DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

V.K. Bhatia
I.A.S.R.I., Library Avenue, New Delhi – 110012
vkbhatia@iasri.res.in

"Garbage in, garbage out " is the rule of data processing. This means that wrong input data or data with serious flaws will always leads to incorrect conclusions, and often, incorrect or harmful actions. In most of practical situations, it is hard to get good basic data, even in simple, non controversial situations and with the best of intentions. With the available basic data, the job of its processing, statistician needs help of various computing equipments such as computer, calculator and mathematical tables etc. Due to the limitations of computing capabilities of these equipments the calculations performed are not always accurate and are subject to some approximations. This means that howsoever fine techniques a statistician may use, if computations are inaccurate, the conclusions he draws from an analysis of numerical data will generally be wrong and very often misleading. It is essential therefore, to look into the sources of inaccuracies in numerical computations and the way to avoid them. In addition to this, before the data is actually processed, it must be ensured that the underlying assumptions for the desired analysis are satisfied because it is well known that the classical statistical techniques behave in the optimum manner under predefined set of conditions and perform badly for the practical situations where they depart significantly from the ideal described assumptions. For these situations thus there is a need to look at the data carefully before finalising the appropriate analysis. This involves checking the quality of the data for the errors, outliers, missing observations or other peculiarities and underlying assumptions. For these rectifications, the question also arises whether the data need to be modified in any way. Further, the main purpose of classification of data and of giving graphical and diagrammatical representation is to indicate the nature of the distribution i.e. to find out the pattern or type of the distribution. Besides the graphical and diagrammatical representation, there are certain arithmetical measures which give a more precise description of the distribution. Such measures also enable us to compare two similar distributions and are helpful for solving some of the important problems of statistical inference.

Thus there is a need to look into these aspects i.e. inaccuracies, checking of abnormal observations, violation of underlying assumptions of data processing and summarization of data including graphical display.

Types of Inaccuracies

It is convenient at the start to make a distinction between different types of accuracies in computational work. A ***blunder*** is a gross inaccuracy arising through ignorance. A statistician who knows his theory rarely commits a blunder. But even when he knows the procedure in detail and use machines for computations, he sometimes makes ***mistakes***. There is a third type of inaccuracy, which we shall call an ***error***. This is different from the other two types in that it is usually impracticable and sometimes even impossible to avoid. In other words an error is an observation which is incorrect, perhaps because it was recorded wrongly in the first place

or because it has been copied or typed incorrectly at some stage. An ***outlier*** is a 'wild' or extreme observation which does not appear to be consistent with the rest of the data. Outliers arise for a variety of reasons and can create severe problems. Errors and outliers are often confused. An error may or may not be an outlier, while an outlier may not be an error.

The search for errors and outliers is an important part of **Initial Data Analysis** (IDA). The terms **data editing** and **data cleaning** are used to denote procedures for detecting and correcting errors. Generally this is an iterative and ongoing process.

Some checks can be made 'by hand', but a computer can readily be programmed to make other routine checks and this should be done. The main checks are for **credibility**, **consistency** and **completeness**. Credibility checks include carrying out a **range test** on each variable. Here a credible range of possible values is pre specified for each variable and every observation is checked to ensure that it lies within the required range. These checks pick up gross outliers as well as impossible values. Bivariate and multivariate checks are also possible. A set of checks, called 'if-then' checks, can be made to assess credibility and consistency between variables.

Another simple, but useful, check is to get a printout of the data and examine it by eye. Although it may be impractical to check every digit visually, the human eye is very efficient at picking out suspect values in a data array provided they are printed in strict column formation in a suitably rounded form. When a suspect value has been detected, the analyst must decide what to do about it. It may be possible to go back to the original data records and use them to make any necessary corrections. In some cases, such as occasional computer malfunctions, correction may not be possible and an observation which is known to be an error may have to be treated as a missing observation.

Extreme observations which, while large, could still be correct, are more difficult to handle. The tests for deciding whether an outlier is significant provide little information as to whether an observation is actually an error. Rather external subject-matter considerations become paramount. It is essential to get advice from people in the field as to which suspect values are obviously silly or impossible, and which, while physically possible, are extremely unlikely and should be viewed with caution. Sometimes additional and further data may resolve the problem. It is sometimes sensible to remove an outlier, or treat it as a missing observation, but this outright rejection of an observation is rather drastic, particularly if there is evidence of a long tail in the distribution. Sometimes the outliers are the most interesting observations.

An alternative approach is to use robust methods of estimation which automatically downweight extreme observations. For example, one possibility for univariate data is to use **Winsorization,** in which an extreme observation is adjusted towards the overall mean, perhaps to the second most extreme value (either large or small as appropriate). However, many analysts prefer a diagnostic approach which highlights unusual observations for further study. Whatsoever amendments are required to be made to the data, there needs to be a clear, and preferably simple, sequence of steps to make the required changes in data.

Missing observations arise for a variety of reasons. A respondent may forget to answer all the questions, an animal may be killed accidentally before a treatment has shown any effect, a scientist may forget to record all the necessary variables or a patient may drop out of a clinical trial etc. It is important to find out why an observation is missing. This is best done by asking 'people in the field'. In particular, there is a world of difference between observations lost through random event, and situations where missing observations are created deliberately. Further the probability that an observation, y, is missing may depend on the value of y and/or on the values of explanatory variables. Only if the probability depends on neither then the observations are said to be **missing completely at random (MCAR)**. For multivariate data, it is sometimes possible to infer missing values from other variables, particularly if redundant variables are included (e.g. age can be inferred from date of birth).

Errors may arise from one or more of the following sources : (a) the mathematical formulation is only an idealized and very seldom an exact description of reality; (b) parameters occurring in mathematical formulae are almost always subject to errors of estimation; (c) many mathematical problems can only be solved by an infinite process, whereas all computations have to be terminated after a finite number of steps; (d) because of the limited digit capacity of computing equipment, computations have to be carried with numbers rounded off conveniently. However, it is not necessary to try to avoid all errors, because usually the final answer need be correct only to a certain number of figures. The theory of calculations with approximate numbers will be subjected to the following errors:

Rounding Off: Because of the limited digit capacity of all computing equipments, computations are generally be carried out with numbers rounded off suitably. To round off a number to n digits, replace all digits to the right of the n-th digit by zeros. If the discarded number contributes less than half a unit in the n-th place, leave the n-th digit unaltered; if it is greater than half a unit, increase the n-th digit by unity; if it is exactly half a unit, leave the n-th digit unaltered when it is an even number and increase it by unity when it is an odd number. For example, the numbers 237.582, 46.85, 3.735 when rounded off to three digits would become 238, 46.8 and 3.74, respectively.

Significant Figures: In a rounded-off number, significant figures are the digits 1, 2,..., 9. Zero (0) is also a significant figure except when it is used to fix the decimal point or to fill the places of unknown or discarded digits. Thus in 0.002603, the number of significant figures is only four. Given a number like 58,100 we cannot say whether the zeros are significant figures or not; to be specific we should write it in the form 5.81×10^4, 5.810×10^4 or 5.8100×10^4 to indicate respectively that the number of significant figures is three, four or five.

Error Involved in the Use of Approximate Numbers : If u is the true value of a number and u_0 an approximation to it, then the error involved is $E = u - u_0$. The relative error is $e = \frac{u - u_0}{u}$ and the percentage error is $p = \frac{100(u - u_0)}{u}$.

Mistakes in Computation

How and Where Mistakes Arise: The only way to avoid mistakes is, of course, to work carefully but a general knowledge about the nature of mistakes and how they arise helps us to work carefully. Most mistakes arise at the stage of copying from the original material to the worksheet or from one worksheet to another, transferring from the worksheet onto the calculating machine or vice versa, and reading from mathematical tables. It is a good idea in any computational program to cut down copying and transferring operations as much as possible. A person who computes should always do things neatly in the first instance and never indulge in the habit of doing "rough work" and then making a fair copy. Computational steps should be broken up into the minimum possible number of unit operations - operations that can be carried out on the calculating machine without having to write down any intermediate answer. Finally the work should be so arranged that it is not necessary to refer to mathematical tables every now and then. As far as possible, all references to such tables should be made together at the same time : this minimizes the possibility of referring to a wrong page and of making gross mistakes in reading similar numbers from the same table. In many mathematical tables, when the first few digits occur repeatedly, they are separated from the body of the table and put separately in a corner; a change in these leading digits in the middle of a row is indicated by a line or some other suitable symbol. We should be careful to read the leading digits correctly from such tables.

Classification of Mistakes: Mistakes in copying, transferring and reading fall into three broad classes: digit substitution, juxtaposition and repetition. One mistake is to substitute hurriedly one digit for another in a number, for instance, 0 for 6, 0 for 9, 1 for 7, 1 for 4, 3 for 8, or 7 for 9. The only remedy is to write the digits distinctly. Another mistake is to alter the arrangement of the digits in a number, to write 32 for 23 or 547 for 457. The third type of mistake occurs when the same number or digit occurs repeatedly. For instance, 12,225 may be copied as 1225 or in the series of numbers 71, 63, 64, 64, 64, . one or more of the 64's may be forgotten. We should be especially careful to avoid these mistakes.

Precautions: Certain general precautions should be taken to avoid mistakes in computations. Whenever possible, we should make provision for checking the accuracy of computation. One way is to make use of mathematical identities and compute the same quantity by different methods. Computations should be properly laid out, in tabular form, with check columns whenever possible. Further before starting on the detailed computations a few extra minutes may be taken for computing mentally as rough answer. This serves a check on the final computation. To summarize, we may lay down the following five principles for avoiding mistakes in computation:

- Write the digits distinctly.
- Cut down copying and transferring operations.
- Use tabular arrangement for computations.
- Keep provision for checking.
- Guess the answer beforehand.

A last word of warning may be helpful. If a mistake is made, it is almost impossible to locate and correct the mistake by going through the original computation, even if this is done a

number of times. The best way out is to work the whole thing afresh, perhaps using a different computational layout altogether.

Data Quality

The quality of the data is of paramount importance and needs to be assessed carefully particularly for the suspicious-looking values, missing observations etc. If there are missing observations then the reasons for their missing and what can be done about them? needs to answered properly. The next question asked is that How were the data collected? What was the format of the questionnaire designed for the sample survey? Were the questions included in the questionnaire practicable to get the reliable information from the respondents? This will help to a great extent in scanning the data for its guanine ness.

Data processing and data editing require careful attention to ensure that the quality of the data is as high as possible. However, it is important to realise that some errors may still get through, particularly with large data sets. Thus diagnostic procedures at the later model-building stage should be carried out to prevent a few data errors from substantially distorting the results. With 'dirty' data containing outliers and missing observations, limited but useful inference may still be possible, although it requires a critical outlook, a knowledge of the subject matter and general resourcefulness on the part of the statistician.

Summary Statistics

After the data have been properly checked for its quality, the first and foremost analysis is usually for the descriptive statistics. The general aim is to summarize the data, iron out any peculiarities and perhaps get ideas for a more sophisticated analysis. The data summary may help to suggest a suitable model which in turn suggests an appropriate inferential procedure. The first phase of the analysis will be described as the initial examination of the data or initial data analysis. It has many things in common with **explanatory data analysis** (EDA) which includes a variety of graphical and numerical techniques for exploring data. Thus EDA is an essential part of nearly every analysis. It provides a reasonably systematic way of digesting and summarizing the data with its exact form naturally varies widely from problem to problem. In general, under initial and exploratory data analysis, the following are given due importance.

Measures of Central Tendency

One of the most important aspects of describing a distribution is the central value around which the observations are distributed. Any arithmetical measure which is intended to represent the center or central value of a set of observations is known as measure of central tendency.

The Arithmetic Mean (or simply Mean)

Suppose that n observations are obtained for a sample from a population. Denote the values of the n observations by $x_1, x_2 ... x_n$; x_1 being the value of the first sample observation, x_2 that of second observation and so on. The arithmetic mean or mean or average denoted by $\bar{x}$ is given by

$$\bar{x}\,(read\ as\ '\ x\ bar'\ ) = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{Sum of\ observations}{No. of\ observations} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The symbol Σ (read as 'sigma') means sum the individual values $x_1, x_2, ..., x_n$ of the variable, X. Usually the limits of the summations are not written, since it is always understood that the summation is over all n values. Hence we can write

$$\bar{x} = \frac{\sum x}{n}$$

The above formula enables us to find the mean when values $x_1, x_2, ..., x_n$ of n discrete observations are available. Sometimes the data set are given in the form of a frequency distribution table then the formula is as follows:

Arithmetic Mean of Grouped Data

Suppose that there are k classes or intervals. Let $x_1, x_2, ..., x_k$ denote the class mid-points of these k intervals and let $f_1, f_2, ..., f_k$ denotes the corresponding frequencies of these classes. Then the arithmetic mean $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{Sum of\ \{(Mid\ point\ of\ class\ interval) \times (frequency\ of\ the\ class)\}}{Total\ frequency}$$

Properties of the arithmetic mean

(a) The Sum of the deviations of a set of n observations $x_1, x_2, ..., x_n$ from their mean $\bar{x}$ is zero. Let d_i as deviation of x_i from $\bar{x}$ then

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

(b) If $x_1, x_2, ..., x_n$ are n observations, $\bar{x}$ is their mean and $d_i = x_i - A$ is the deviation of x_i from a given number A, then

$$\bar{x} = A + \frac{\sum d_i}{n}$$

(c) If the numbers $x_1, x_2, ..., x_n$ occur with the frequencies $f_1, f_2, ..., f_n$ respectively and $d_i = x_i - A$, then

$$\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$$

(d) If in a frequency distribution all the k class intervals are of the same width c, and $d_i = x_i - A$ denote the deviation of x_i from A, where A is the value of a certain mid-point and $x_1, x_2, ..., x_k$ are the class mid-points of the k-classes, then $d_i = c\,u_i$ where $u_i = 0, \pm 1, \pm 2, ...$

and $\bar{x} = A + (\frac{\sum f_i u_i}{\sum f_i})c$

Example 1: For the frequency distribution of weights of sorghum ear-heads given below, calculate the mean value.

Frequency distribution of weights of 190 sorghum ear-heads

Weight of ear-head (in g) (X)	No. of ear-heads(f)
40-60	6
60-80	28
80-100	35
100-120	55
120-140	30
140-160	15
160-180	12
180-200	9
Total	**190**

Computation of mean by direct method

Mid point x	f	f.x.
50	6	300
70	28	1960
90	35	3150
110	55	6050
130	30	3900
150	15	2250
170	12	2040
190	9	1710
Total	**190**	**21360**

The mean weight of ear-heads is given by,

$$\bar{x} = \frac{\sum fx}{n}$$
$$= \frac{21360}{190}$$
$$= 112.4g$$

Computation of mean by short-cum method

x	f	u	fu
50	6	-3	-18
70	28	-2	-56
90	35	-1	-35
110	55	0	0
130	30	1	30
150	15	2	30
170	12	3	36
190	9	4	36
Total	**190**	**(-109 + 132 = 23)**	

In this example, the maximum frequency is 55 and the mid-value against this value is 110. Hence, A = 110.

The mean weight of sorghum ear-heads is then,

$$\bar{x} = A + \left(\frac{\sum fu}{n} \times c \right)$$

$$= 110 + \left(\frac{23}{190} \times 20 \right)$$

$$= 110 + \frac{46}{19} \quad = \quad 112.4 \text{ g}$$

The result is same as in the direct method.

Example 2: For the frequency distribution of seed yield of sesamum given in the following table. Calculate the mean yield per plot.

Frequency distribution of seed yield of sesamum from 100 plots

Yield per plot (in g)	No. of plots
65-84	3
85-104	5
105-124	7
125-144	20
145-164	24
165-184	26
185-204	12
205-224	2
225-244	1
Total	**100**

In this example, the classes are not continuous. Since the yield is given in nearest gram we may take the classes continuous in the following manner. Take the upper-limit of the first class and the lower-limit of the next class and divide their sum by 2. Thus we have (84 + 85)/2 = 84.5. This value will be the upper limit of the first class and lower limit of the next

class. Care should be taken to keep the class intervals unaltered. Thus the lower limit of the first class will be 64.5 and the upper limit of the last class will be 244.5.

Computation of mean for grouped data

Yield (in g) X	Mid value x	f	u	fu
64.5-84.5	74.5	3	-4	-12
84.5-104.5	94.5	5	-3	-15
104.5-124.5	114.5	7	-2	-14
124.5-144.5	134.5	20	-1	-20
144.5-164.5	154.5	24	0	0
164.5-184.5	174.5	26	1	26
184.5-204.5	194.5	12	2	24
204.5-224.5	214.5	2	3	6
224.5-244.5	234.5	1	4	4
Total		**100**		**-61+60 = -1**

The mean yield per plot is

$$\bar{x} = A + \left(\frac{\sum fu}{n} \times c \right) \qquad = 154.5 + \left(\frac{(-1)}{100} \times 20 \right) \quad = \quad 154.5 - 0.2 \text{ or } 154.3 \text{ g}$$

Example 3: The data in the following table are the number of seeds germinated out 5 in each of 50 pots. Find the mean number of seeds that germinate.

Number of seeds germinated out of 5 in each of 50 pots

No. of seeds germinated X	No. of pots(f)	d	fd
0	4	-2	-8
1	13	-1	-13
2	16	0	0
3	9	1	9
4	5	2	10
5	3	3	9
Total	**50**		**7**

$$\bar{x} = A + \left(\frac{\sum fd}{n} \right) \quad = 2 + \left(\frac{7}{50} \right) = 2.14g$$

The Median

The median of a set of n measurements or observations x_1 , x_2 ,..., x_n is the middle value when the measurements are arranged in an array according to their order of magnitude. If n is odd, the middle value is the median. If n is even, there are two middle values and the average of these values is the median. The median is the value which divides the set of observations into two equal halves, such that 50% of the observations lie below the median and 50% above the median. The median is not affected by the actual values of the observations but rather on their positions.

The Median of Grouped Data

The formula of median of grouped data is as

$$\text{Median} = L_m + \left(\frac{N/2 - (\sum f)_o}{f_m} \right) xc$$

where N = Total frequency $= \sum f_i$

f_m = frequency of the class where the median lies.

L_m = Lower class boundary of the class where the median lies.

$(\sum f)_0$ = Sum of frequencies of classes below (or lower than) the class where the median lies.

c = Width of the median class interval.

Example 4: If weights of sorghum ear-heads are 45, 60, 48, 100, 65 g, then the data arrangement will be 45, 48, 60, 65, 100. Since, there are 5 items, the median is $\frac{(5+1)}{2}$ th item, that is, 3rd item. It is 60 g.

Example 5: If the weights of sorghum ear-heads are 45, 48, 60, 65, 65, 100 g, then the median is $\frac{(6+1)}{2}$ = (3.5)th item. That is,

Median = 3rd item + (4^{th} item - 3^{rd} item) × (0.5)

= 60 + (65 - 60) ((0.5)

= 60 + 2.5 or 62.5 g

Example 6: For the frequency distribution of weights of sorghum ear-heads given in the following table, calculate the median.

The frequency distribution of weights of 190 sorghum ear-heads

Weight of ear-head (in g)	No. of ear-heads
40 - 60	6
60 - 80	28
80 - 100	35
100 - 120	55
120 - 140	30
140 - 160	15
160 - 180	12
180 - 200	9
Total	**190**

This table can further be arranged as

Weight of ear-head (in g)	No. of ear-heads	Less than class	Cumulative frequency
40 - 60	6	< 60	6
60 - 80	28	< 80	34
80 - 100	35	< 100	69
			$\rightarrow \frac{n}{2} = 95$
100 - 120	55	< 120	124
120 - 140	30	< 140	154
140 - 160	15	< 160	169
160 - 180	12	< 180	181
180 - 200	9	< 200	190
Total	**190**		

In our example, (n / 2) = (190 / 2) = 95. This value lies in between 69 and 124, and less than classes corresponding to these values are 100 and 120, respectively. Hence the median class is 100 - 120 and the lower limit of this class is 100. The cumulative frequency upto 100 is 69 and the frequency of the median class, 100 - 120 is 55.

Therefore,

$$\text{Median} = 100 + \left[\frac{(95-69)}{55} \times 20\right] \quad = \quad 100 + \left[\frac{26}{55} \times 20\right]$$

$$= 100 + 9.45 \text{ or } 109.45 \text{ g}$$

The Mode

The mode is the observation which occurs most frequently in a set. In grouped data mode is worked out as

$$\text{Mode} = l + \left(\frac{f_s}{f_p + f_s} \times c\right)$$

where

l = Lower limit of the modal class.

f_s = the frequency of the class succeeding the modal class .

f_p = the frequency of the class preceding the modal class

c = Width of the class interval.

The mode can be determined analytically in the case of continuous distribution. For a symmetrical distribution, the mean, median and mode coincide. For a distribution skewed to the left (or negatively skewed distribution), the mean, the median and the mode are in that order (as they appear in the dictionary) and for a distribution skewed to the right (or positively skewed distribution) they occur in the reverse order, mode, median and mean. There is an empirical formula for a moderately asymmetrical skewed distribution, it is given by

Mean - Mode = 3 (Mean - Median)

Example 7: If the yield of paddy from different fields are 6.0, 4.9, 6.0, 5.8, 6.2, 6.0, 6.3, 4.8, 6.0, 5.7 and 6.0 tonnes per hectare, the modal value is 6.0 tonnes per hectare.

Example 8: For the frequency distribution of weights of sorghum ear-heads given in the following table, calculate the mode.

Weight of ear-head (in g)	**No. of ear-heads**	
40 - 60	6	
60 - 80	28	
80 - 100		35 $\rightarrow f_p$
100 - 120	55	
120 - 140		30 $\rightarrow f_s$
140 - 160	15	
160 - 180	12	
180 - 200	9	
Total	**190**	

For calculating the mode for grouped data first find out the modal class. The modal class is the class against the maximum frequency. In our example, the maximum frequency is 55 and hence the modal class is 100 - 120.

In our example, l = 100, f_p = 35 and f_s = 30. Hence,

$$\text{Mode} = 100 + \left[\frac{30}{35+30} \times 20\right] = 100 + \left[\frac{600}{65}\right] = 109.23 \text{ g}$$

The Geometric Mean

There are two other averages, the geometric mean and harmonic mean which are sometimes used. The Geometric Mean (GM) of a set of observations is such that its logarithm equals the arithmetic mean of the logarithms of the values of the observations.

$$\mathbf{GM = (x_1\ x_2 ..... x_n)^{1/n}}$$

log *GM = 1/n* ($\sum$ *log* x_i) or in frequency distribution, log *GM* = 1/n ($\sum f_i$ *log* x_i)

Example 9: Let 2, 4, 8, 16 be the 4 items. Their geometric mean is calculated as follows:

x	**Log x**
2	0.3010
4	0.6021
8	0.9031
16	1.2041
Total	**3.0103**

Mean (of log values) = $\frac{\sum \log x}{n}$ = $\frac{3.0103}{4}$ = 0.7526

Antilog(0.7526) = 5.66 Therefore, G.M. = 5.66

In case of frequency distribution,

$$\text{G.M.} = \text{Antilog}\left[\frac{\sum f(\log x)}{n}\right]$$

The geometric mean can be obtained only if the values assumed by the observation are positive(greater than zero).

Harmonic mean

The Harmonic Mean (HM) of a set of observations is such that its reciprocal is the arithmetic mean of the reciprocals of the values of the observation

$$\frac{1}{HM} = \frac{\sum_{i=1}^{n} 1/x_i}{n}$$

or in frequency distribution

$$\frac{1}{HM} = \frac{\sum f_i(1/x_i)}{f_i}$$

Example 10: There are 5 agricultural labourers. They can complete weeding operations on a 100 square meter land in 4, 5, 5, 6 and 7 hours, respectively. If these 5 labourers are employed for weeding in 500 square meter area, in how many hours will they complete the work?

For such problems we compute harmonic mean.

$$\text{H.M.} = \frac{5}{(1/4)+(1/5)+(1/5)+(1/6)+(1/7)}$$

$$= \frac{5}{(0.25+0.20+0.20+0.167+0.143}$$

$$= \frac{5}{0.960} = 5.21 \text{ hours}$$

The harmonic mean is rarely computed for a frequency distribution.

Weighted Mean

If there are n observations, x_1, x_2, $x_{3,\ldots,}x_n$ with corresponding weights w_1, w_2, $w_{3,\ldots,}w_n$, then the weighted mean is given by,

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

Example 11: The average yield of IR 20 paddy from two localities are 55 quintals and 65 quintals per hectare respectively. The averages were based on 20 hectares and 10 hectares respectively. What is the combined average yield per hectare?

The simple average of 55 and 65 quintals will give us (55+65)/2 = 60 quintals per hectare. This is not correct. We have to compute the weighted mean.

$$\bar{x}_w = \frac{(55\times 20)+(65\times 10)}{20+10}$$

$$= \frac{1100+650}{30} = \frac{1750}{30}$$

$$= 58.3 \text{ quintals per hectare}$$

In computing the mean, we take the frequency of a class as its weight. That is $\bar{x} = \frac{\sum fx}{\sum f}$.

Hence, it is a special case of weighted mean. The three means are related by

A.M. $\geq$ G.M. $\geq$ H.M.

Important characteristics of a good average

Since an average is a representative item of a distribution it should possess the following properties :

1. It should take all items into consideration.
2. It should not be affected by extreme values.
3. It should be stable from sample to sample.
4. It should be capable of being used for further statistical analysis.

Mean satisfies all the properties excepting that it is affected by the presence of extreme items. For example, if the items are 5, 6, 7, 7, 8 and 9 then the mean, median and mode are all equal to 7. If the last value is 30 instead of 9, the mean will be 10, whereas median and mode are not changed. Though median and mode are better in this respect they do not satisfy the other properties. Hence mean is the best average among these three.

When to use different averages

The proper average to be used depends upon the nature of the data, nature of the frequency distribution and the purpose.

If the data is qualitative one, only mode can be computed. For example, when we are interested in knowing the typical soil type in a locality or the typical cropping pattern in a region we can use mode. On the other hand, if the data is quantitative one, we can use any one of the averages

If the data is quantitative, then we have to consider the nature of the frequency distribution. When the frequency distribution is skewed (not symmetrical) the median or mode will be proper average. In case of raw data in which extreme values, either small or large, are present, the median or mode is the proper average. In case of a symmetrical distribution either mean or median or mode can be used. However, as seen already, the mean is preferred over the other two.

When we are dealing with rates, speed and prices we use harmonic mean. If we are interested in relative change, as in the case of bacterial growth, cell division etc., geometric mean is the most appropriate average.

Measures of Dispersion

We know that averages are representatives of a frequency distribution but they fail to give a complete picture of the distribution. They do not tell anything about the scatterness of observations within the distribution.

Suppose that we have the distribution of the yields (kg per plot) of two paddy varieties from 5 plots each. The distribution may be as follows:

Variety I	45	42	42	41	40
Variety II	54	48	42	33	30

It can be seen that the mean yield for both varieties is 42 kg. But we can not say that the performance of the two varieties is same. There is greater uniformity of yields in the first variety whereas there is more variability in the yields of the second variety. The first variety may be preferred since it is more consistent in yield performance. From the above example, it is obvious that a measure of central tendency alone is not sufficient to describe a frequency distribution. In addition to it we should have a measure of scatterness of observations. The scatterness or variation of observations from their average is called the *dispersion*. There are different measures of dispersion like the range, the quartile deviation, the mean deviation and the standard deviation.

Range

The simplest measure of dispersion is the range. The range is the difference between the minimum and maximum values in a group of observations for example, suppose that the *kapas* yields (kg per plot) of a cotton variety from five plots are 8, 9, 8, 10 and 11. The range is (11 - 8) = 3 kg. In practice the range is indicated as 8 - 11 kg.

Range takes only the maximum and minimum values into account and not all the values. Hence it is a very unstable or unreliable indicator of the amount of deviation. It is affected by extreme values. In the above example, if we have 15 instead of figure 11, the range will be (8 - 15) = 7 kg. In order to avoid these difficulties another measure of dispersion called quartile deviation is preferred.

Quartile Deviation

We have already defined the quartiles. We can delete the values below the first quartile and the values above the third quartile. It is assumed that the unusually extreme values are eliminated by this way. We can then take the mean of the deviations of the two quartiles from the second quartile (median). That is,

$$\frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{(Q_3 - Q_1)}{2}$$

This quantity is known as the quartile deviation (Q.D.).

Example 12: The following are the paddy yields (kg/plot) from 14 plots: 30, 32, 35, 38, 40, 42, 48, 49, 52, 55, 58, 60, 62 and 65 (after arranging in ascending order).

$$\text{Q. D.} = \frac{(Q_3 - Q_1)}{2} = \frac{55.75 - 37.25}{2} = \frac{18.5}{2} = 9.25g$$

The quartile deviation is more stable than the range as it depends on two intermediate values. This is not affected by extreme values since the extreme values are already removed. However, quartile deviation also fails to take the values of all deviations.

Mean Deviation

Mean deviation is the mean of the deviations of individual values from their average. The average may be either mean or median. For raw data the mean deviation from the median is the least. Therefore, median is considered to be most suitable for raw data. But usually the mean is used to find out the mean deviation. The mean deviation is given by

M.D. = $\frac{\sum |x - \bar{x}|}{n}$ for raw data and M.D. = $\frac{\sum f|x - \bar{x}|}{n}$ for grouped data

All positive and negative differences are treated as positive values. Hence we use the modulus symbol $|\ |$. We have to read $|x - \bar{x}|$ as "modulus *x* - $\bar{x}$". If we take *x* - $\bar{x}$ as such, the sum of the deviations, $\sum (x - \bar{x})$ will be 0. Hence, if the signs are not eliminated the mean deviation will always be 0, which is not correct.

Example 13: The *kapas* yields (in kg per plot) of a cotton variety from seven plots are 5, 6, 7, 7, 7, 8 and 9. The mean deviation for this data is computed as follows:

Mean = (5 + 6 + 7 + 7 + 7 + 8 + 9) / 7 = 49 / 7 = 7 kg

$$\text{M.D.} = \frac{|5-7| + |6-7| + |7-7| + |7-7| + |7-7| + |8-7| + |9-7|}{7} = \frac{6}{7}kg$$

Example 14: From field with large number of seasamum plants of a variety, 100 plants were selected at random. The seed yields (in g) per plant were recorded. The results are presented in the following Table. Find the mean deviation for the data.

Seed yield (X) in g	**Number of plants (f)**	**Mid class (x)**	$\lvert x - \bar{x}\rvert = \lvert x - 6.7\rvert$	f $\lvert x - \bar{x}\rvert$
2.5 - 3.5	4	3	3.7	14.8
3.5 - 4.5	6	4	2.7	16.2
4.5 - 5.5	10	5	1.7	17.0
5.5 - 6.5	26	6	0.7	18.2
6.5 - 7.5	24	7	0.3	7.2
7.5 - 8.5	15	8	1.3	19.5
8.5 - 9.5	10	9	2.3	23.0
9.5 - 10.5	5	10	3.3	16.5
Total	**100**			**132.4**

M.D. = $\frac{\sum f|x - \bar{x}|}{n}$ = 132.4 / 100 = 1.324 g

The steps of computation are as follows :

Step 1: If the classes are not continuous we have to make them continuous. In this case they are continuous.
Step 2: Find out the mid values of the classes (mid - $X = x$).
Step 3: Compute the mean.
Step 4: Find out $|x - \bar{x}|$ for all values of x.
Step 5: Multiply each $|x - \bar{x}|$ by the corresponding frequencies.
Step 6: Use the formula.

The mean deviation takes all the values into consideration. It is fairly stable compared to range or quartile deviation. Since, the mean deviation ignors signs of deviations, it is not possible to use it for further statistical analysis and it is not stable as standard deviation which is defined as:

Standard Deviation

Ignoring the signs of the deviations is mathematically not correct. We may square the deviation to make a negative value as positive. After calculating the average squared deviations, it can be expressed in original units by taking its square root. This type of the measure of variation is known as Standard Deviation.

The standard deviation is defined as the square root of the mean of the squared deviations of individual values from their mean. Symbolically,

Standard Deviation (S.D.) or $\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$ or $= \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n}}$

This is called standard deviation because of the fact that it indicates a sort of group standard spread of values around their mean. For grouped data it is given as

Standard Deviation (S.D.) or $\sigma = \sqrt{\frac{\sum f x^2 - \frac{\left(\sum f x\right)^2}{n}}{n}}$

The sample standard deviation should be an unbiased estimate of the population standard deviation because we use sample standard deviation to estimate the population standard deviation. For this we substitute n - 1 for n in the formula. Thus, the sample standard deviation is written as

$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}}$

For grouped data it is given by

$$s = \sqrt{\frac{\sum f x^2 - \frac{\left(\sum f x\right)^2}{n}}{n-1}} \qquad \text{Alternatively,} \quad s = \sqrt{\left[\frac{\sum f d^2 - \frac{\left(\sum f d\right)^2}{n}}{n-1}\right]} \times C$$

where,

C = class interval

$d = (x - A) / C$ as given under mean.

The square of the standard deviation is known as the variance. In the analysis of variance technique, the term $\sum x^2 - \frac{(\sum x)^2}{n}$ is called the sum of squares, and the variance is called the mean square. The standard deviation is denoted by s in case of sample, and by σ (read 'sigma') in case of population.

Example 15: For the data in example 1, the standard deviation is computed as follows :

$$\bar{x} = \frac{5+6+7+7+7+8+9}{7} = 7kg$$

The deviations $(x - \bar{x})$ are (5 - 7), (6 - 7), (7 - 7), (7 - 7), (7 - 7), (8 - 7) and (9 - 7). Therefore,

$$s = \sqrt{\frac{(-2)^2 + (-1)^2 + (0)^2 + (0)^2 + (0)^2 + (1)^2 + (2)^2}{7-1}} = \sqrt{\frac{10}{6}} = 1.29 \text{ kg}$$

The second method is,

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}}$$

$$= \frac{(5^2 + 6^2 + 7^2 + 7^2 + 7^2 + 8^2 + 9^2) - \frac{(49)^2}{7}}{7-1}$$

$$= \sqrt{\frac{353 - \frac{2401}{7}}{6}} = \sqrt{\frac{353-343}{6}} = \sqrt{\frac{10}{6}} = 1.29 \text{ kg}$$

The variance is given by, $s^2 = 1.67 \text{ kg}^2$

Example 16: For the data given in the example 2 compute the standard deviation and variance.

Seed yield (X) in g	Number of plants (f)	Mid class (x)	$(x - A)/C = d$	fd	fd^2
2.5 - 3.5	4	3	-3	-12	36
3.5 - 4.5	6	4	-2	-12	24
4.5 - 5.5	10	5	-1	-10	10
5.5 - 6.5	26	6	0	0	0
6.5 - 7.5	24	7	1	24	24
7.5 - 8.5	15	8	2	30	60
8.5 - 9.5	10	9	3	30	90
9.5 - 10.5	5	10	4	20	80
Total	**100**			**70**	**324**

Here A = 6 and C = 10. The standard deviation is given by

$$s = \sqrt{\left[\frac{\sum fd^2 - \frac{\left(\sum fd\right)^2}{n}}{n-1}\right]} \times C$$

$$= \sqrt{\left[\frac{324 - (70^2/100)}{100-1}\right]} \times 1 = \sqrt{\left[\frac{324-49}{99}\right]} = \sqrt{\frac{275}{99}} = \sqrt{2.7676} = 1.66g$$

Therefore the variance is $s^2 = 2.77\ g^2$

The standard deviation is the most widely used measure of dispersion. It takes all the items into consideration. It is more stable compared to other measures. However, it will be inflated by extreme items as is the mean.

The standard deviation has some additional special characteristics. It is not affected by adding or subtracting a constant value to each observed value. It is affected by multiplying or dividing each observation by a constant. When the observations are multiplied by a constant, the resulting standard deviation will be equivalent to the product of the actual standard deviation and the constant. (Note that division of all observations by a constant, C is equivalent to multiplication by its reciprocal, 1/C. Subtracting a constant C is equivalent of adding a constant, - C.)

Example 17: Suppose we have a set of numbers 1, 2, 3, 4 and 5. Then we have the following results

Original values			**After adding a constant, 2**			**After multiplying a constant, 2**		
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-2	4	3	-2	4	2	-4	16
2	-1	1	4	-1	1	4	-2	4
3	0	0	5	0	0	6	0	0
4	1	1	6	1	1	8	2	4
5	2	4	7	2	4	10	4	16
Total=15		10	25		10	30		40
$s = \sqrt{10/4} = 1.58$			$s = \sqrt{10/4} = 1.58$			$s = \sqrt{40/4} = 3.16 = 2(1.58)$		

When the values are multiplied by 2, the resultant standard deviation is twice the actual one.

The standard deviations can be pooled. If the sum of squares for the first distribution with n_1 observations is SS_1, and the sum of squares for the second distribution with n_2 observations is SS_2, then the pooled standard deviation is given by,

$$s \text{ (pooled)} = \sqrt{\frac{SS_1 + SS_2}{n_1 + n_2 - 2}}$$

Measures of Relative Dispersion

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they are expressed in different units of measurement, we can not use the standard deviations as such for comparing their variability. We have to use the relative measures of dispersion in such situations.

There are relative dispersions in relation to range, the quartile deviation, the mean deviation, and the standard deviation. Of these, the coefficient of variation which is related to the standard deviation is important. The coefficient of variation is given by,

C.V. = (S.D. / Mean) x 100

The C.V. is a unit-free measure. It is always expressed as percentage. The C.V. will be small if the variation is small of the two groups, the one with less C.V. is said to be more consistent.

The coefficient of variation is unreliable if the mean is near zero. Also it is unstable if the measurement scale used is not ratio scale. The C.V. is informative if it is given along with the mean and standard deviation. Otherwise, it may be misleading.

Example 18: Consider the distribution of the yields (per plot) of two paddy varieties. For the first variety, the mean and standard deviation are 60 kg and 10 kg, respectively. For the second variety, the mean and standard deviation are 50 kg and 9 kg, respectively. Then we have, for the first variety,
C.V. = (10/60) x 100 = 16.7 %
For the second variety,
C.V. = (9/50) x 100 = 18.0 %
It is apparent that the variability in first variety is less as compared to that in the second variety. But in terms of standard deviation the interpretation could be reverse.

Example 19: Consider the measurements on yield and plant height of a paddy variety. The mean and standard deviation for yield are 50 kg and 10 kg respectively. The mean and standard deviation for plant height are 55 cm and 5 cm, respectively.

Here the measurements for yield and plant height are in different units. Hence, the variability can be compared only by using coefficient of variation. For yield,
C.V. = (10 / 50) x 100 = 20 %
For plant height,
C.V. = (5 / 55) x 100 = 9.1 %
The yield is subject to more variation than the plant height.

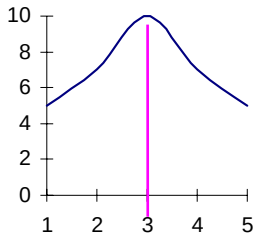
Skewness and Kurtosis
The average and measure of dispersion can describe the distribution but they are not sufficient to describe the nature of the distribution. For this purpose we use other concepts known as Skewness and Kurtosis.

Skewness
Skewness means lack of symmetry. A distribution is said to be symmetrical when the values are uniformly distributed around the mean. For example, the following distribution is symmetrical about its mean 3.

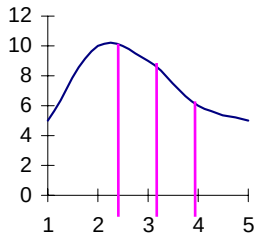
x	:	1	2	3	4	5
frequency (f)	:	5	9	12	9	5

In a symmetrical distribution the mean, median and mode coincide, that is,
mean = median = mode.

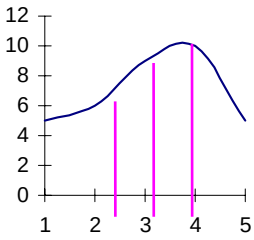
The symmetrical and skewed distributions are shown by curves as



Mean = Median = Mode



Mode >Med> Mean



Mean< Med<Mode

Several measures are used to express the direction and extent of skewness of a dispersion. The important measures are that given by Pearson. The first one is the Coefficient of Skewness:

$$S_k = \frac{3(mean - median)}{S\tan dard\, deviation}$$

For a symmetric distribution $S_k = 0$. If the distribution is negatively skewed then S_k is negative and if it is positively skewed then S_k is positive. The range for S_k is from -3 to 3.

The other measure uses the β (read 'beta') coefficient which is given by, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ where, μ_2 and μ_3 are the second and third central moments. The second central moment μ_2 is nothing but the variance. The sample estimate of this coefficient is $b_1 = \frac{m_3^2}{m_2^3}$ where m_2 and m_3 are the sample central moments given by m_2 = variance =

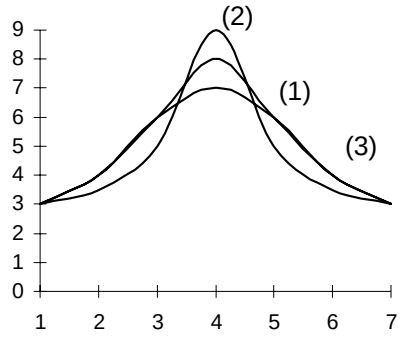
$$\frac{\sum (x-\bar{x})^2}{n-1} or \frac{\sum f(x-\bar{x})^2}{n-1}$$

and $m_3 = \frac{\sum (x-\bar{x})^3}{n-1} or \frac{\sum f(x-\bar{x})^3}{n-1}$

For a symmetrical distribution $b_1 = 0$. Skewness is positive or negative depending upon whether m_3 is positive or negative.

Kurtosis

A measure of the peakness or convexity of a curve is known as Kurtosis.



It is clear from the above figure that all the three curves, (1), (2) and (3) are symmetrical about the mean. Still they are not of the same type. One has different peak as compared to that of others. Curve (1) is known as mesokurtic (normal curve); Curve (2) is known as leptocurtic (leading curve) and Curve (3) is known as platykurtic (flat curve). Kurtosis is measured by Pearson's coefficient, β_2 (read 'beta - two').It is given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$. The sample estimate of this coefficient is $b_2 = \frac{m_4}{m_2^2}$ where, m_4 is the fourth central moment given by $m_4 = \frac{\sum (x - \bar{x})^4}{n-1}$ or $\frac{\sum f\,(x - \bar{x})^4}{n-1}$

The distribution is called normal if $b_2 = 3$. When b_2 is more than 3 the distribution is said to be leptokurtic. If b_2 is less than 3 the distribution is said to be platykurtic.

Example 20: The measures of skewness and kurtosis are given as below:

x	**f**	**(x - $\bar{x}$) = d**	**f d**	**f d^2**	**f d^3**	**f d^4**
3	4	-3.7	-14.8	54.76	-202.612	749.6644
4	6	-2.7	-16.2	43.74	-118.098	318.8646
5	10	-1.7	-17.0	28.90	- 49.130	83.5210
6	26	-0.7	-18.2	12.74	- 8.918	6.2426
7	24	0.3	7.2	2.16	0.648	0.1944
8	15	1.3	19.5	25.35	32.955	42.8415
9	10	2.3	23.0	52.90	121.670	279.8410
10	5	3.3	16.5	54.45	179.685	592.9605
Total	**100**		**0**	**275.00**	**- 43.800**	**2074.1300**

m_2 or Variance = $\frac{\sum f(x-\bar{x})^2}{n-1} = \frac{\sum f d^2}{n-1}$ = 275.000 / 99 = 2.7777

$m_3 = \frac{\sum f(x-\bar{x})^3}{n-1} = \frac{\sum f d^3}{n-1}$ = - 43.800 / 99 = - 0.4424

$$m_4 = \frac{\sum f(x-\bar{x})^4}{n-1} = \frac{\sum f d^4}{n-1} = 2074.1300 / 99 = 20.9508$$

$$b_1 = \frac{(-0.4424)^2}{(2.7777)^3} = \frac{0.1957}{21.4376} = 0.0091$$

$$b_2 = \frac{20.9508}{(2.7777)^2} = \frac{20.9508}{7.7156} = 2.7153$$

Since b_1 is 0.0091, it is only slightly skewed. It is negatively skewed since m_3 is negative. The value of b_2 is 2.7153 which is less than 3. Hence the distribution is platykurtic.

Exploring of data

The first step of data analysis is the detailed examination of the data. There are several important reasons for examining data carefully before the actual analysis is carried out. The first reason for examination of data is for the mistakes which occur at various stages right from recording to entering the data on computer. The next step is to explore the data. The technique of exploratory data analysis is very useful in getting quick information, behaviour and structure of the data. Whereas the classical statistical techniques are designated to be best when stringently assumptions hold true. However it is seen that these techniques fail miserably in the practical situation where the data deviate from the ideal described conditions. Thus the need for examining data is to look into methods which are robust and resistant instead of just being the best in a narrowly defined situation. The aim of exploratory data analysis is to look into a procedure which is best under broad range of situations. The main purpose of exploratory data analysis is to isolate patterns and features of the data which in turn are useful for identifying suitable models for analysis. Another feature of exploratory approach is flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive steps of analysis uncover.

Graphical Representation of Data

The most common data structure is a collection of batch of numbers. This simple structure, in case of large number of observations, is sometimes difficult to study and scan thoroughly with just looking into it. In order to concise the data, there are number of ways by which the data can be represented graphically. The histogram is a commonly used display. The range of observed values is subdivided into equal intervals and then the cases in each interval are obtained. The length of the interval is directly proportional to the number of cases within it. A display closely related to the histogram is the stem-and-leaf plot.

Stem-and-leaf Display

The stem-and-leaf plot provides more information about the actual values than does a histogram. As in the histogram, the length of each bar corresponds to the number of cases that fall into a particular interval. However, instead of representing all cases with a same symbol, the stem-and-leaf plot represents each case with a symbol that corresponds to the actual observed value. This is done by dividing observed values into two components - the

leading digit or digits, called the stem and the trailing digit called the leaf. The main purpose of stem-and leaf display is to throw light on the following :

(1) Whether the pattern of the observation is symmetric.
(2) The spread or variation of observation.
(3) Whether a few values are far away from the rest.
(4) Points of concentration in data.
(5) Areas of gaps in the data.

Example 21: For the data values 22.9, 26.3, 26.6, 26.8, 26.9, 26.9, 27.5, 27.6, 27.6, 28.0, 28.4, 28.4, 28.5, 28.8, 28.8, 29.4, 29.9, 30.0. Display stem and leaf diagram.

For the first data value of 22.9

Data value	Split	Stem	and	Leaf
22.9	22/9	22		9

Then we allocate a separate line in the display for each possible string of leading digits (the stem), the necessary lines run from 22 to 31. Finally we write down the first trailing digit (the leaf) of each data value on the line corresponding to its leading digits.

(Unit = 1 day)

22	:	9					
23	:						
24	:						
25	:						
26	:	3	6	8	9	9	
27	:	5	6	6			
28	:	0	4	4	5	8	8
29	:	4	9				
30	:	0	3				
31	:	2	8				

Sometimes, there are too many leaves per line (stem) then in that case it is desired to split lines and repeat each stem.

0	*	(Putting leaves 0 through 4)
0	.	(Putting 5 through 9)
1	*	
1	.	
2	*	
2	.	

In such a display, the interval width is 5 times a power of 10. Again, even if for two lines it is crowded then we have a third form, five lines per stem.

0*
t
f
s
0.

With variables 0 and 1 on the * line, 2 (two) and 3 (three) on the t line, 4 (four) and 5 (five) on the *f* line, 6 (six) and 7 (seven) on the *s* line and 8 and 9 on the . line.

The Box-plot

Both the histogram and the stem-and-leaf plots are useful for studying the distribution of observed values. A display that further summarizes information about the distribution of the values is the box-plot. Instead of plotting the actual values, a box plot displays summary statistics for the distribution. It plots the median, the 25th percentile, 75th percentile and values that are deviating from the rest. Fifty percent of the cases lie within the box. The length of the box corresponds to the interquartile range, which is the difference between the Ist and 3rd quartiles. The box plot identifies extreme values which are more than 3 box-lengths from the upper or lower edge of the box. The values which are more than 1.5 box-lengths are characterized as outliers. The largest and the smallest observed values are also part of the box-plot in terms of edges of lines. The median which is a measure of location lies within the box. The length of box depicts the spread or variability of observations. If the median is not in the center of the box, the values are skewed. If the median is closer to the bottom of the box than the top, the data are positively skewed. If the median is closer to top then the data are negatively skewed.

Spread-versus-level plot

When a comparison of batches shows a systematic relationship between the average value or level of a variable and the variability or spread associated with it, then it is of interest to search for a re-expression, or transformation of the raw data that reduces or eliminates this dependency. If such a transformation can be found, the re-expressed data will be better suited both for visual exploration and for analysis. This will further make analysis of variance techniques valid and more effective, when there is exactly or approximately equal variance across groups. The spread-versus-level plot is useful for searching an appropriate power transformation. By power transformation it is meant as power i.e. searching a power (or exponent) p as the transformation that replaces x by x^p . The power can be estimated from the slope of line in the plot of log of the median against the log of the interquartile range i.e. IR α M_d = c M_d or log IR = log c + B log M_d . The power is obtained by subtracting the slope from 1. (i.e. Power = 1 - slope). This is based on the concept that transformation Z = x^{1-b} of the data given re-expressed value Z whose interquartile range or spread does not depend at least approximately on the level. In addition to this graphical method of judging the independence of spread and level, there is a test known as Levene Test for testing the homogeneity of variances.

Although there is a wide variety of tests available for testing the equality of variances, but many of them are heavily dependent on the data being samples from normal populations. Analysis of variance procedures on the other hand are reasonably robust to departures from normality. The Levene test is a homogeneity of variance test that is less dependent on the assumption of normality than most tests and thus is all the more important with analysis of variance. It is obtained by computing for each case the absolute difference from its cell mean and then performing a one-way analysis of variance on these differences.

Examination of Normality

As the normal distribution is very important for statistical inference point of view so it is desired to examine the assumption to test whether the data is from a normal distribution. The

normality can be tested by plotting a normal plot. In a normal probability plot each observed value is paired with its expected value from the normal distribution. In a situation of normality, it is expected that points will fall on straight line. In addition to this a plot of deviation from straight line can also be plotted as detrended normal plot. A structure-less detrended normal plot confirms normality. These two plots give a visual basis for examining normality. Besides these visual displays, the statistical tests are Shappiro-Wilks and the Lilliefors. The Lilliefors test is based on the modification of the Kolmogorov-Smirov test for the situation when means and variances are not known but are estimated from the data. The Shapiro-Wilks test is more powerful in many situations as compared to other tests.

Robust Location Estimation for Comparing Groups

In comparison of groups based on exploratory data analysis, median (a resistant location estimator) and spread play a very crucial role because wild observations have no effect on them significantly. Besides estimating median or spread for simply exploring the data for comparison purpose, it is also desirable to consolidate them in very concise manner in terms of one or two reliable estimates of parameters which take into account all the observations. The simplest parameter one can think of, is that of arithmetic mean to estimate central tendency or location. Further it is well known that mean is heavily influenced by outliers. One very small or large observation may change mean drastically or in other words mean is not a resistant estimator of location. In addition to the resistant to wild observations, it is also of interest to look into a estimator which is robust against underlying distributional assumptions.

A simple robust estimator of location is trimmed mean which is obtained by excluding the extreme values. Like median it is also not affected by extremes. The advantage of trimmed mean in comparison to median is that it utilises more observations or in other words it makes better use of the data. The trimmed mean is also thought of weighted mean in which observations included are attached weight of unity and observations excluded are attached with zero weights. For robust estimation an another alternative is to include the extremes with smaller weights than the cases that are closer to the center. There could be various ways of assigning weights to the observations and this leads to thus many generalized maximum-likelihood estimators of location popularly known as M-estimators. Thus the logic of the M-estimators is assigning of weights to observations which is inversely proportional to the distance from the center (or any measure of location).

In robust estimation, the numerical values of estimates are essentially obtained by an iterative process because the estimators do not have closed forms like the mean or variance. Estimates are obtained by minimizing a function. This is achieved by assuming some starting value, say m_0 of location and then computing a new estimate m_*. Then this new estimate can become the starting value for another round, and the process continues until the estimates at two different iterations do not differ much. In the course of minimising a function, different weights are attached to different observations depending upon their deviation or relative deviation from the location estimate. For instance in Huber's estimator the observations are attached equal weights of unity upto a critical point c, and thereafter weights decrease as standardised distance from the location estimate increase.

The different M-estimators differ from each other in respect of weights assign to observations. The Tukey biweight does not have a point at which weights shift abruptly from 1. Instead weights decline to 0. Cases with values greater than c standardised units from the estimates are assigned zero weights. Hampel's M-estimator has a more complicated weighting scheme than the Huber or Tukey biweights. In this case the range is divided by 3 cutoff points say a, b, c. Cases below 'a' are given weights unity and cases above 'c' are given zero weights. Cases between 'a' and 'b', and 'b' and 'c' are assigned weights according to standardised distances. In case of Andrew's M-estimator, there is no abrupt change in the assignment of weights. A smooth function replaces the separate pieces.

There are mainly three categories of robust estimation of location. Their logic and underlying concepts are

L-Estimators: A class of estimators, called L-estimators is defined by a linear combination of the order statistics. This class includes the sample mean, median and trimmed means as special cases.

R-Estimators: R-Estimators are derived from rank tests for a shift between samples, estimating the shift by moving one sample along until the test is least able to detect a shift. In estimating location, the second sample is the mirror image of the actual sample, reflected about the estimate, so that an R-estimator of location minimises the shift between the sample and its mirror image, as measured by the rank test.

M-Estimators: M-estimators minimise functions of the deviations of the observations from the estimate that are more general than the sum of squared deviations or the sum of absolute deviations. In this way the class of M-estimators includes the mean and the median as special classes. Viewed in another way, M-estimators generalise the idea of the maximum-likelihood estimator of the location parameter in a specified distribution. Thus it is reasonable to expect that a suitably chosen M-estimator will have good robustness of efficiency in large samples. In fact, the original theoretical development was motivated by achieving robustness in a neighbourhood of the normal distribution.