i. ***Blank entries*:** Studies involve collection of data for many variables. It is possible that all variables are not recorded for all participants. This results in blanks which may be interpreted as "zero". Reason for blanks could be: variable not measured or variable measured but not entered. Too many missing entries reduce the effective sample size for that variable.

ii. ***Check for distribution of each variable*:** This involves finding minimum and maximum values for variables measured on ratio/interval scale. This may show an illegal entry. An outlier is an extreme value. Such observations need to be verified.

iii. ***Check for total*:** This is useful especially for nominal and ordinal variables. For example, if you are undertaking study on birth weight, then A. Total of male /female children and B. Total of underweight, normal weight and overweight children should be same. Sometimes these total may not match. For example, a number of deliveries and live births in a hospital during month may not match due to possibility of still-births and twins.

iv. ***Checking two variables at a time*:** For example, age of mother and a number of children, gender and type of operation.

Computer software can restrict recording of invalid data. They give an alert message moment an invalid input is attempted.

Use of computer and computer programs greatly assists data editing. Illegal entries and unusual entries are either not accepted or are accepted after confirmation. A sophisticated data entry program can also check for consistency between some key variables, and can avoid potential inconsistencies.

***Data reduction*:** This involves data summarization. In effect, it includes preparation of a contingency table in which the frequency distribution of subjects (or units) with specific combination of variable/characteristic. It gives frequency distribution of some key variables. Data reduction requires collapsing of the data into categories. Two decisions are required to be made at this stage:

i. Number of categories
ii. The boundaries of these categories.

Making too many categories stretches the data in such a way that the number of observations in each category becomes too few to be of any value. This makes the categories "statistically unstable" and produces "large random errors." Making too few categories defeats the purpose of data reduction. Following equation comes handy

$$K = 1 + (3.332 \times \text{Log } 10\ (n)) \quad \ldots \text{(Eq. 3.1)}$$

Here,

K = Number of categories

n is the number of observations

For $n = 200$, K will be 8.33 (= 9)

In general, additional benefit of more than five categories is marginal. The second decision is about the boundaries of the categories. There can be "natural" categories if the distribution has more than one mode. Secondly, it is desirable to retain extreme categories in the analysis without merging them into adjacent categories. These are often those groups that permit biologically most informative contrasts. The grouping should be such that starting point and end point of group can be clearly inferred. Thus, while grouping for one data category like 60+ will not be able to tell the maximum age in the group.

Chapter

5

Data Presentation

***Objectives of data presentation*:** Data presentation can have one or more of the following objectives:

i. It is a step before analysis/interpretation of the data.

ii. It involves reduction in the volume of the data. This facilitates the better understanding.

iii. When the data is presented in the form of tables, graphs or pictures it makes the data interesting.

Methods of data presentation: The data can be presented fully/partially in the form of text. Similarly, it can be presented in the form of centering constants, rates, ratios, proportions, etc. However, traditionally these are not considered as methods of data presentation. The common methods of data presentation are:

- *Tabular method*: Frequency table, association table, correlation table and master table.
- *Graphical method*: Bar diagram, line diagram, scatter diagram, histogram, frequency polygon, pie diagram.
- *Pictorial method*: Pictogram.

5.1 TABULAR PRESENTATION

Here the data is compiled into groups and subgroups. Due compression of vast data in groups, it is reduced in bulk, increases its attractiveness and makes it easier to understand.

***Anatomy of table*:** A typical table has number, title, columns, rows, cells and footnote. Tables are numbered chronologically as they appear in the text. There are various systems of numbering the tables like using simple numbers, roman numbers. In this book, tables are numbered according to chapter. For example, first table in Chapter 5 will be Table 5.1, second table in Chapter 5 will be numbered as Table 5.2 and so on. The title of the table should be short and self-explanatory.

The "body" of the table consists of columns, rows, cells and total. Columns and rows indicate the demarcations of various groups/subgroups compiled out of the data. The figures in the "cells" pertain to the corresponding row and column. The total is given for each row and column. Usually these are given in the last column and row. In the 'cell' corresponding to the last column and last row grand total is given. Units of measurement are usually indicated in the title or in the column/row itself. In a few cases it may be indicated in the 'footnote'. The "footnote" of the table is used for:

- Indicating the source of the data.
- Explaining the discrepancies, if any, in the data.
- Providing additional information not given in the title and body of the table (e.g. explanation of the abbreviations used).

Frequency polygon emerges when the center of the top of bars in histogram is joined. It reaches baseline through center of height of first and last bars.

Pie diagram

***Indication*:** Used for presenting qualitative data and quantitative discrete data.

***Method*:** Frequency of the attribute/variate is shown by dividing a circle into segments. Each segment represents an attribute/variate. The area of segment is proportional to frequency.

***Example*:** Distribution of students as per marks obtained in an examination is given below:

<50:25, 50–70:57, >70:18

Related pie diagram is shown in Fig. 5.8. Note that, the area of circle allocated for each mark's category is proportional to the number of students in the respective category.

The procedure for pie chart in MS Excel is explained in Box 5.5.

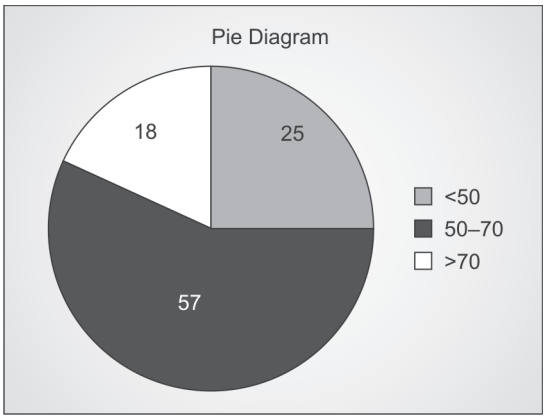


Fig. 5.8: Number of students as per marks obtained

Box 5.5: Creating pie diagram with MS Excel

i. Type the data related to Fig. 5.8 in excel sheet. Start from cell A_1 (Marks). You should end in cell B_4 wherein value 18 will be entered.
ii. Select array A_1:B_4.
iii. Click Home->Insert->Pie. You will have numerous options of Pie. Select first option.
iv. Default pie diagram will be displayed.
v. To change color of segments click on the segment and follow Format->Shape Fill and give desired color. Repeat for all segments.
vi. To show frequency of the group in respective segment, select entire circle and follow Layout-> Data Labels.
vii. To give customized chart title follow Layout-> Chart Title.

Pictogram

***Indication*:** It is suitable for quantitative discrete variate or qualitative characteristics. The presentation is meant lay persons.

***Method*:** Each attribute/variate is shown in pictorial form and its frequency is shown against the picture.

***Example*:** The percent distribution of expenditure for various items in a hospital for a year is Tablets: 45%, Inj: 40%, Solid-Liquid: 15% shown in Fig. 5.9.

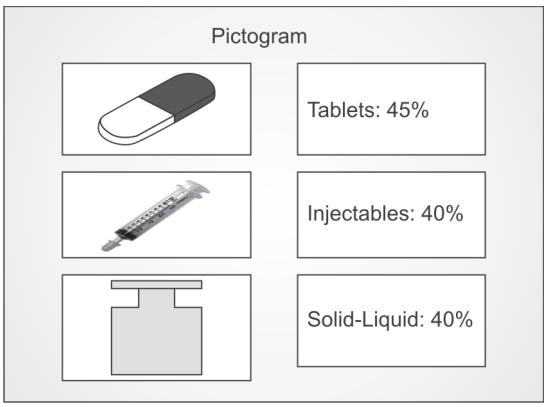


Fig. 5.9: Share of different items in expenditure of a hospital

$$AD = \frac{\Sigma | x_i - m |}{n}$$

$$= \frac{23.4}{20} = 1.17$$

With MS Excel Formula

i. Enter 20 values in Excel sheet starting from cell A_1 as shown in Fig. 8.4.

ii. Select any empty cell (say F_2), find and activate formula AVEDEV (procedure PR1, page 21).

iii. A blank form will be displayed as shown in Fig. 8.2. In box "Number 1", enter the array details which is B_2:B_{21} for our example.

iv. Click OK. Answer displayed will be 1.17.

Using Syntax

The syntax for average deviation is:

= AVEDEV(<Array>) … (Syn 8.4)

This syntax is entered in any empty cell for getting average deviation (in Fig. 8.2, it is shown in cell G_2 and actually entered in cell F_2.

= AVEDEV(B_1:B_{21})

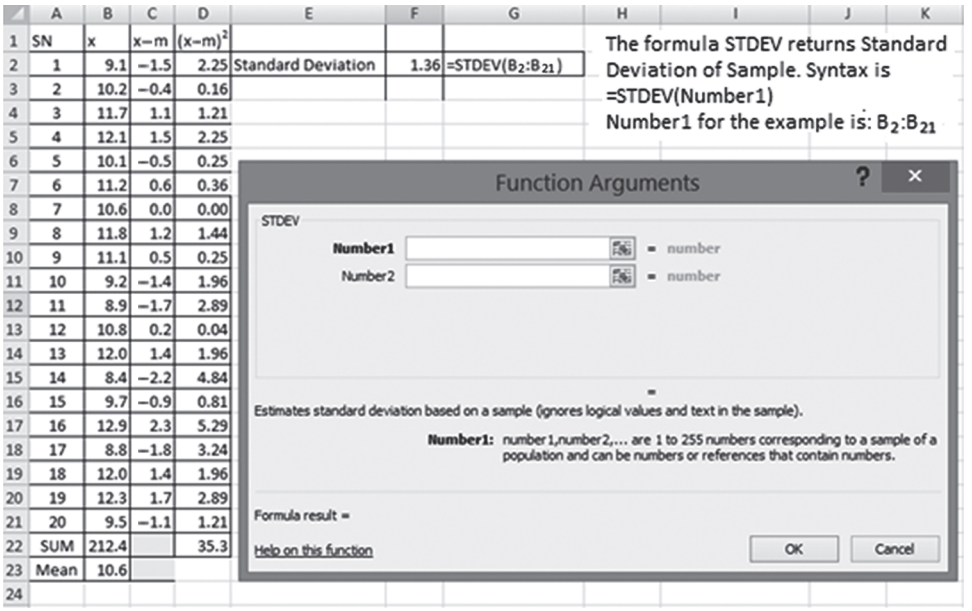
For our example, <Array> is A_1:A_{20}.

8.4 STANDARD DEVIATION (SD)

By definition, *SD* is root mean squared deviation. It is the most commonly used measure of variation.

Example 8.4: We will use data in Fig. 8.4 and calculate *SD*. The steps are as below:

Step	*Result for example*
i. Calculate mean (m)	10.6 (Cell B_{23})
ii. Calculate difference of each observation from mean (x–m). Do not ignore positive/negative sign.	(shown in column *C*, Fig. 8.6)
iii. Calculate square of each difference from mean [$(x–m)^2$]	(shown in column *D*, Fig. 8.6)
iv. Sum all values in step iii	Cell D_{22} = 35.3
v. Divide value at step iv by n–1 = 32.252/(20–1)	1.76 1.36
vi. Calculate square root of value at step v: $\sqrt{\{\Sigma[(x-m)^2]/n-1\}}$	



	A	B	C	D	E	F	G
1	SN	x	x–m	$(x-m)^2$			
2	1	9.1	–1.5	2.25	Standard Deviation	1.36	=STDEV(B2:B21)
3	2	10.2	–0.4	0.16			
4	3	11.7	1.1	1.21			
5	4	12.1	1.5	2.25			
6	5	10.1	–0.5	0.25			
7	6	11.2	0.6	0.36			
8	7	10.6	0.0	0.00			
9	8	11.8	1.2	1.44			
10	9	11.1	0.5	0.25			
11	10	9.2	–1.4	1.96			
12	11	8.9	–1.7	2.89			
13	12	10.8	0.2	0.04			
14	13	12.0	1.4	1.96			
15	14	8.4	–2.2	4.84			
16	15	9.7	–0.9	0.81			
17	16	12.9	2.3	5.29			
18	17	8.8	–1.8	3.24			
19	18	12.0	1.4	1.96			
20	19	12.3	1.7	2.89			
21	20	9.5	–1.1	1.21			
22	SUM	212.4		35.3			
23	Mean	10.6					
24							

Fig. 8.4: Calculation of standard deviation with MS Excel: Example 8.4