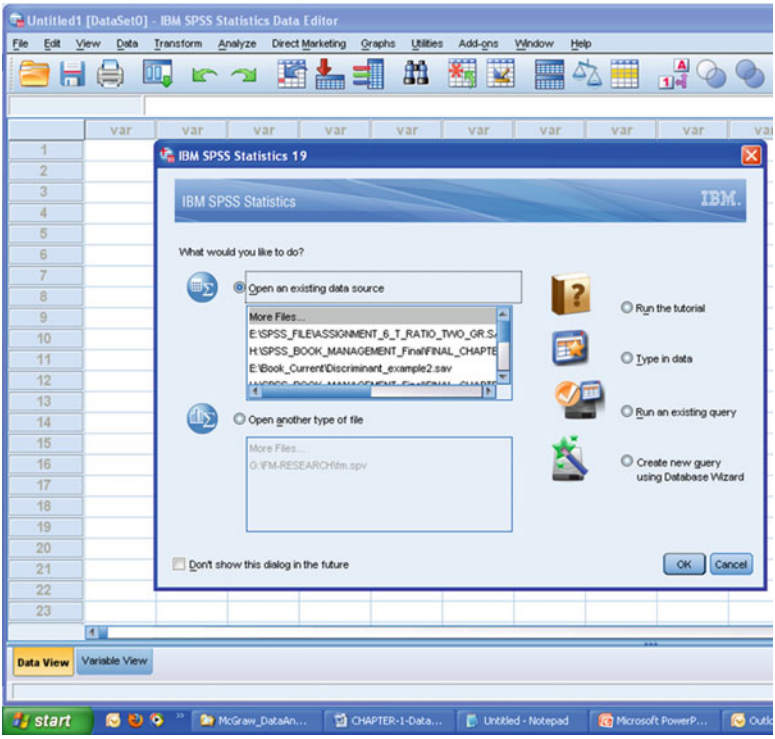


Fig. 1.3 Screen showing the option for creating/opening file

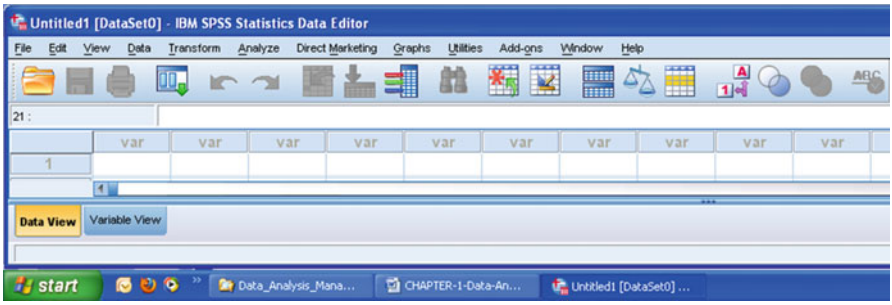


Fig. 1.4 Blank format for defining the variables in SPSS

Column 3: Under the column heading "Width," number of digits a variable can have may be altered.

Column 4: In this column, number of decimal a variable can have may be altered.

Column 5: Under the column heading "Label," full name of the variable can be defined. The user can take advantage of this facility to write the expanded name of the variable the way one feels like.

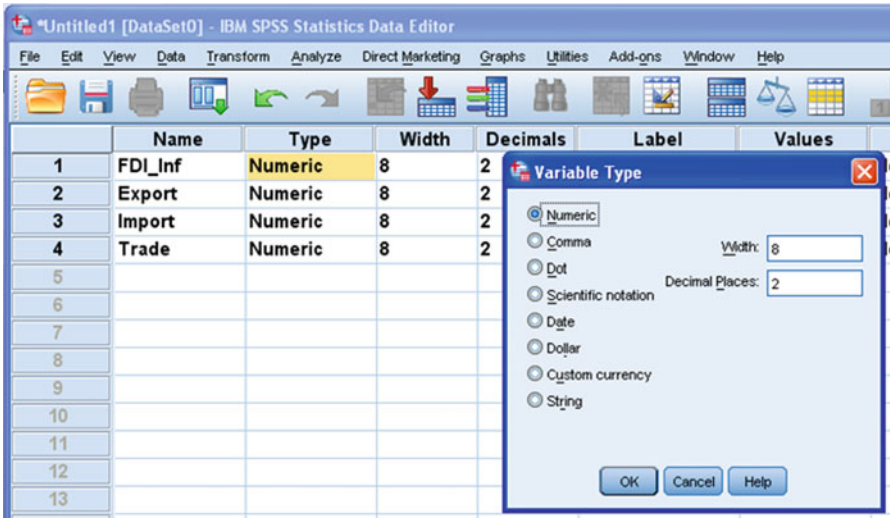


Fig. 1.5 Option showing defining of variable as numeric or nonnumeric

Column 6: Under the column heading "Values," the coding of the variable may be defined by double clicking the cell. Sometimes, the variable is of classificatory in nature. For example, if there is a choice of choosing any one of the following four departments for training

(a) Production
(b) Marketing
(c) Human resource
(d) Public relation

then these departments can be coded as 1 = production, 2 = marketing, 3 = human resource, and 4 = public relation. While entering data into the computer, these codes are entered, as per the response of a particular subject. SPSS window showing the option for entering code has been shown in Fig. 1.6.

Column 7: In survey study, it is quite likely that for certain questions the respondent does not reply, which creates the problem of missing value. Such missing value can be defined under column heading "Missing."

Column 8: Under the heading "Columns," width of the column space where data is typed in Data View is defined.

Column 9: Under the column heading "Align," the alignment of data while feeding may be defined as left, right, or center.

Column 10: Under the column heading "Measure," the variable type may be defined as scale, ordinal, or nominal.

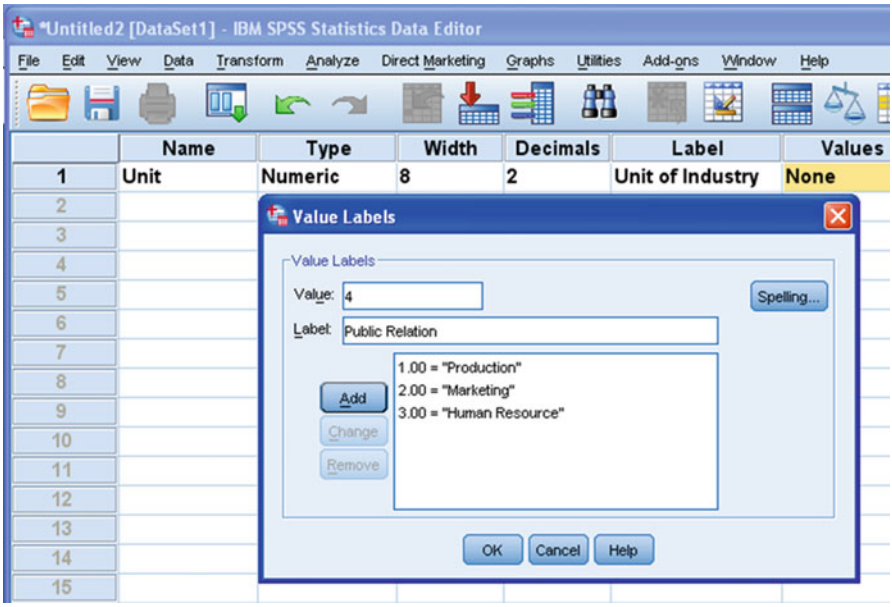


Fig. 1.6 Screen showing how to define the code for the different labels of the variable

Defining Variables for the Data in Table 1.1

1. Write short name of all the five variables as *States, FDI_Inf, Export, Import, and Trade* under the column heading "Name."
2. Under the column heading "Label," full name of these variables may be defined as *FDI Inflows, Export Data, Import Data,* and *Trade Data*. One can take liberty of defining some more detailed name of these variables as well.
3. Use default entries in rest of the columns.

After defining variables in the variable view, the screen shall look like Fig. 1.7.

Entering the Data

After defining all the five variables in the **Variable View**, click **Data View** on the left bottom of the screen to open the format for entering the data. For each variable, data can be entered column wise. After entering the data, the screen will look like Fig. 1.8. Save the data file in the desired location before further processing.

After preparing the data file, one may use different types of statistical analysis available under the tag **Analyze** in the SPSS package. Different types of statistical analyses have been discussed in different chapters of the book along with their interpretations. Methods of data entry are different in certain applications; for

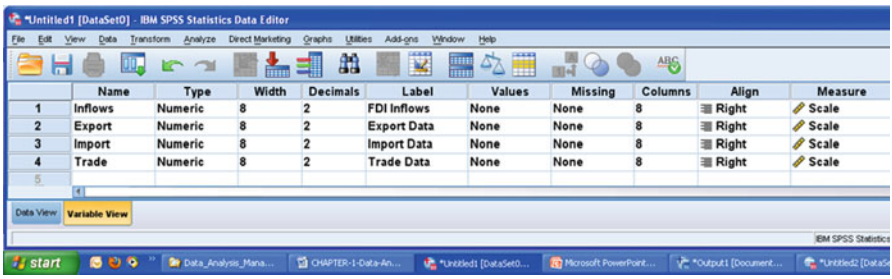
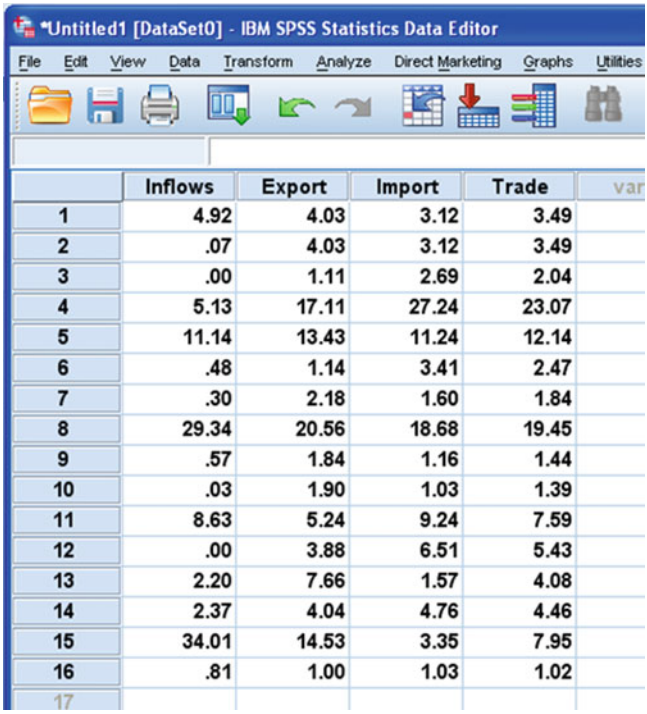


Fig. 1.7 Variables along with their characteristics for the data shown in Table 1.1



	Inflows	Export	Import	Trade	var
1	4.92	4.03	3.12	3.49	
2	.07	4.03	3.12	3.49	
3	.00	1.11	2.69	2.04	
4	5.13	17.11	27.24	23.07	
5	11.14	13.43	11.24	12.14	
6	.48	1.14	3.41	2.47	
7	.30	2.18	1.60	1.84	
8	29.34	20.56	18.68	19.45	
9	.57	1.84	1.16	1.44	
10	.03	1.90	1.03	1.39	
11	8.63	5.24	9.24	7.59	
12	.00	3.88	6.51	5.43	
13	2.20	7.66	1.57	4.08	
14	2.37	4.04	4.76	4.46	
15	34.01	14.53	3.35	7.95	
16	.81	1.00	1.03	1.02	
17					

Fig. 1.8 Screen showing entered data for all the variables in the data view

instance, readers are advised to note carefully the way data is entered for the application in Example 6.2 in Chap. 6. Relevant details have been discussed in that chapter.

Importing Data in SPSS

In SPSS, data can be imported from ASCII as well as Excel file. The procedure of importing these two types of data files has been discussed in the following sections.

Importing Data from an ASCII File

In ASCII file, data for each variable may be separated by a space, tab, comma, or some other character. The Text Import Wizard in SPSS facilitates you to import data from an ASCII file format. Consider the following set of data in ASCII file saved on the desktop by the file name Business data:

File name: *Business data*

S.N.	FDI	Exports inflows	Imports	Trade
1	4.92	4.03	3.12	3.49
2	0.07	4.03	3.12	3.49
3	0.00	1.11	2.69	2.04
4	5.13	17.11	27.24	23.07
5	11.14	13.43	11.24	12.14

The sequence of commands is as follows:

1. For importing the required ASCII file into SPSS, follow the below-mentioned sequence of commands in **Data View.**

$$\textbf{File} -> \textbf{Open} -> \textbf{Data} -> \textbf{Businessdata}$$

   – Choose "Text" as the "File Type" if your ASCII file has the .txt extension. Otherwise, choose the option "All files."
   – After selecting the file that you want to import, click **Open** as shown in Fig. 1.9.

2. After choosing the ASCII file from the saved location in Fig. 1.9, the Text Import Wizard will pop up automatically as shown in Fig. 1.10 that will take you for further option in importing the file. Take the following steps:

   – If your file does not match a predefined format, which is usually not, so click **Next**.

3. After clicking the **Next** option above, you will get the screen as shown in Fig. 1.11. Take the following steps:

   – Define delimiter and check the option "Delimited" as the data in the file is separated by either space or comma.
   – If variable names are written in the first row of your data file, check the header row option as "Yes," otherwise "No." In this, the option "Yes" will be selected because variable names have been written in the first row of the data file. Click **Next.**

4. After clicking the option **Next,** you will get the screen as shown in Fig. 1.12. Enter the line number where the first case of your data begins. If there is no

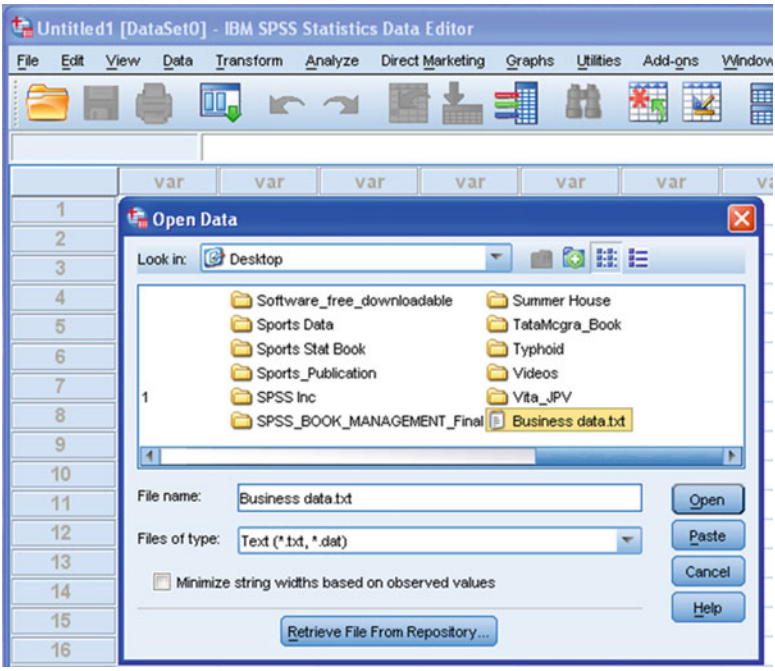


Fig. 1.9 Selecting an ASCII file saved as text file for importing in SPSS

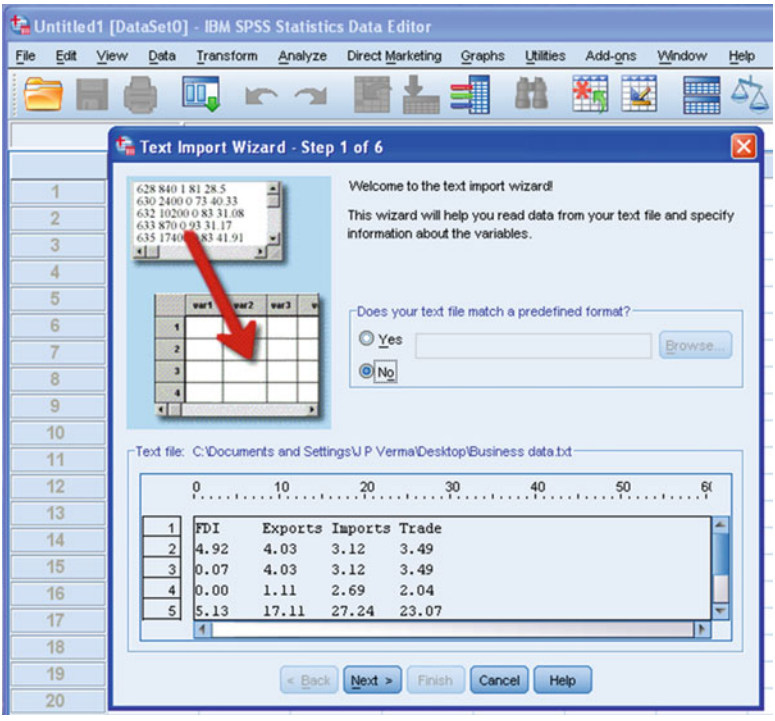


Fig. 1.10 Import text wizard for opening an ASCII file in SPSS

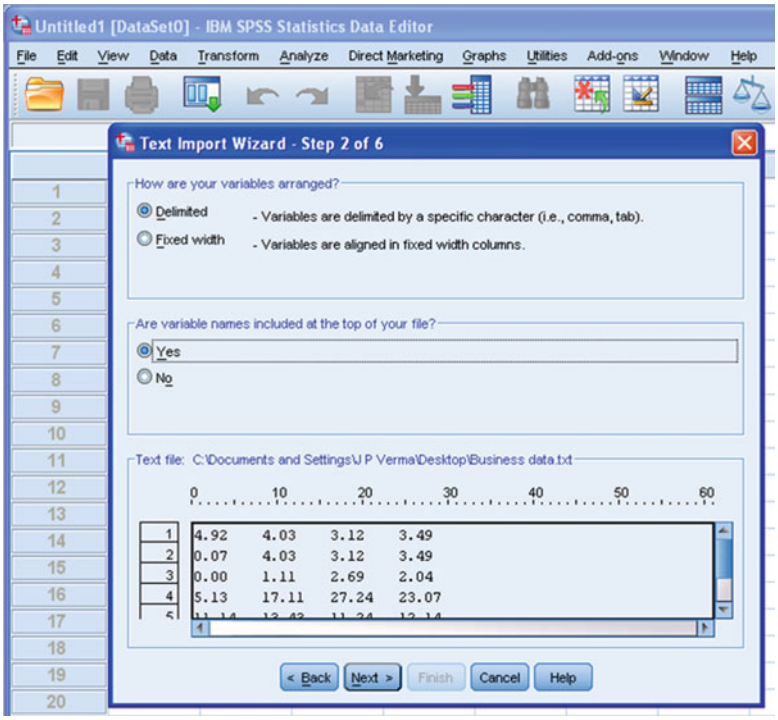


Fig. 1.11 Defining option for delimiter and header row

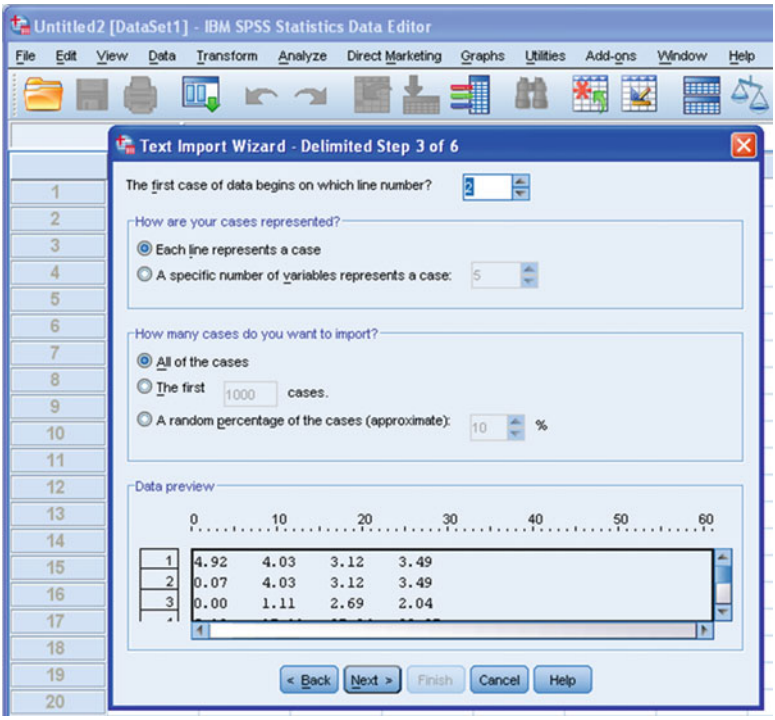


Fig. 1.12 Defining option for beginning line of data and number of cases to be selected

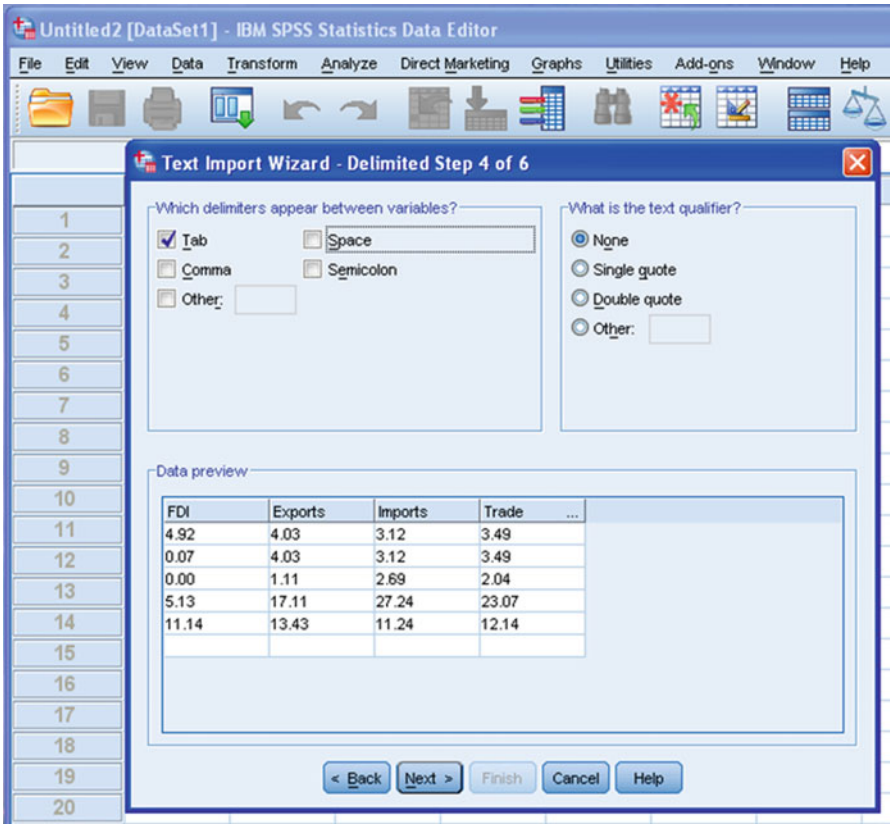


Fig. 1.13 Defining option for delimiters and text qualifier

variable name in the first line of the data file, line 1 is selected; otherwise, line 2 may be selected as the data starts from line 2 in the data file. Take the following steps:

- Check the option "Each line represents a case." Normally in your data file, each line represents a case.
- Check the option "All of the cases." Usually, you import all the cases from the file. Other option may be tried if only few cases are imported from the file. Click **Next** to get the screen as shown in Fig. 1.13.

5. In Fig. 1.13, delimiters of the data file (probably comma or space) are set:

- Check the delimiters as "Coma" as the data is separated by comma. Other delimiters may be selected if used in the data file.
- Check the "Double quote" as text qualifier. Other options may be checked if the variables are flanked other than double quote.
- On the basis of the options chosen by you, SPSS formats the file in the small screen in the bottom. There you can check if everything is set correctly. Click **Next** when everything is ok to get the screen as shown in Fig. 1.14.

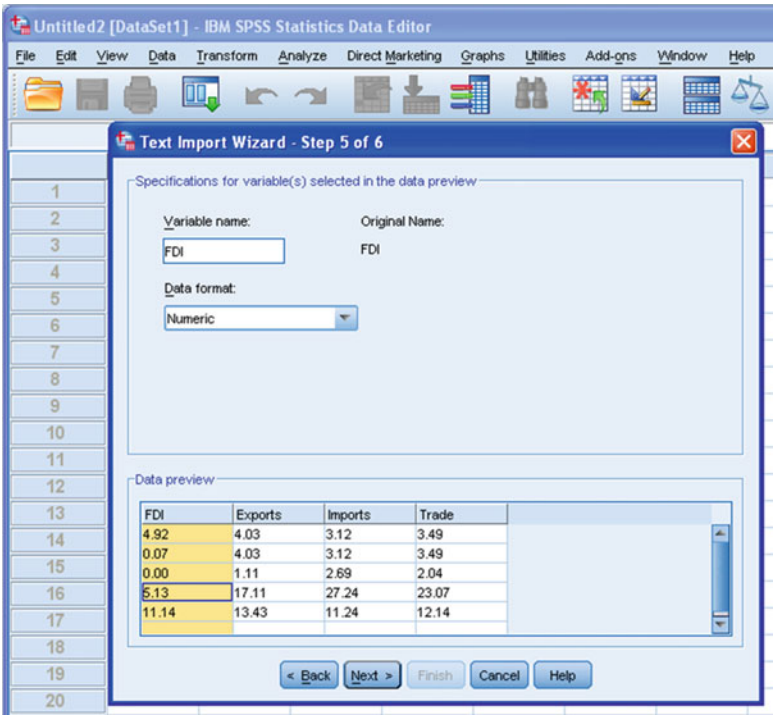


Fig. 1.14 Defining specifications for the variables

6. In Fig. 1.14, you can define the specifications for the variables, but you may just ignore it if you have already defined your variables or want to do it later. Click **Next** to get the screen as shown in Fig. 1.15.
7. In Fig. 1.15, select all the default options and ensure that your actual data file has been shown in the window or not. Once your data is shown in the window, click **Finish**. This will import your file successfully in SPSS.

Importing Data File from Excel Format

The data prepared in Excel file can be imported in SPSS by simple command. While importing Excel data file, one must ensure that it is not open. The sequence of commands for importing Excel data file is as follows:

1. **File** $->$ **Open** $->$ **Data** $->$ **requiredfile**
   - Choose "Excel" as the File Type if your ASCII file has the .xls extension. Otherwise, choose the option "All files."
   - After selecting the file that you want to import, click **Open** as shown in Fig. 1.16.

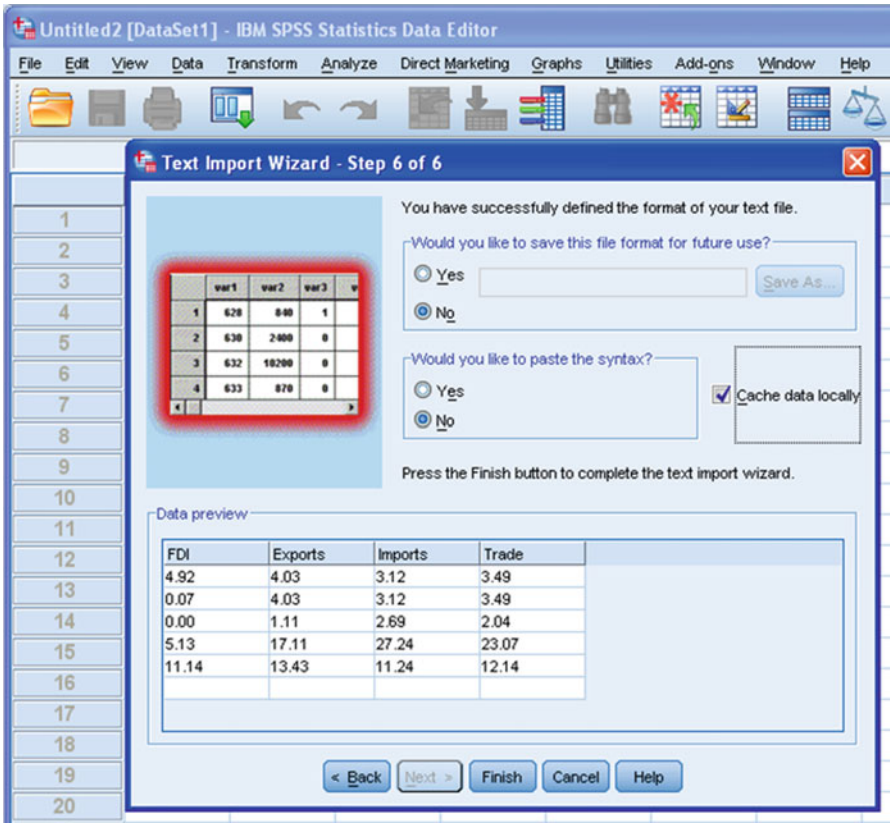


Fig. 1.15 Option for saving the format

2. After choosing the required Excel file from the saved location in Fig. 1.16, you will get the pop screen called "Opening Excel Data Source" as shown in Fig. 1.17. Take the following steps:
   - Check the option "Read variable names from the first row of data" if you are using the header row in the data file.
   - Select the right worksheet from which you want to import the data. The screen will show you all the worksheets of the file containing data. If you have data only in the first worksheet, leave this option as it is.
   - If you want to use only a portion of data from the file, define the fields in "**Range**" option like A3:E8. This means that the data from the A3 row till column E8 shall be selected.
   - Press **Continue** to get the Excel file opened in SPSS.

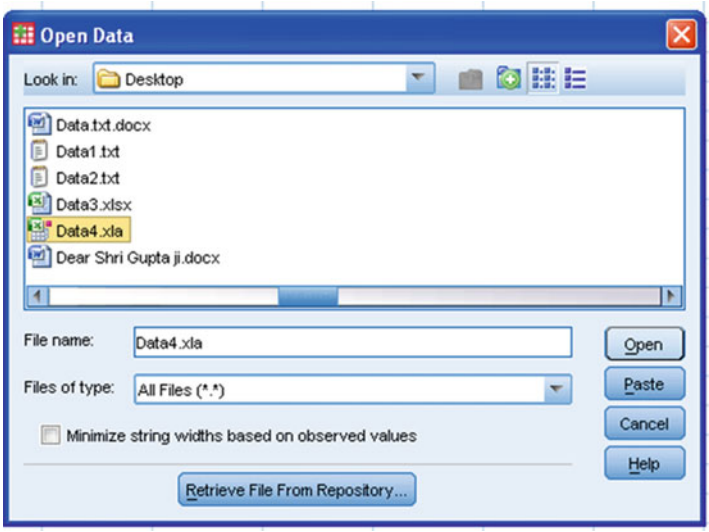


Fig. 1.16 Selecting an Excel file for importing in SPSS

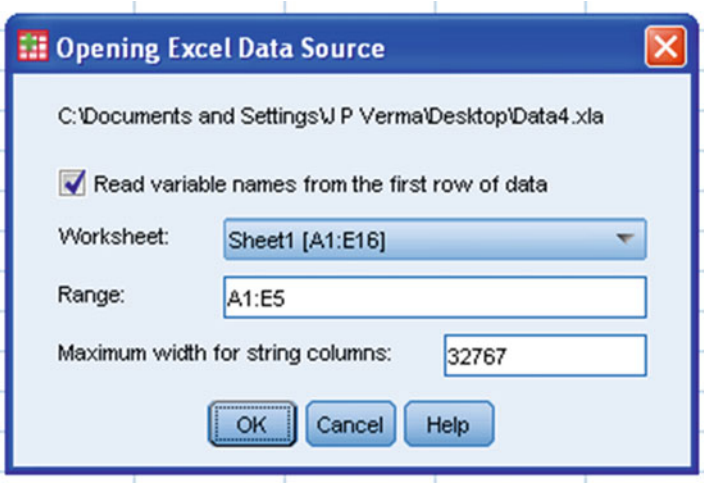


Fig. 1.17 Option for defining the range of data in Excel sheet to be imported in SPSS

Exercise

Short-Answer Questions

Note: Write the answer to each of the questions in not more than 200 words.

Q.1. What do you mean by inductive and inferential statistics? What is the difference between them? Explain by means of example.
Q.2. What do you mean by metric and nonmetric data? Discuss them by means of example.
Q.3. Under what situation analytical studies should be conducted? Discuss a situation where it can be used.
Q.4. What do you mean by mutually exclusive and independent attributes? Give two examples where the attributes are not mutually exclusive.
Q.5. What is an extraneous variable? How it affects the findings of an experiment? Suggest remedies for eliminating its effects.
Q.6. While feeding the data in SPSS, what are the possible mistakes that a user might commit?
Q.7. Explain in brief as to how an error can be identified in data feeding.

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers. Tick mark the one that you consider the closest to the correct answer.

1. Given the following statements,

   I. Parametric tests do not assume anything about the form of the distribution.
   II. Nonparametric tests are simple to use.
   III. Parametric tests are the most powerful if their assumptions are satisfied.
   IV. Nonparametric tests are based upon the assumption of normality.

   choose the correct statements from the above-listed ones.

   (a) (I) and (II)
   (b) (I) and (III)
   (c) (II) and (III)
   (d) (III) and (IV)

2. If the respondents were required to rate themselves on emotional strength on a 9-point scale, what type of data would be generated?

   (a) Ratio
   (b) Interval
   (c) Nominal
   (d) Ordinal

3. The variable measured on which of the following scales are termed as categorical.

   (a) Ratio and interval
   (b) Interval and ordinal
   (c) Interval and nominal
   (d) Ordinal and nominal

4. In tossing an unbiased coin, one can get the following events:
   E_1: getting a head, E_2: getting a tail. Choose the correct statement.

   (a) E_1 and E_2 are independent.
   (b) E_1 and E_2 are mutually exclusive.
   (c) E_1 and E_2 are not equally likely.
   (d) E_1 and E_2 are independent and mutually exclusive.

5. While creating a new data file in SPSS, which option should be used?

   (a) Type in data
   (b) Open an existing data source
   (c) Open another type of file
   (d) None

6. Identify valid name of the variable.

   (a) SalesData
   (b) Cust No
   (c) Outlet " Center"
   (d) Sales-Data

7. While defining the types of the variable under the heading "Measure" in SPSS, what are the valid options out of the following?

   (i) Interval (ii) Scale
   (iii) Nominal (iv) Ordinal

   (a) (i),(ii), and (iii)
   (b) (i),(ii), and (iv)
   (c) (i),(iii), and (iv)
   (d) (ii),(iii), and (iv)

8. For analyzing the data, the commands are selected while being in the

   (a) Variable View
   (b) Data View
   (c) Data and Variable View
   (d) Neither in Data nor in Variable View

9. Runs scored in a cricket match is

   (a) Interval data
   (b) Ratio data
   (c) Nominal data
   (d) Ordinal data

10. In an experiment, effect of three types of incentives on satisfaction level has to be seen on the subjects. Choose the correct statement.

    (a) Incentive is a dependent variable, and satisfaction level is an independent variable.
    (b) Incentive is an independent variable, and satisfaction level is a dependent variable.
    (c) Incentive and satisfaction level are independent variables.
    (d) Incentive and satisfaction level both are dependent variables.

Answers to Multiple-Choice Questions

Q.1	c	Q.2	b
Q.3	d	Q.4	b
Q.5	a	Q.6	a
Q.7	d	Q.8	b
Q.9	b	Q.10	b