1

INTRODUCTION TO DATA TYPES AND SPSS OPERATIONS

LEARNING OBJECTIVES

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

- Understand different data types generated in research
- Learn the nature of variables
- Know various data cleaning methods
- Learn to install SPSS package in computer
- Prepare data file in SPSS

1.1 INTRODUCTION

Due to large stake involved in sports, research in this area is gaining momentum in different universities of the world. Even developing countries have started introducing sports sciences in different universities. The sole purpose is to create specific knowledge required for enhancing sports performance. Everyday, enormous data is being generated in the area of sports all over the world, which can be used to draw meaningful conclusions. Scientists have started organizing experiments by taking athletes as subjects. It is therefore required to support these scientists with analytical skill

Sports Research with Analytical Solution using SPSS®, First Edition. J. P. Verma.

Companion website: www.wiley.com/go/Verma/Sportsresearch

set to carry out their business. Since they deal with the data, it is essential that they are aware of its nature. Depending upon the data types, one identifies the relevant analytical technique for addressing research issues. Sports research can broadly be classified into two categories: descriptive and analytical. In descriptive research, the nature of dataset is investigated from different perspectives. Several statistics like mean, standard deviation, coefficient of variation, skewness, kurtosis, and percentiles are used to describe the characteristics of the dataset. Many interesting facts about the population can be investigated by using these descriptive statistics. Analytical research broadly follows two approach: exploratory and confirmatory. In explorative research, focus is on discovering the hidden relationships. It is done by hypothesis testing, data modeling, and using multivariate analysis. On the other hand, in confirmatory studies, some of the facts are either confirmed or denied on the basis of hypothesis testing.

Numerous statistical techniques are available to the researchers for analyzing their research data. Selection of an appropriate technique depends upon the research questions being investigated in the study. Due to complexities of different analytical solutions in sports research, one needs to use some user-friendly software package. This chapter will acquaint you with different types of data that are generated in sports research and some of the widely used statistical techniques by the research scholars to solve them for answering different research questions by using the most popular IBM SPSS Statistics package.

1.2 TYPES OF DATA

It is essential to know the types of data generated in research studies because choosing statistical test for analyzing data depends upon its type. Data can be classified into two categories: metric and nonmetric. Metric data is analyzed by using parametric tests such as t, F, Z, correlation coefficient, etc., whereas nonparametric tests such as Wilcoxon signed-ranked, Chi-square, Mann–Whitney *U*, and Kruskal Wallis are used to analyze nonmetric data.

Parametric tests are more reliable than the nonparametric, but to use such tests certain assumptions must be satisfied. On the other hand, nonparametric tests are more flexible, easy to use, and not many assumptions are required to use them.

Nonmetric and metric data are also known as qualitative and quantitative data, respectively. Nonmetric data is further classified into nominal and ordinal. On the other hand, metric data is classified into interval and ratio. These classification is based on the level of measurements. The details of these four types of data have been discussed under two categories: qualitative data and quantitative data.

1.2.1 Qualitative Data

Qualitative data is a categorical measurement and is expressed not in terms of numbers, rather by means of a natural language description. It is often known as "categorical" data. For instance, smoking habit = "smoker" and gender = "male" are

the examples of categorical data. These data can be measured on two different scales: nominal and ordinal.

1.2.1.1 Nominal Scale Variables measured on this scale are known as categorical variables. Categorical variables result from a selection of categories. Examples might be response (agree, disagree), sports specialization, race, religion, etc. If in a class 30 subjects are male and 20 are female, no gradation is possible. In other words, 30 do not indicate that the males are better than the female in some sense.

1.2.1.2 Ordinal Scale Variables that are assessed on the ordinal scale are also known as categorical variables, but here the categories are ordered. Such variables are also called "ordinal variables." Categorical variables that assess performance (good, average, poor, etc.) are ordinal variables. Similarly, the variables that measure attitude (strongly agree, agree, undecided, disagree, and strongly disagree) are also ordinal variables. On the basis of the order of these variables, we may not know the magnitude of the measured phenomenon of an individual, but we can always grade them. For instance, if A's playing ability in soccer is good and B's is average, we can always conclude that the A is better than B, but how much is not known. Moreover, the distance between the ordered categories is also not same and measurable.

1.2.2 Quantitative Data

Quantitative data is a numerical measurement expressed in terms of numbers. It is not necessary that all numbers are continuous and measurable. For instance, the roll number is a number, but not something that one can add or subtract. Quantitative data are always associated with a scale measure. These data can be measured on two different types of scales: interval and ratio.

1.2.2.1 Interval Scale The interval scale is a quantitative measure. It also has an equidistant measure. But the doubling principle breaks down in this scale. The 4 marks given to an individual for his creativity do not explain that his nature is twice as good as the person with 2 marks. This is so because on this scale zero cannot be exactly located. Thus, variables measured on an interval scale have values in which differences are uniform, but ratios are not.

1.2.2.2 Ratio Scale 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 in a test is twice that of 30. This is so because zero exists in the ratio scale. Height is another ratio scale quantitative measure. Observations that are counted or measured are ratio data (e.g., number of goals, runs, height, and weight).

1.3 IMPORTANT DEFINITIONS

1.3.1 Variable

A variable is a phenomenon that changes from time to time, place to place, and individual to individual. It can be numeric or attribute. Numeric variable can further be classified into discrete and continuous. *Discrete variable* is a numeric variable that assumes value from a limited set of numbers and is always represented in whole number. Examples of such variables are number of goals, runs scored in cricket, scores in basketball match, etc. *Continuous variable* is also a numeric variable, but it can take any value within a range and is usually represented in fraction. Examples of such variables are height, weight, and timings.

On the other hand, an attribute is a qualitative variable that takes sub-values of a variable, such as "male" and "female," "student" and "teacher," etc. An attribute is said to be mutually exclusive if its sub-values do not occur at the same time. For instance, gender is a mutually exclusive variable because it can take value either "male" or "female" but not both. Similarly in a survey, a person can choose only one option from a list of alternatives (as opposed to selecting as many that might apply).

1.3.1.1 Independent Variable An independent variable can be defined as the one that can be manipulated by a researcher. In planning a research experiment to see the effect of different intensities of exercise on the performance, exercise intensity is an independent variable because the researcher is free to manipulate it.

1.3.1.2 Dependent Variable A variable is said to be dependent if it changes as a result of the change in the independent variable. In the previous example, performance is a dependent variable because it is affected by the change in exercise intensity. In fact dependent variable can be defined as the variable of interest. In creating the graph, the dependent variable is taken along the Y-axis, whereas the independent variable is plotted on the X-axis.

1.3.1.3 Extraneous Variable Any additional variable that may provide alternative explanation or create some doubt on the conclusions in an experimental study is known as extraneous variable. If the effect of three different teaching methods on the performance is to be compared, IQ of the subjects may be considered as an extraneous variable as it might affect the final outcomes in the experiment, if IQ of all the groups are not equal initially.

1.4 DATA CLEANING

Data needs to be organized before preparing a data file. There are more chances that a dataset may contain unusual data due to wrong feeding or due to extreme cases. And if it is so, the analyzed results may lead to erroneous conclusions. Analysts tend to waste lots of time in drawing valid conclusions if the data is erroneous. Thus, it is utmost important that the data must be cleaned before analysis. In cleaned data, analysis becomes straightforward and valid conclusions can be drawn from it.

In data cleaning, first an unusual data is detected 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

1.5 DETECTION OF ERRORS

The wrongly fed data can be detected by means of descriptive statistics. Some useful approaches in this regard are given in the text.

1.5.1 Using Frequencies

One of the methods of cleaning data is to use frequency of each score obtained in descriptive statistics. Since most of the behavioral parameters are normally distributed; therefore, if any anthropometric or physical variable shows large frequency for any values, it must be checked for any systematic error.

1.5.2 Using Mean and Standard Deviation

Normally, the value of standard deviation is less than the mean, except in case of the distribution like negative binomial. Thus, if the value of standard deviation for any of the variables like age, height, or cardio-respiratory index is more than its mean, then some of the values of these variables must be negative. However, the value of these variables cannot be negative, and thus one may identify the wrongly fed data.

1.5.3 Logic Checks

Data error may also be detected by observing whether responses are logical or not? For example, one would expect to see 100% of responses, not 110%. Another example would be if a question is asked to female respondents about their periods and the reply is marked "yes," but you notice that the respondent is coded "MALE." Logical approach is to be adopted with full justification, to avoid the embarrassing situation like in reporting that 10% of the men in the sample had periods during training.

1.5.4 Outlier Detection

The unusual data can also be identified by detecting the outliers. Any data that lies outside the two sigma limits can be considered to be outlier. In other words, data lying outside the range mean $\pm$ 2SD may be identified as an outlier and may be removed from the dataset. If a liberal view is adopted, then one can take mean $\pm$ 3SD

limits to detect the unusual data. The outlier can be detected in a dataset by means of Boxplot discussed in Chapter 2.

1.6 HOW TO START SPSS?

This book has been written by using the IBM SPSS software. The SPSS package needs to be activated on the computer before entering the data. This can be done by clicking the left button of the mouse on SPSS icon in the SPSS directory in the **Start** and **All Programs** option (if the SPSS directory has been created in the programs file). Using the command sequence shown in Figure 1.1, SPSS can be activated. The last box is marked SPSS, but usually it will be followed by the version you are using.

FIGURE 1.1 Sequence of commands for starting SPSS package.

By using the aforementioned command sequence and clicking **IBM SPSS Statistics 20** in the window shown in Figure 1.2, you will get the screen as shown in Figure 1.3 to prepare the data file or open the existing data file.

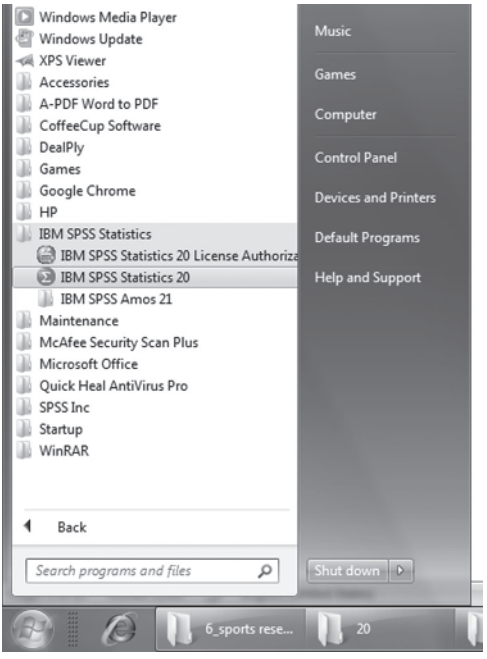


FIGURE 1.2 Commands for initiating SPSS.

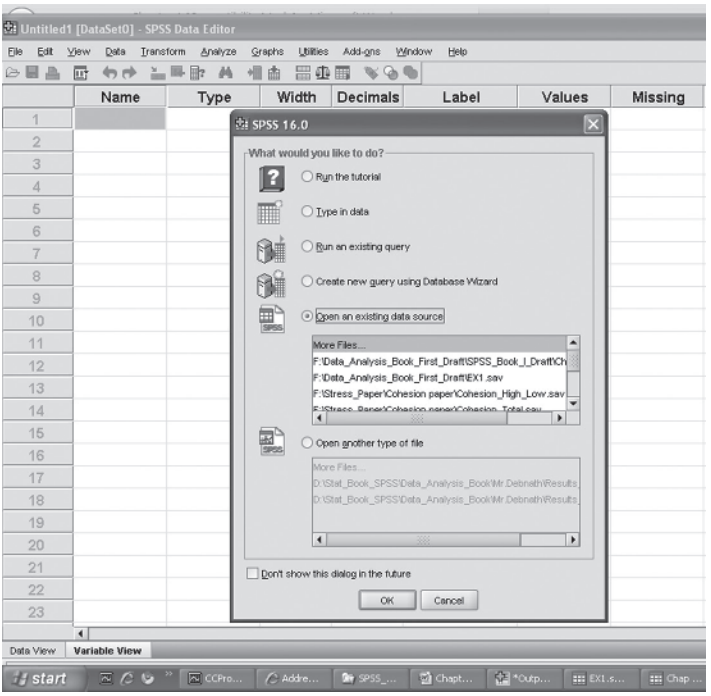


FIGURE 1.3 Screen showing the option for creating/opening data file.

If you are entering the data for a new problem and the file is to be created for the first time, check 'Type in data' option and if the existing file is to be opened or edited, then select the 'Open an existing data source' option in the window shown in Figure 1.3.

Click on **OK** to get the screen for defining variables in the **Variable View**.

1.6.1 Preparing Data File

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

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

Now you are ready for defining variables row-wise.

1.6.1.1 Procedure for Defining Variables and Their Properties

Column 1: Under the column heading "Name," short names of the variables are defined. The variable name should essentially start with an alphabet only and may include underscore and numerals in between without any gap. If at all the variable needs to be defined in two words, then they must be joined by using the underscore such as Playing_Ability or Muscular_Strength.

TABLE 1.1 Data on Anthropometric Parameters Obtained on College Badminton Players

S.N.	Height (cm)	Weight (kg)	Arm Length (cm)	Leg Length (cm)	Trunk Length (cm)	Thigh Girth (cm)	Shoulder Width (cm)
1	177	66	82	89	91	50	36
2	172	75	74	90	85	52	41
3	180	68	85	87	91	51	44
4	189	49	81	96	91	54	48
5	180	55	75	95	86	47	37
6	175	74	82	89	88	51	43
7	187	73	86	93	92	52	42
8	181	69	73	96	84	50	44
9	171	68	75	87	86	54	43
10	180	62	78	92	91	48	39
11	177	66	72	91	85	53	44
12	163	68	71	88	77	52	45
13	162	65	73	87	76	54	46
14	168	67	74	89	78	53	48
15	165	69	75	91	79	51	47

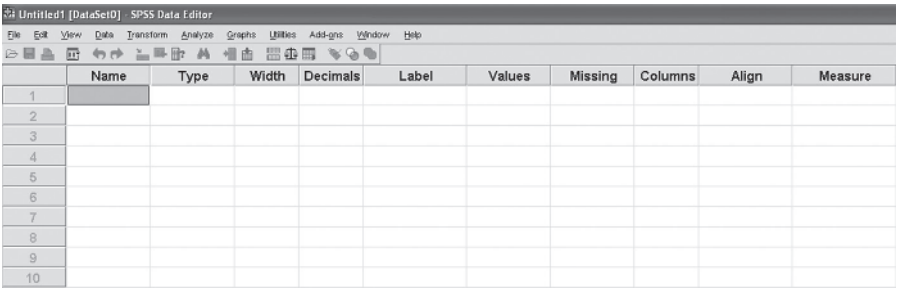


FIGURE 1.4 Blank format for defining variables.

Column 2: Under the column heading "Type," format of the variables (numeric or non-numeric) is defined. This can be done by double clicking the cell. The screen shall look like as shown in Figure 1.5.

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

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

Column 5: Under the column heading "Label," full name of the variable can be written. User can take advantage of this facility to write expanded name of the variable.

Column 6: Under the column heading "Values," coding of the variable is defined by double clicking the cell if the variable is of classificatory in nature. For example, if there is a choice of choosing any one of the four sports—cricket, gymnastics, swimming, and athletics—then these sport categories can

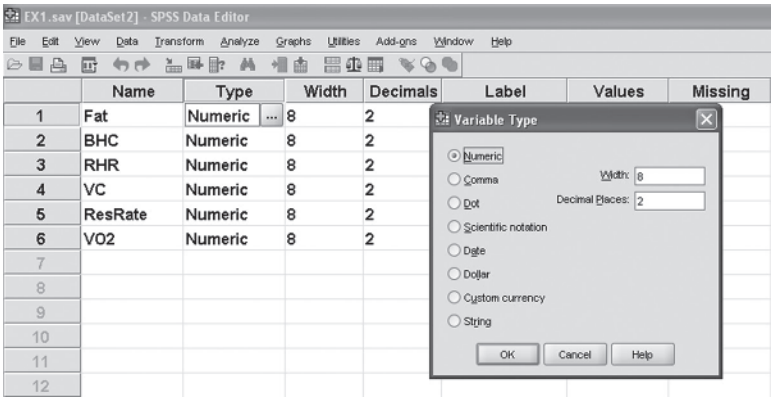


FIGURE 1.5 Defining variables and their characteristics.

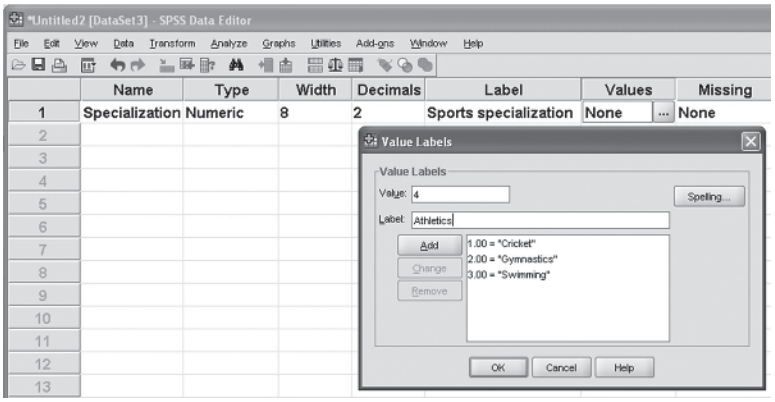


FIGURE 1.6 Defining code of nominal variable.

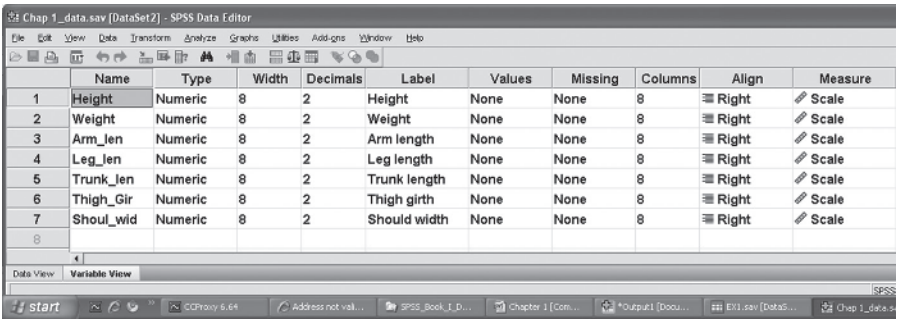
be coded as 1 = cricket, 2 = gymnastics, 3 = swimming, and 4 = athletics. While entering data into computer, these codes are entered as per the response of a particular subject. SPSS window showing the option for entering the code has been shown in Figure 1.6.

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

Column 8: Under the heading "Columns," the width of the column space where data is typed in **Data View** is defined (Figure 1.7).

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. Scale is used for interval and ratio data both.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Height	Numeric	8	2	Height	None	None	8	Right	Scale
2	Weight	Numeric	8	2	Weight	None	None	8	Right	Scale
3	Arm_len	Numeric	8	2	Arm length	None	None	8	Right	Scale
4	Leg_len	Numeric	8	2	Leg length	None	None	8	Right	Scale
5	Trunk_len	Numeric	8	2	Trunk length	None	None	8	Right	Scale
6	Thigh_Gir	Numeric	8	2	Thigh girth	None	None	8	Right	Scale
7	Shoul_wid	Numeric	8	2	Should width	None	None	8	Right	Scale
8										

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

1.6.1.1.1 Defining Variables

1. Write short name of each of the seven variables as *Height, Weight, Arm_len, Leg_len, Trunk_len, Thigh_Gir,* and *Shoul_wid* under the column heading "Name."
2. Under the column heading "Label," full names of these variables may be defined as *Height, Weight, Arm length, Leg length, Trunk length, Thigh girth,* and *Shoulder width*. One may define some other names as well.
3. Use default entries in rest of the columns.

After defining variables in **Variable View**, screen shall look like as shown in Figure 1.7.

1.6.1.1.2 Entering Data After defining the variables, click the **Data View** option on the left corner in the bottom of the screen to open the format for entering data. For each variable, data can be entered column-wise. After entering data, the screen will look like as shown in Figure 1.8. Save the data file in the desired location before further processing.

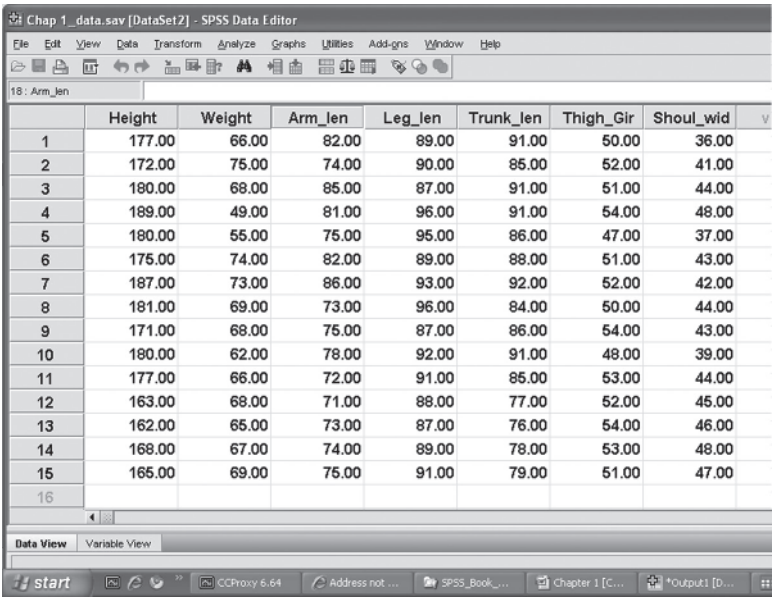
After preparing the data file, one may use it for different types of statistical analysis available under the **Analyze** in SPSS. Various types of statistical analyses have been discussed along with their interpretations in different chapters of the book. Methods of data entry differ in different applications. Relevant details have been discussed in different chapters.

1.7 EXERCISE

1.7.1 Short Answer Questions

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

Q.1 What do you mean by exploratory data analysis? Explain any one situation in research where such analysis can be applied.



	Height	Weight	Arm_len	Leg_len	Trunk_len	Thigh_Gir	Shoul_wid
1	177.00	66.00	82.00	89.00	91.00	50.00	36.00
2	172.00	75.00	74.00	90.00	85.00	52.00	41.00
3	180.00	68.00	85.00	87.00	91.00	51.00	44.00
4	189.00	49.00	81.00	96.00	91.00	54.00	48.00
5	180.00	55.00	75.00	95.00	86.00	47.00	37.00
6	175.00	74.00	82.00	89.00	88.00	51.00	43.00
7	187.00	73.00	86.00	93.00	92.00	52.00	42.00
8	181.00	69.00	73.00	96.00	84.00	50.00	44.00
9	171.00	68.00	75.00	87.00	86.00	54.00	43.00
10	180.00	62.00	78.00	92.00	91.00	48.00	39.00
11	177.00	66.00	72.00	91.00	85.00	53.00	44.00
12	163.00	68.00	71.00	88.00	77.00	52.00	45.00
13	162.00	65.00	73.00	87.00	76.00	54.00	46.00
14	168.00	67.00	74.00	89.00	78.00	53.00	48.00
15	165.00	69.00	75.00	91.00	79.00	51.00	47.00
16							

FIGURE 1.8 Format of data entry in most of the applications.

Q.2 What do you mean by ratio scale, and how is it different from interval scale?

Q.3 Under what situations should qualitative data be preferred? Explain its types with examples.

Q.4 Explain a situation in research where responses can be obtained on mutually exclusive attributes.

Q.5 What is an extraneous variable? How does it affect findings in an experiment? Suggest remedies for eliminating its effects.

Q.6 While feeding data in SPSS, what are the possible mistakes that a user may commit?

Q.7 Explain in brief as to how an error can be identified in data feeding.

1.7.2 Multiple Choice Questions

Note: Questions 1–10 have four alternative answers for each question. Tick mark the one that you consider the closest to the correct answer.

1 Read the following statements carefully:

(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 assumptions of normality.

Choose the correct statements.
(a) (i) and (ii)
(b) (i) and (iii)
(c) (ii) and (iii)
(d) (iii) and (iv)

2 If 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 term "categorical variables" are used for the data measured on
(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 E1: getting a head, E2: getting a tail. Choose the correct statement.
(a) E1 and E2 are independent.
(b) E1 and E2 are mutually exclusive.
(c) E1 and E2 are not equally likely.
(d) E1 and E2 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 a variable.
(a) CardioRes
(b) My Flexibility
(c) My Height
(d) Cardio-Res

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 Choose the correct statement.
(a) t-test and chi-square tests are parametric
(b) t-test is parametric and chi-square test is nonparametric
(c) t-test and chi-square tests are nonparametric
(d) t-test is nonparametric and chi-square test is parametric

9 Runs scored in a cricket match is
(a) Interval data
(b) Ratio data
(c) Nominal data
(d) Ordinal data

10 In an experiment, the effect of different intensities of aerobic exercises on cardio-respiratory endurance has to be seen on the subjects. Choose the correct statement.
(a) Aerobic intensity is a dependent variable and cardio-respiratory endurance is an independent variable.
(b) Aerobic intensity is an independent variable and cardio-respiratory endurance is a dependent variable.
(c) Aerobic intensities and cardio-respiratory endurance both are independent variables.
(d) Aerobic intensities and cardio-respiratory endurance both are dependent variables.