Basic Statistics for Libraries and for Analysis of Research Data

M S Sridhar
Indian Space Research Organization

BASIC STATISTICS FOR LIBRARIES AND FOR ANALYSIS OF RESEARCH DATA

M S Sridhar

Head, Library & Documentation

ISRO Satellite Centre

Bangalore 560017

E-mail: sridhar@isac.gov.in &

sridharmirle@yahoomail.com

Synopsis – Research Methodology

1. *Introduction to Research & Research methodology*
2. *Selection and formulation of research problem*
3. *Research design and plan*
4. *Experimental designs*
5. *Sampling and sampling strategy or plan*
6. *Measurement and scaling techniques*
7. *Data collection methods and techniques*
8. *Testing of hypotheses*
9. *Statistical techniques for processing & analysis of data*
10. *Analysis, interpretation and drawing inferences*
11. *Report writing*

SYSNOPSIS

2.0 Objectives

This unit introduces you to basics of statistics in general and procedures and techniques required for processing and analysing research data in particular. After studying this unit you will be able to (i) appreciate the importance of statistical techniques in research (ii) how to go about processing enormous data collected in the process of research (iii) understand elementary statistical methods and techniques to process and analyse data and (iv) get brief guidance about steps involved in processing and analysis of data. In particular, after studying this unit you should be able to:

i. Organize observations to aid understanding of what you have observed.
ii. Construct tallies from raw data.
iii. Construct simple frequency distributions from tallies.

iv. Able to employ simple coding to the data.
v. Use percentages to compare distributions.
vi. Translate frequency distributions into percentage distributions.
vii. Change simple (discrete data) frequency distributions into cumulative distributions.
viii. Identify grouped data distributions, translate simple distributions into grouped distributions; Construct of grouped data distributions.
ix. Construct frequency polygons, histograms and other charts for simple (discrete) and grouped data distributions.
x. Understand the concepts central tendency, dispersion and asymmetry.
xi. Compute values for the mode, mean. median, range, mean deviation, and standard deviation for simple (discrete) and grouped data and understand their use and limitations.
xii. Determine the midpoint of simple (discrete) and grouped data categories.
xiii. Compare two different distributions and know what differences might exist between them based on given values for central tendency and dispersion.
xiv. Describe a distribution in terms of skewedness and kurtosis.

2.1 Introduction

Statistics is both singular and plural. As plural it means numerical facts systematically collected and as singular it is the science of collecting, classifying and using statistics.

What statistics does ?

1. Enables to present facts on a precise definite form that helps in proper comprehension of what is stated. Exact facts are more convincing than vague statements.
2. Helps to condensing the mass of data into a few numerical measures, i.e., summarises data and presents a meaningful overall information about a mass of data.
3. Helps in finding relationship between different factors in testing the validity of assumed relationship.
4. Helps in predicting the changes in one factor due to the changes in another factor.
5. Helps in formulation of plans and policies which require the knowledge of further trends; Statistics plays vital role in decision making

2.1.1 Significance of Statistics in Research

Statistics, basically, helps to organise and summarise numerical information, to relate findings on a small sample to the large population, to test hypothesis, to explain variations, to build model and predict the result, to randomise the sample, etc. Science of statistics cannot be ignored by researcher. It is an essential tool for designing research, processing & analysing data and drawing inferences / conclusions. It is also called a double edged tool, as it can be abused and misused. It can be abused with poor data coupled with sophisticated statistical techniques to obtain unreliable results and it can be misused when honest, hard facts are combined with poor/inappropriate statistical techniques to create false impressions and conclusions. A number of funny examples like applying percentage for very small sample, using wrong average, playing with probabilities, inappropriate scale, origin and proportion between ordinate & abscissa, funny correlations like mind and beauty, one-dimensional figures, unmentioned base, etc., can be seen in literature.

Statistics for researcher can be broadly grouped into:

A. Descriptive statistics (& causal analysis) concerned with development of certain indices from the raw data and causal analysis for which the following aspects of statistics are required.

1. Uni-dimension analysis (mostly one variable)

(I) Central tendency - mean, median, mode, Geometric Mean (GM) & Harmonic Mean (HM)
(ii) Dispersion - variance, standard deviation , mean deviation & range
(iii) Asymmetry (Skeweness) & Kurtosis

(iv) Relationship - Pearson’s product moment correlation, Spearman’s rank order correlation and Yule's coefficient of association

(v) Others - One way Analysis of Variation (ANOVA), index numbers, time series analysis, simple correlation & regression

2. Bivariate analysis

   (I) Simple regression & correlation

   (ii) Association of attributes

   (iii) Two-way ANOVA

3. Multivariate analysis

   (I) Multiple regression & correlation / partial correlation

   (ii) Multiple discriminant analysis

   (iii) Multi-ANOVA

   (iv) Canonical analysis

   (v) Factor analysis, cluster analysis, etc.

(As far as this unit is concerned, you will be studying selected techniques listed under ‘Uni-dimensional analysis).

B. Inferential (sampling / statistical) Analysis (used in testing of hypothesis) is concerned with the process of generalization, particularly (a) estimation of parameter values (point and interval estimates) (b) testing of hypothesis (using parametric / standard tests and non-parametric / distribution-free tests) and (c) drawing inferences consisting of data processing, analysis, presentation (in table, chart or graph) & interpretation (is to expound the meaning) should lead to drawing inference. In other words, validation of hypotheses and realisation of objectives of research with respect to (a) relationship between variables (b) discovering a fact (c) establishing a general or universal law are involved in inferential analysis.

The concepts parametric and non-parametric tests just mentioned recur quite often and need to be understood correctly. **Parametric / standard tests** require measurements equivalent to at least an interval scale and assume certain properties of parent population like i) observations are from a normal population ii) large sample iii) population parameters like mean, variance, etc. **Non-parametric / distribution-free tests** do not depend on any assumptions about properties/ parameters of the parent population. Most non-parametric tests assume nominal or ordinal data. Non-parametric tests require more observations than parametric tests to achieve the same size of Type I and Type II errors (as explained in the adjacent table for given Null hypothesis H_0).

This unit being elementary introduction, only basic concepts are presented with few techniques for a beginner. These basic concepts are also useful in other units concerning experimental design, sampling, measurement and scaling techniques, testing of hypothesis as well as analysis, interpretation and drawing inferences.

	Decision	**Decision**
	Accept	**Reject**
H_0 (true)	**Correct decision**	**Type I error**
H_0 (false)	**Type II error**	**Correct decision**

2.1.2 Quantitative and Qualitative Data:

A quantitative or numerical data is an expression of a property or quality in numerical terms. In other words, it is data measured and expressed in quantity. It enables (i) precise measurement (ii) knowing trends or changes over time, and (iii) comparison of trends or individual units. On the other hand, qualitative (or categorical) data involves quality or kind with subjectivity. When feel & flavour of the situation become important, researchers resort to qualitative data (some times called attributes). *Qualitative data describe attributes of a single or a group of persons that is important to record as accurately as possible even though they cannot be measured in quantitative terms*. More time & efforts are needed to collect & process qualitative data. Such data are not amenable for statistical rules & manipulations. As said above, scaling techniques help converting some qualitative data into quantitative data. Usual data reduction, synthesis and plotting trends are required but differ substantially in case of qualitative data. Above all, extrapolation of finding is difficult in case of qualitative data. It also calls for sensitive interpretation & creative presentation. For example, quotation from interviews, open remarks in questionnaires, case histories bringing evidence, content analysis of verbatim material, etc. provide abundant qualitative data requiring sensitive use.

Identifying & coding recurring answers to open ended questions in questionnaires help categorising key concepts & behaviour for limited quantitative processing. Such counting and cross analysis require pattern discerning skills. Even unstructured depth interviews can be coded to summarise key concepts & present in the form of master charts.

The process and analysis of qualitative data involves:

(i) Initial formalisation with issues arising (i.e., build themes & issues)
(ii) Systematically describing the contents (compiling a list of key themes)
(iii) Indexing the data (noting reflections for patterns, links, etc.) in descriptions, interpreting in relation to objectives and checking the interpretation
(iv) Charting the data themes
(v) Refining the charted material
(vi) Describing & discussing the emerging story

Qualitative coding is concerned with classifying data which are not originally created for research purpose and having very little order. As such, very less of statistical processing and analysis are applicable to qualitative data. We shall be discussing in this unit application of statistical techniques to processing and analysis of quantitative data.

2.1.3 Basic Concepts

2.1.3.1 Data types

Data is the central thread of any research. While formulating a research problem, choosing type of research and designing research (i.e., preparing a blue print for research), researcher has to have a clear and unambiguous understanding of what kind of data he is intending to collect. Designing a suitable measuring scale appropriate to data collection tools and later choice of statistical techniques for processing presuppose the knowledge of nature of data to be gathered by the researcher.

You have already noticed that, data are of broadly two kinds, namely, qualitative and quantitative. Qualitative data can also be transformed into quantitative data to a limited extent by choosing or designing measurement scale though their accuracy cannot be compared to quantitative data. But it is often better to measure and analyse attributes with some degree of quantification than leaving completely as a qualitative and non-quantifiable.

Based on their mathematical properties, data are divided into four groups. An understanding of properties of these four fundamental groups are essential for both study of scales of measurement and also for processing and analysis of data. They are Nominal, Ordinal, Interval and Ratio. One can remember the four with the acronym **NOIR.** They are ordered with their increasing accuracy, powerfulness of measurement, preciseness and wide application of statistical techniques.

NOMINAL (meaning name and count) data are alphabetic or numerical in name only. Even if the numbers are used they are only symbols used to label and number has no arithmetic significance. For example, PC112 has no arithmetic significance. There use is restricted to keep track of people, objects and events. For example, subject of specialisation of users like Physics (P), Chemistry (C), Mathematics (M), Biology (B), etc can be represented by alphabets. They can also be named as 1,2,3 & 4 and yet they are nominal. No statistical manipulation is possible with nominal data. The order is of no consequence in this kind of data. As such nominal data and scale are least powerful in measurement with no arithmetic origin, order or distance relationship. Hence nominal data is of restricted or limited use. Chi-square is the most common test used for nominal data and correlations are possible through contingency coefficients.

ORDINAL (meaning rank or order) data allows for setting up inequalities and nothing much. Hence ordinal scales place events in order. As intervals of the ordinal scales are not equal adjacent ranks need not be equal in their differences. Ordinal data has no absolute value (only relative position in the inequality) and hence more precise comparisons are not possible. For example, educational qualification of a population can be arranged as undergraduate (U), Graduate (G), Postgraduate (P) and Doctorate (D) with their respective alphabets or as 1, 2 3 and 4 and they allow for setting inequalities like U < G < P < D but the differences between any two of them (say U and G or G and P) cannot be said to be same. Median is appropriate measure of central tendency and percentile or quartile is used for measuring dispersion of ordinal data. As ordinal data allows for setting up inequalities, rank order correlations can be worked out. However, only non-parametric (or distribution-free) tests are possible. Such ranked ordinal data is used extensively in qualitative research.

INTERVAL (OR SCORE/ MARK) data further allows for forming differences in addition to setting up inequalities. Interval scale adjusts intervals in such a way that a rule can be established as a basis for making the units equal. As no absolute zero or unique origin exists, the scale fails to measure the complete absence of a trait or characteristic. Measure for temperature is a good example for interval data (10^o, 20^o etc). Interval scale is more powerful than ordinal scale due to equality of intervals. Mean is appropriate measure of central tendency and standard deviation (SD) is most widely used measure of dispersion for interval data. Product moment correlations can be worked out with 't' test & 'F' test for significance in case of interval data.

RATIO data allows for forming quotients in addition to setting up inequalities and forming differences. In other words, all mathematical operations (manipulations with real numbers) are possible on ratio data. Geometric mean (GM) and Harmon mean (HM) can also be used with ratio data. The coefficient of variation can be worked out for this kind of data. The ratio scale has an absolute or true zero and represents the actual amounts of variables. It is the most precise scale and it allows for application of all statistical techniques.

2.1.3.2 Variables, Experiment and Hypothesis

The concepts variable, experiment and hypothesis will be dealt in detail in experimental design and hypothesis testing parts of research methods. However, it is necessary to know the terminology in this unit.

***VARIABLE** is a* concept which can take on different quantitative values. That means the absence of non-quantifiable attributes is important.
DEPENDENT VARIABLE depends upon or a consequence of other variable.
INDEPENDENT VARIABLE is an antecedent to dependent variable.
EXTRANEOUS VARIABLE: Independent variable not related to the purpose but may affect the dependent variable & its effect on dependent variable is called 'experimental error'.
CONTROL (of experiment) is to minimise the effects of extraneous independent variable.
***CONFOUNDED RELATIONSHIP** is* relationship between dependent & independent variable. Confounded relationship means that the experiment is not free from the influence of extraneous variable.
RESEARCH HYPOTHESIS is a predictive statement relating independent variable to dependent variable . At least one dependent & one independent variable should be there for a hypothesis to exist.
***EXPERIMENTAL HYPOTHESIS TESTING** involves* manipulating independent variable in the experiment.
***NON-EXPERIMENTAL HYPOTHESIS TEST** is where* independent variable is not manipulated in the experiment.
EXPERIMENTAL GROUP is a group exposed to some novel or special condition, stimuli or treatment.
***CONTROL GROUP** is* group exposed to usual conditions.
TREATMENTS (STIMULI) is the different conditions under which experimental & control groups are subjected to.
***EXPERIMENT** is the p*rocess of examining the truth of a statement or hypothesis (absolute or comparative experiment).
***EXPERIMENTAL UNITS** are* pre-determined plots or blocks, where different treatments are applied**.**

2.1.3.3 Normal Distribution

One of the most fundamental concepts in the statistical analysis is normal distribution. Hence you need to know basics of normal distribution. Normal distribution is a special continuous distribution of items in a series which is perfectly symmetrical. If a curve is drawn form normal frequency distribution it will be bell-shaped symmetric curve (see figure). Great many techniques used in applied statistics are based on this distribution. Many populations encountered in the course of research in many fields seems to have a normal distribution to a good degree of approximation. In other words, in research nearly normal distributions are encountered quite frequently. The normal curve being a perfect symmetrical curve the mean, median and the mode of the distribution are one and the same (X = M = Z). The sampling distributions based on a parent normal distributions are manageable analytically.

A researcher or experimenter after understanding what type or kind of data has been collected must know, at least approximately, the general form of the distribution function which his data follow. If it follows normal distribution, he can straight away use many established statistical methods and techniques. If the distribution is not normal, he must explore the possibilities of transforming his data so that the transformed observations follow a normal distribution. If a researcher does not know the form of his population distribution and the kind of distribution his data follows, then he must use other more general but usually less powerful methods of analysis called non-parametric methods.

Definition: The random variable x is said to be normally distributed if density function is given by

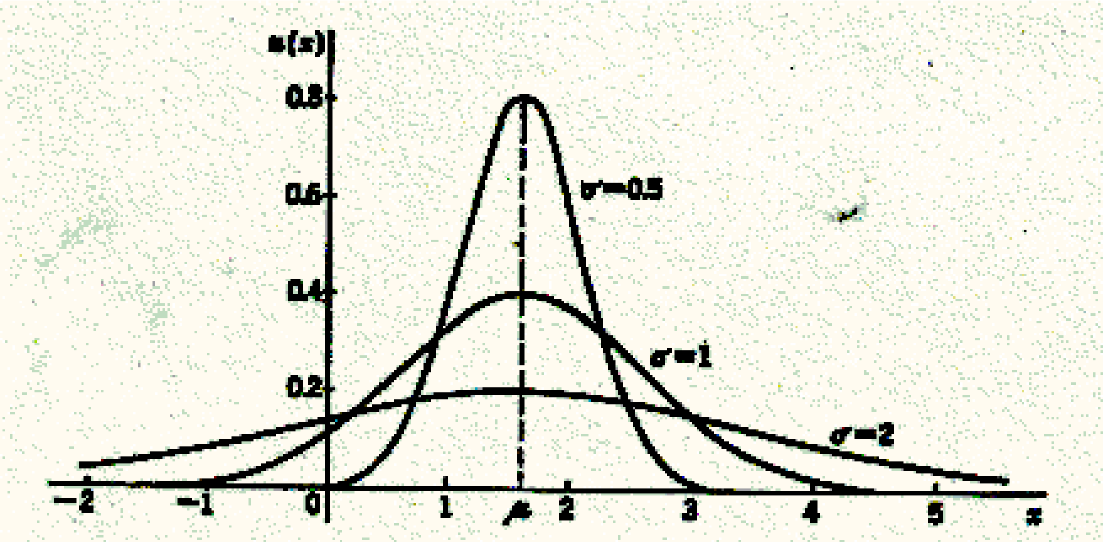
$$n(x) = \frac{1}{\sqrt{\Pi}\sigma} e^{-(x-\mu)2} / 2\sigma^2$$

Where $\int_{-\infty}^{\infty} n(x)\,dx = 1$ and $-\infty < x < \infty$

(Since n (x) is given to be a density function , it implied that $n(x)\,dx = 1$)

When the function is plotted for several values of σ (standard deviation) , a bell shaped curve as shown below can be seen. Changing μ (mean) merely shifts the curves to the right or left without changing their shapes. The function given actually represents a two-parameter family of distributions, the parameters being μ and σ^2 (mean and variance).

Figure:



2.2 Processing & Analysis of (Quantitative) Data

Quantitative data (i.e., numbers representing counts, ordering or measurements) can be described, summarised (data reduction), aggregated, compared and manipulated arithmetically & statistically. As described earlier, types or levels of measurement (ie., nominal, ordinal, interval & ratio) determine the kind of statistical techniques to be used. Further, use of computer is necessary in processing and analysis of data and it varies from simple Excel (MS Office) to more extensive SSPS package.

The following steps are often identified for processing and analysis of data :

1. Data reduction, aggregation or compression (Reducing large batches & data sets to numerical summaries, tabular & graphical form in order to enable to ask questions about observed patterns).
2. Data presentation

3. Graphical presentation
4. Exploratory data analysis
5. Looking for relationships & trends

For the purpose of this unit, let us briefly see the processing and analysis of quantitative data in two parts, namely, processing and analysis.

2.2.1 Processing (Aggregation & Compression) of Data

The processing of data involves editing, coding, classification, tabulation and graphical presentation. Let us see each one of them with examples.

2.2.1.1 Editing

Data collected in research require certain amount of editing for making it unambiguous and clear as well as for maintaining consistency and accuracy. Editing can be done either in the field while collecting data or centrally after bringing it to a place and they are respectively called field editing and central editing.

2.2.1.2 Coding

Assigning data measured and collected to a limited number of mutually exclusive but exhaustive categories or classes is coding. It is required as a basic step in processing.

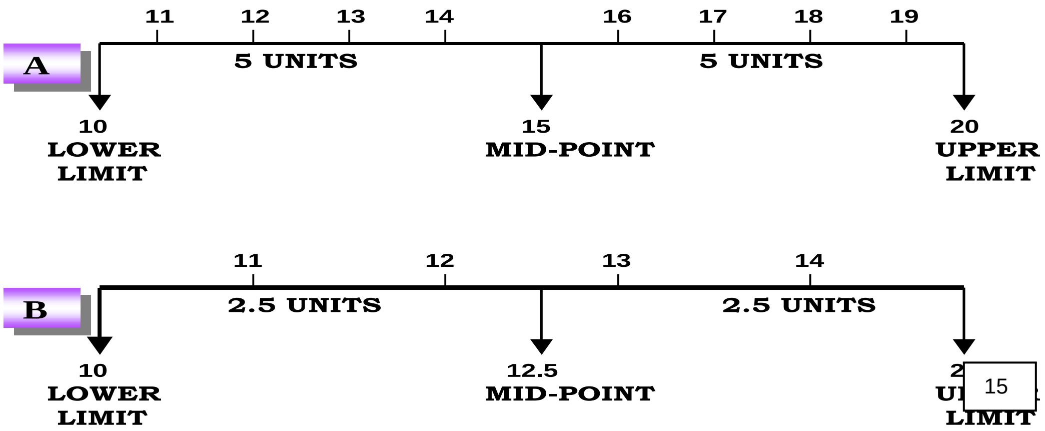
2.2.1.3 Classification

To create such mutually exclusive but exhaustive classes, it is necessary to do classification of data by arranging data in groups or classes on the basis of common characteristics. This can be done by using attributes (statistics of attributes) for qualitative data and by using class intervals, class limits, magnitude & frequencies for quantitative data (statistics of variables). The number of classes normally should be between 5 and 15. The mathematical way to work out size of class is given by the formula $i = R / 1+3.3 \log N$, where i is size of class interval, R is Range, N is number of items to be grouped.

The class limits are chosen in such a way that midpoint is close to average.

Class intervals: are the various intervals of the variable chosen for classifying data. The class intervals could be exclusive or inclusive (i.e., type of class intervals). The procedure to determine mid point of even and odd class-intervals are shown in the diagram below:

A - Even class-interval & its mid point ; B - Odd class-interval & its mid point



Frequency of an observation: The number of times a certain observation occur is the frequency of observation. Determination of frequency of each class is done by marking a vertical line in tally sheet for each observation and a diagonal across for the fifth score so that frequencies are easily consolidated by looking at the group of five as shown in the table 9.1 .

Frequency table: A table which gives the class intervals and the frequencies associated with them in a summarized way. Tally (tabular) sheets or charts showing frequency distribution are given below:

Table 2.1 (Quantitative data) Frequency distribution of citations in technical reports

<table>
<tr><th><i>No. of citations</i></th><th><i>Tally</i></th><th><i>Frequency</i>
<i>(No. of technical reports)</i></th></tr>
<tr><td>0</td><td>||</td><td>2</td></tr>
<tr><td>1</td><td>||||</td><td>4</td></tr>
<tr><td>2</td><td>||||</td><td>5</td></tr>
<tr><td>3</td><td>||||</td><td>4</td></tr>
<tr><td>4</td><td>|||| ||</td><td>7</td></tr>
<tr><td>5</td><td>|||| |||</td><td>8</td></tr>
<tr><td></td><td></td><td><b>Total 30</b></td></tr>
</table>

Table 2.2 (Qualitative data) Frequency distribution of qualification (educational level) of users

Qualification	***Tally***	***Frequency (No. of users)***
Undergraduates		6
Graduates		9
Postgraduates		7
Doctorates		3
		Total 25

2.2.1.4 Tabulation (Tabular Presentation)

Summarising and displaying data in a concise / compact and logical order for further analysis is the purpose of tabulation. It is a statistical representation summarising and comparing frequencies, determining bases for and computing percentages, etc. and presenting simple or complex table based on accepted general principles. It is important to remember while tabulating responses to questionnaire that problems concerning the responses like 'Don't know' and not answered responses, computation of percentages, etc. have to be handled carefully. Some examples of frequency tables are given below to familiarize the task.

Table 2.3 Frequency distribution of age of 66 users who used a branch public library during an hour (Grouped/ interval data of single variable)

(Note that the raw data of age of individual users is already grouped here).

Age in years (Groups/Classes)	*Tally*	*Frequency (No. of users)*
< 11		11
11 – 20		14
21 – 30		16
31 - 40		12
41 - 50		6
51 - 60		3
> 60		**4**
		Total 66

Table 2.4: No. of books acquired by a library over last six years

Year	***No. of Books acquired***
(Qualitative)	***(Quantitative)***
2000	772
2001	910
2002	873
2003	747
2004	832
2005	891
Total	5025

Table 2.5: The daily visits of users to a library during a week are recorded and summarised

Day	***Number of users***
(Qualitative)	***(Quantitative)***
Monday	391
Tuesday	247
Wednesday	219
Thursday	278
Friday	362
Saturday	96
Total	1593

Table 2.6: The frequency distribution of number of authors per paper of 224 sample papers

No. of Authors	***No. of papers***
1	43
2	51
3	53
4	30
5	19
6	15
7	6
8	4
9	2
10	1
Total	224

Table 2.7: Total books (B), journals (J) and reports (R) issued out from a library counter in one hour are recorded as below:

B	B	B	J	B	B
B	B	J	B	B	B
B	B	B	B	B	B
B	B	B	B	B	J
B	R	B	B	B	J

A frequency table can be worked out for above data as shown below:

Document	***Tally***	***Frequency***	***Relative***	***Cumulative***	***Cumulative***
Type		***(Number)***	***frequency***	***frequency***	***relative frequency***
Books		20	0.8	20	0.8
Journal		4	0.16	24	0.96
Reports		1	0.04	25	1.0
	Total	25	1.0		

Note: If the proportion of each type of document (category) are of interest rather than actual numbers, the same can be expressed in percentages or as proportions as shown below:

Proportions of books, journals and reports issued from a library in one hour is 20:4:1 **OR**

Type of document	Proportion of each type of document (%)
Books	80
Journal	16
Reports	4
Total	100

Table 2.8: Given below is a summarized table of the relevant records retrieved from a database in response to six queries

Search No.	***Total Documents Retrieved***	***Relevant Documents Retrieved***	***% of relevant records Retrieved***
1	79	21	26.6
2	18	10	55.6
3	20	11	55.0
4	123	48	39.0
5	16	8	50.0
6	109	48	44.0
Total	375	146	

Note: Percentage of relevant records retrieved for each query gives better picture about which query is more efficient than observing just frequencies.

Table 2.9:: Frequency distribution of borrowed use of books of a library over four years

No. Times borrowed *(Quantitative)*	**No. of Books *(Quantitative)***	**Percentage**	**Cumulative Percentage**
0	19887	57.12	57.12
1	4477	12.56	69.68
2	4047	11.93	81.61
3	1328	3.81	85.42
4	897	2.57	87.99
5	726	2.02	90.1
6	557	1.58	91.68
7	447	1.28	92.96
8	348	1	93.96
9	286	0.92	94.78
10	290	0.84	95.62
>10	1524	4.38	100

Table 2.10: The raw data of self-citations in a sample of 10 technical reports are given below:

5	0	1	4	0
3	8	2	3	0
4	2	1	0	7
3	1	2	6	0
2	2	5	7	2

Frequency distribution of self-citations of technical reports:

No. of self-citations	***Frequency (No. of reports)***		***Less than (or equal) cumulative frequency***	***More than (or equal) cumulative frequency***
	No.	***%***	***%***	***%***
0	5	20	20	100
1	3	12	32	80
2	6	24	56	68
3	3	12	68	44
4	2	8	76	32
5	2	8	84	24
6	1	4	88	16
7	2	8	96	12
8	1	4	100	4
TOTAL	25			

Table 2.11:: (Qualitative Data) Responses in the form of True (T) or False (F) to a questionnaire (opinionnaire) is tabulated and given along with qualitative raw data

True	17
False	8
No response	5
Total	30

T	T	T	F	F
F	T	T	T	T
T	T	F	F	T
F	T	F	T	T
T	T	T	T	F

Grouped or interval data: So far, except Table 9.3 (which gives distribution of grouped/ interval data),

Table 2.12 : (Grouped or interval data) Raw data of prices (in Rs.) of a set of 50 popular science books in Kannada

30	80	100	12	40
50	60	40	30	45
40	30	70	43	40
25	50	10	30	35
18	35	60	35	25
27	25	25	30	30
35	35	14	32	35
25	30	40	15	30
20	16	13	30	60
20	65	60	40	10

you have seen data which is discrete and number of cases/ items are limited . Many a times continuous data (like height of people, which varies minutely from person to person, say 5' 1", 5' 2", etc.) has to be collected in group or interval like between 5' and 5'5", for any meaningful analysis. In research, often, we assemble large quantity of discrete data which is required to be compressed and reduced for meaningful observation, analysis and inferences. Table 9.11 present 50 observations and if we create a frequency table of these discrete data it will have 22 line table as there are 22 different values ranging fro Rs.10/- to Rs.100/-. Such large tables are undesirable as they not only take more time but also the resulting frequency table is less appealing. In such situation we transform discrete data into grouped or interval data by creating manageable number of classes or groups. Such data compression is inevitable and worth despite some loss of accuracy (or data).

Frequency distribution of above discrete data

Price in Rs.	No. of books
10	1
12	1
13	1
14	1
15	1
16	1
18	1
20	2
25	5
27	1
30	9
32	1
35	6
40	6
43	1
45	1
50	2
60	4
65	1
70	1
80	1
100	1

Mean = Rs. 35.9
Median = Rs. 33.5
Mode = Rs. 30

From the above raw data, a frequency table with class intervals can be worked out easily given below.

Frequency distribution of grouped or interval data of price (in Rs.) of popular science books in Kannada (Table 9.12) :

Price (in Rs.) (class)	***Frequency (f) (No. of books)***
1 -- 10	2
11 -- 20	8
21 -- 30	15
31 -- 40	13
41 -- 50	4
51 -- 60	4
61 -- 70	2
71 -- 80	1
81 -- 90	0
91 -- 100	1
Total	50

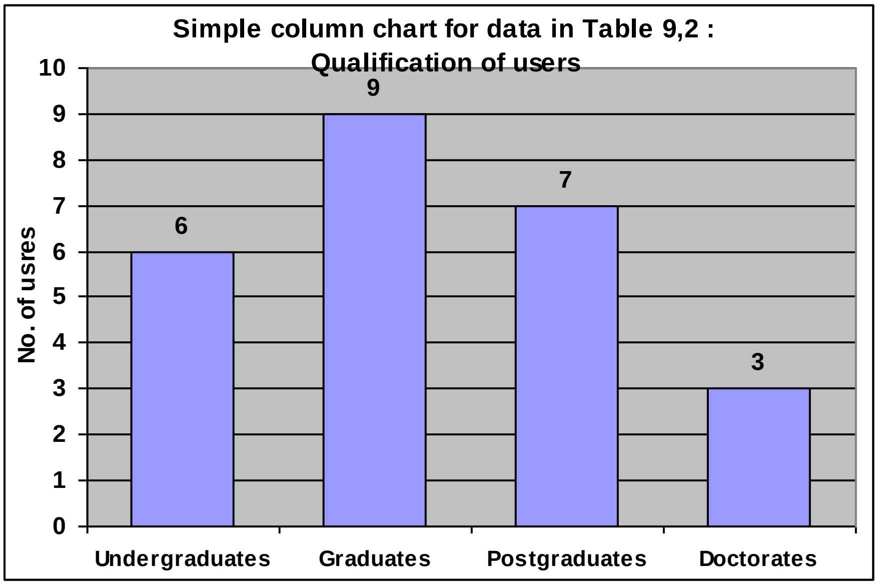
Home work: 1. Work out a frequency table with less than cumulative and more than cumulative frequencies for the raw data of number of words per line in a book given below :

12	10	12	09	11	10	13	13
07	11	10	10	09	10	12	11
01	10	13	10	15	13	11	12
08	13	11	10	08	12	13	11
09	11	14	12	07	12	11	10

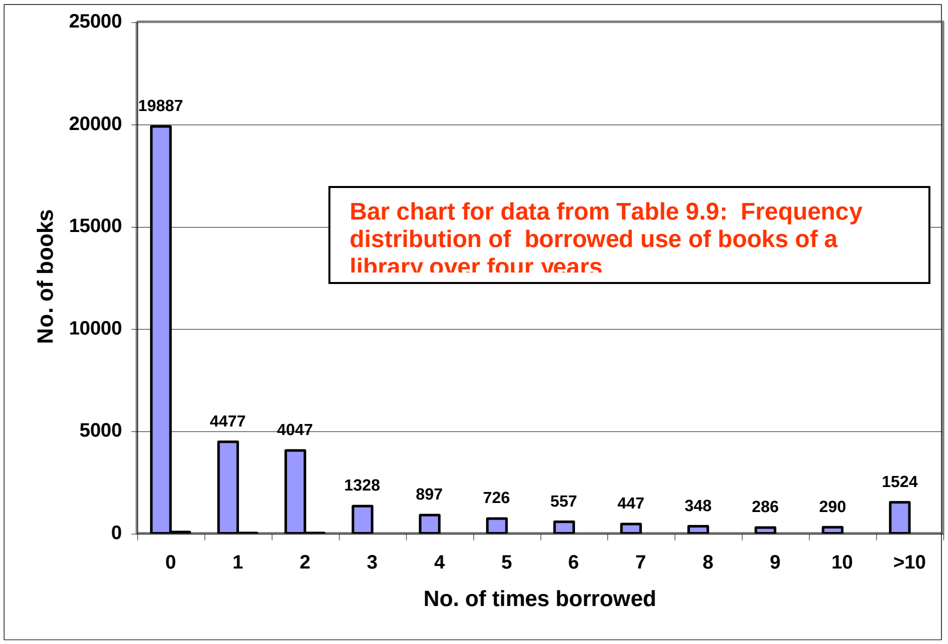
2.2.1.5 Graphical / Diagrammatic Presentation

In this section, instead of explaining how to prepare and how they look like actual sample charts are constructed and presented using data from previously presented tables.

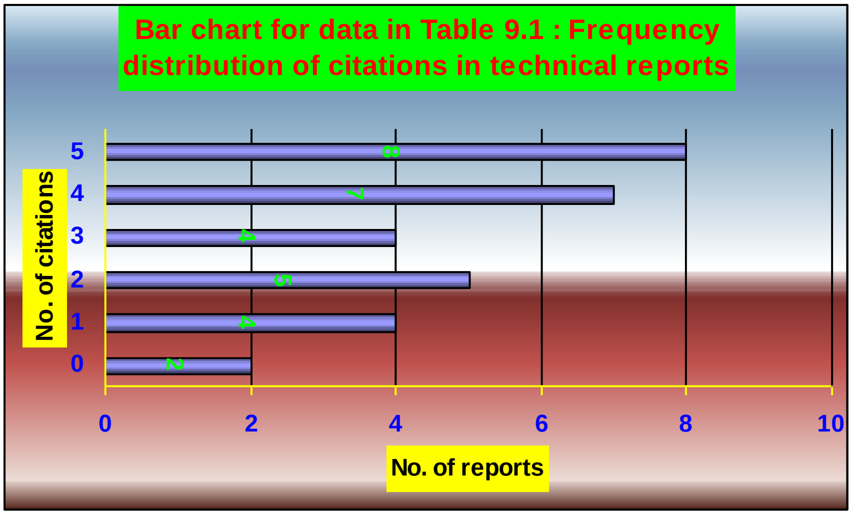
2.2.1.5 .1 Column Chart



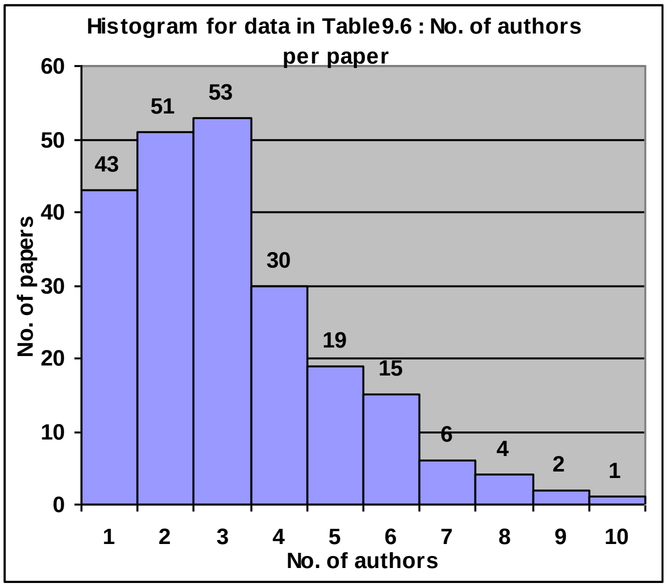
Note: More than one series of data can also be plotted to show pairs of columns in the same diagram to enable comparison.



2.2.1.5.2 Bar Chart



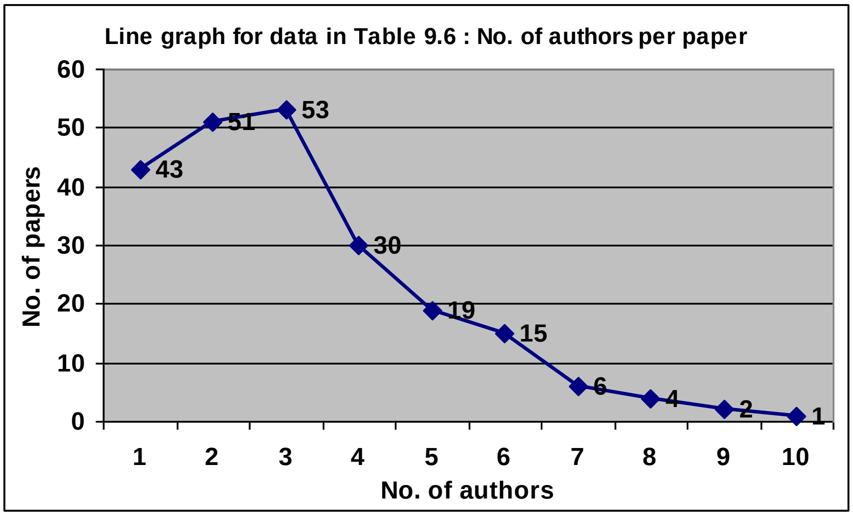
2.2.1.5.3 Histogram



Note; 1. Frequency is represented as area in the histogram

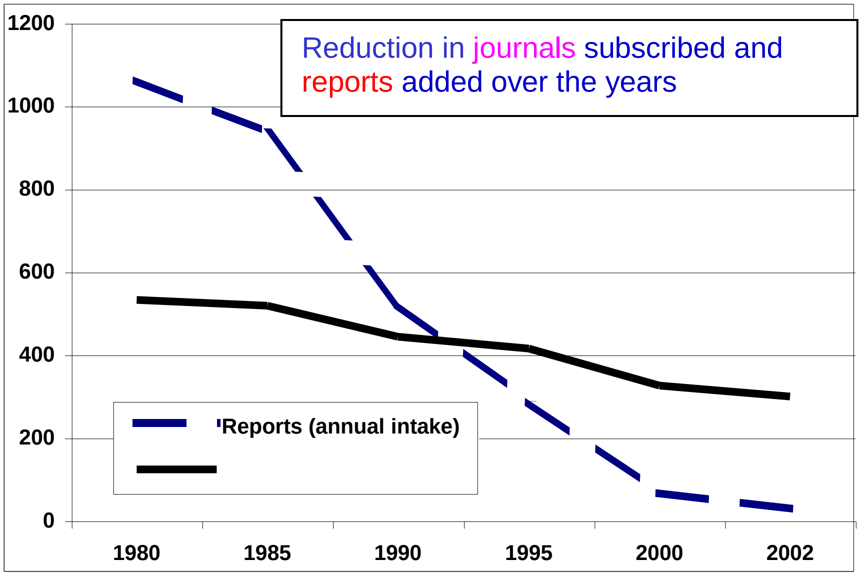
2. In case of grouped or interval data, frequency is represented by area only when equal class intervals are used.

2.2.1.5.4 Simple Line Graph/ Frequency Graph

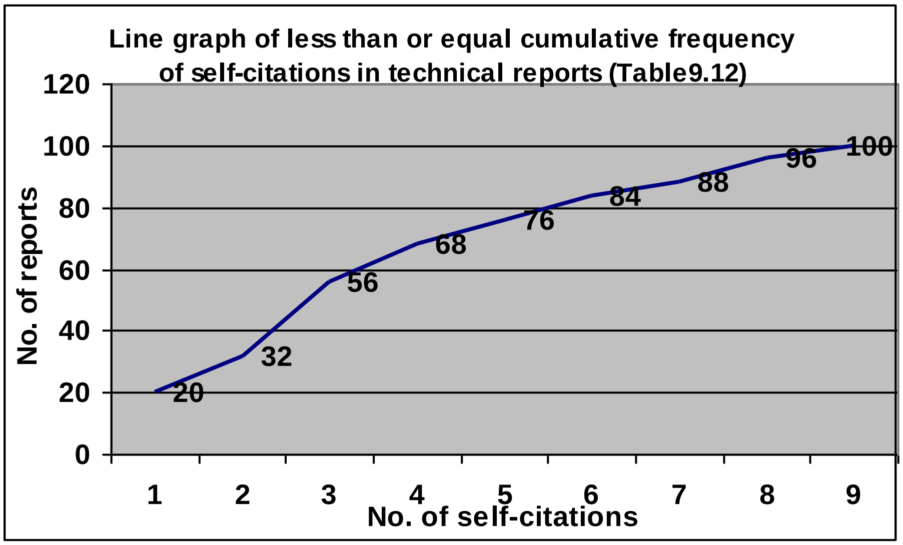


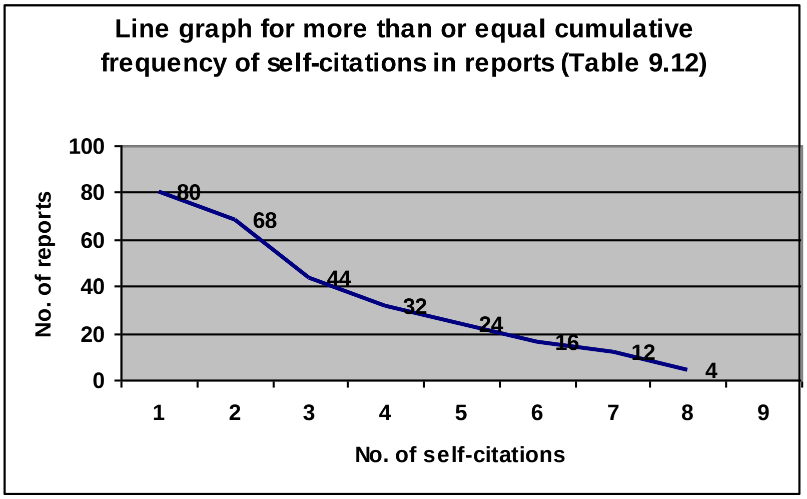
Note: 1. If only dots are shown without joining line in the graph it is called dot graph.

2. A "Frequency Polygon" will have more than one series of data plotted and hence will have more than one line graph in the same diagram to enable comparison. Chart below is an example of how reduction in number of journals subscribed and number of reports added over the years by a library can be compared while depicting the individual trend of journals subscribed and reports added.

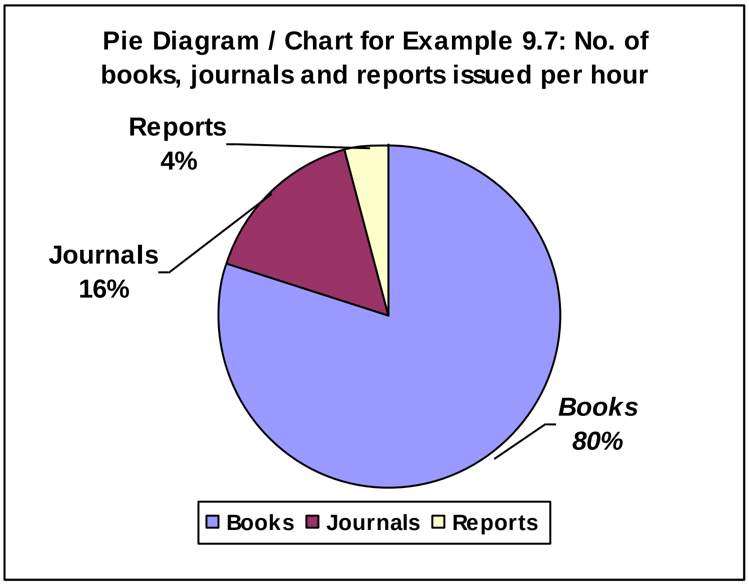


2.2.1.5.5 Cumulative Frequency Graph





2.2.1.5.6 Pie Diagram/ Chart



Note: The other slightly varied types of graphs and charts like component bar chart, 100% component bar charts, grouped column chart, comparative 100% columnar chart, charts with figures and symbols, semi-logarithmic graph, surface graph, 100% surface graph, etc may be observed in books, magazines and newspapers.

Home work: 4. Prepare a less than or equal cumulative line graph for frequency table developed for Exercise 3 (use Excel of MS Office software).

2.2.2 Analysis of Data

The next part of processing and analysis of data involves exploring analysis, computation of certain indices or measures, searching for patterns of relationships, trends, estimating values of unknown parameters & testing of hypothesis for inferences, etc. Analysis of data itself can be grouped into two:

(i) Descriptive analysis which largely deals with the study of distributions of one variable (**univariate**). Distributions of two variables are called **bivariate** and those with more than two variables are called **multivariate.**

(ii) Inferential or statistical analysis is mainly concerned with bivariate analysis, correlational and casual analyses and multivariate analysis.

Correlation & causal analyses: An analysis of joint variation of two or more variables is correlation analysis. Studying how one or more variables affect another variable is causal analysis.

Regression analysis deals with explaining functional relation existing between two or more variables.

As this unit is concerned with basics of statistical processing and analysis, only selected univariate measures are presented. However a selected list of bivariate and multivariate measures are listed at the end.

Univariate Measures: As explained above, univariate analysis is concerned with single variable and we shall see important univariate measures like central tendency, dispersion measures and measure of asymetry.

2.2.2.1 Central Tendency Measures (or Measures of Location)

2.2.2.2.1 MEAN

By dividing the total of items by total number of items we get mean. It tells the point about which items have a tendency to cluster. It is the most representative figure for the entire mass of data. It is also called statistical / arithmetic average and it is unduly affected by extreme items. It is the measure of central tendency most used for numerical variables.

Example (**Discrete data**):

4 6 7 8 9 10 11 11 11 12 13

$$M = 4+6+7+8+9+10+11+11+11+12+13 / 11 = 102/11 = 9.27$$

Note :1. Often fractional figures of mean look meaningless in real life situation (like 2.5 persons) but the same indicate the central tendency more accurately.

2. In order to reduce the tedious computation involved in calculating mean for large values of large number of items using every datum, wherever appropriate weighted mean and standardized mean can be used.

3. Choosing an appropriate measure of location (cental tendency) depends on the nature of distribution of data like unimodal or bimodal, skewness, etc.

For grouped or interval data:

$\bar{X} = \sum f_i \underline{x}_i \,/\, n$ where $n = \sum f_i$

$= f_1X_1 + f_2 X_2 + \ldots.+ f_n X_n \,/\, f_1 + f_2 + \ldots.+ f_n$

Formula for **Weighted Mean:**

$$\bar{X}_W = \sum W_x X_i \,/\, \sum W_i$$

Formula for **Mean Of Combined Sample:**

$$\bar{X} = n\bar{X} + m\bar{Y} \,/\, n + m$$

Formula for **Moving Average (**Shortcut or Assumed Average Method):

$$\bar{X} = f_i\,(X_i - A) \,/\, n \quad : \text{ where } n = \sum f_i$$

NOTE: **Step deviation method** takes common factor out to enable simple working and uses the formula $\bar{X} = g + [\sum f\,d \,/\, n]$ (i)

Calculation of the **mean** ($\bar{X}$) from a **frequency distribution** of grouped or interval data of price (in Rs.) of popular science books in Kannada (Table 9.12) using **Step deviation method** is shown below:

Price (in Rs.) (class)	***Frequency (f) (No. of books)***	***Cumulative less than or equal frequency (cf)***	***Distance of class from the assumed average class (d)***	***fd***	***d^2***	***fd^2***
1 -- 10	2	2	-4	-8	16	32
11 -- 20	8	10	-3	-24	9	72
21 -- 30	15	25	-2	-30	4	60
31 -- 40	13	38	-1	-13	1	13
41 -- 50	***4***	***42***	***0***	***0***	***0***	***0***
51 -- 60	4	46	1	4	1	0
61 -- 70	2	48	2	8	4	8
71 -- 80	1	49	3	3	9	9
81 -- 90	0	50	4	4	16	0
Total	50			-56		194

g = 46

$\Sigma f\text{d} = -56$

$n = 50$

$i = 10$

$\bar{X} = g + [\sum f\, d\, / \,n]\ (i) = 46 + [-56 / 50]\ (10) = 34.6$

Note: Compare answer with mean calculated as discrete data in Table 9.12

2.2.2.1.2 MEDIAN

The median in the layman language is divider like the 'divider' on the road that divides the road into two halves. The median of a set of observations is a value that divides the set of observations into two halves so that one half of observations are less than or equal to the median value and the other half are greater than or equal to the median value. In other words, Median is middle item of a series when arranged in ascending or descending order of magnitude. Thus, in order to find median, observations have to be arranged in order.

M = Value of [N+1] / 2 th item

Example (**Discrete data**) :

Data :	11	7	13	4	11	9	6	11	10	12	8
Ascending order :	4	6	7	8	9	10	11	11	11	12	13
Serially numbered frequencies :	1	2	3	4	5	6	7	8	9	10	11

M = Value of [N+1] / 2 th item = [11 + 1] / 2 = 6th item i.e., 10

Note : 1. For discrete data, mean and median do not change if all the measurements are multiplied by the same positive number and the result divided later by the same constant.

2. As a positional average, median does not involve values of all items and it is more useful in qualitative phenomenon.

For frequency distribution of grouped or interval data

$M = L + W/F\ (i)$

Where, $W = [n/2] - C_f$ (No. of observations to be added to the cumulative total in the previous class in order to reach the middle observation in the array)

L = Lower limit of the median class (the array in which middle observation lies)

C_f = Cumulative frequency of the class preceding the median class

i = Width of the class interval of the median class (the class in which the middle observation of the array lies)

F = Frequency distribution of the median class

Calculation of the **median** (M) from a **frequency distribution** of grouped or interval data of price (in Rs.) of popular science books in Kannada (Table 9.12) is shown below:

Price (in Rs.) (class)	***Frequency (f) (No. of books)***	***Cumulative less than or equal frequency***
1 -- 10	2	2
11 -- 20	8	10
21 -- 30	***15***	***25***
31 -- 40	13	38
41 -- 50	3	41
51 -- 60	3	44
61 -- 70	2	46
71 -- 80	1	47
81 -- 90	0	47
91 -- 100	1	48
> 100	2	50
Total	**50**	

$L = 21$

$C_f = 10$

$I = 10$

$F = 15$

$W = [n/2] – C_f = [50/2] – 10 = 15$

$M = L + W/F (i) = 21 + 15/15 (10) = 31$

Note: Compare answer with median calculated as discrete data in Table 9.12

2.2.2.1.3 MODE

Mode is the most commonly or frequently occurring value in a series. In other words, the mode of a categorical or a discrete numerical variable is that value of the variable which occurs maximum number of times (i.e., with the greatest frequency). The mode or modal category in Table 9.2 is Graduates with greatest frequency of 9. The mode does not necessarily describe the 'most' (for example, more than 50 %) of the cases. Like median, mode is also a positional average and is not affected by values of extreme items. Hence mode is useful in eliminating the effect of extreme variations and to study popular (highest occurring) case. The mode is usually a good indicator of the centre of the data only if there is one dominating frequency. However, it does not give relative importance and not amenable for algebraic treatment (like median).

Note: 1. Median and mode could also be used in qualitative data.
2. Median lies between mean & mode.
3. For normal distribution, mean, median and mode are equal (one and the same).

Example (**Discrete data**) : 4 6 7 8 9 10 11 11 11 12 13
^

As it occurs thrice, 11 is the mode of this discrete data.

For frequency distribution of grouped or interval data

For frequency distribution with grouped (or interval) quantitative data , the model class is the class interval with the highest frequency. This is more useful when we measure a continuous variable which results in every observed value having different frequency. Modal class in Table

9.3 is age group 21-30. Please note that since the notion of the location or central tendency requires order mode is not meaningful for nominal data.

$$Z = L + \frac{\Delta_2}{\Delta_2 + \Delta_1} \; (i) \qquad \text{OR} \qquad L + \frac{f_2}{f_2 + f_1} \; (i)$$

Where, L = Lower limit of the modal class

Δ_1 = Difference in frequency between the modal class and the preceding class

Δ_2 = Difference in frequency between the modal class and the succeeding class

i = Width of the class interval of the modal class

f_1 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

Calculation of the **mode** (Z) from a **frequency distribution** of grouped or interval data of price (in Rs.) of popular science books in Kannada (Table 9.12) is shown below:

Price (in Rs.) *(class)*	*Frequency (f)* *(No. of books)*	*Cumulative less than or equal* *frequency (cf)*
1 -- 10	2	2
11 -- 20	8	10
21 -- 30	15	25
31 -- 40	13	38
41 -- 50	**4**	**42**
51 -- 60	*4*	*46*
61 -- 70	2	48
71 -- 80	1	49
81 -- 90	0	50
Total	50	

$L = 41$

$i = 10$

$f_1 = 13$

$f_2 = 4$

$$Z = L + [f_1 / f_1 + f_2] \; (i) \qquad \text{OR} \qquad L + [\Delta_2 / \Delta_1 + \Delta_2] \; (i)$$

$$Z = 41 + [13 / 13 + 4] \, (10) = 48.65$$

The value 48.65 lies in the class 41-50 and hence the modal class is 41-50 in the grouped data.

Note: Compare answer with mode calculated as discrete data in Table 9.12

2.2.2.2 Dispersion Measures

Dispersion is the scatter of the values of items in the series around the true value of average. Some of the important measures of dispersion are range, mean deviation, standard deviation and quartiles.

2.2.2.2.1 Range

Range is the difference between the values of the extreme items of a series, i.e., difference between the smallest and largest observations. Range gives an idea of the variability very quickly . It is simplest and most crude measure of dispersion. It is greatly affected by the two extreme values and fluctuations of sampling. The range may increase with the size of the set of observations though it can decrease, while ideally a measure of variability should be roughly independent of the number of measurements.

Example: 4 6 7 8 9 10 11 11 11 12 13

Range = 13 - 4 = 9

2.2.2.2 Mean Deviation (MD)

Median is the average of difference of the values of items from some average of the series. In other words, the average of the absolute value of the deviation (i.e., numerical value without regard to sign) from the mean is mean deviation.

$$\delta_x = \frac{\Sigma \mid x_i - x \mid}{n}$$

For grouped or interval data

$$\delta_x = \frac{\Sigma f_i \mid x_i - x \mid}{n}$$

Note: Instead of mean ($\bar{X}$), median (M) or mode (Z) can also be used for calculation of standard deviation.

Example : 4 6 7 8 9 10 11 11 11 12 13

$$\delta_x = \frac{14 - 9.271 + 16\text{-}9.271 + \ldots\ldots\ldots + 113 - 9.271}{11} = \frac{24.73}{11} = 2.25$$

Coefficient of mean deviation is the mean deviation divided by the average. It is a relative measure of dispersion and is comparable to similar measure of other series of data. Mean deviation and its coefficient are used to judge the variability and they are better measure than range.

Coefficient of MD = δ_x / $\bar{X}$

Example: 4 6 7 8 9 10 11 11 11 12 13

Coefficient of MD = δ_x / $\bar{X}$ = 2.25 / 9.27 = 0.24

2.2.2.2.3 Standard Deviation (SD)

Standard Deviation is the square root of the average of squares of deviations (based on mean). It is very satisfactory and widely used measure of dispersion amenable for mathematical manipulation. The merits of this measure ire that it does not ignore the algebraic signs and it is less affected by fluctuations of sampling. It can be considered as a kind of average (the 'root mean square') of the absolute deviations of observations from the mean.

. $\sigma = \sqrt{[\Sigma (x_i - \bar{x})^2 / n]}$.

Example: 4 6 7 8 9 10 11 12 13

$\sigma = \sqrt{[(4-9.27)^2 + (6-9.27)^2 + \ldots\ldots + (13 - 9.27)^2 / 11]} = 2.64$

Coefficient of S D is S D divided by mean. Coefficient of SD is a relative measure and is often used for comparing with similar measure of other series.

Example: 4 6 7 8 9 10 11 12 13

i.e., 2.64 / 9.27 = 0.28

Variance is square of S D. Alternatively, variance is the average of the squared deviations.

i.e., $\mathrm{VAR} = \Sigma (x_i - \bar{x})^2 / n$

Example: 4 6 7 8 9 10 11 12 13

i.e., $(2.64)^2 = 6.97$

Coefficient of Variation is coefficient of S D multiplied by 100

Example: 4 6 7 8 9 10 11 12 13

i.e., 0.28 x 100 = 28

For frequency of grouped or interval data

$\sigma = \sqrt{[\Sigma f_i (x_i - \bar{x})^2 / \Sigma f_i]}$

Indirect method using assumed average uses the formula

. $\sigma = \sigma = \{\sqrt{[(\Sigma f d^2 / n) - (\Sigma f d)^2) / n^2]}\}$ (i)

Where, d = Distance of class from the assumed average class

$n = \Sigma f_i$

Calculation of the **SD** (σ) from a frequency distribution of grouped or interval data of price (in Rs.) of popular science books in Kannada (Table 9.12) using **assumed average method** is shown below:

Price (in Rs.) (class)	Frequency (f) (No. of books)	Cumulative less than or equal frequency (cf)	Distance of class from the assumed average class (d)	fd	d^2	fd^2
1 -- 10	2	2	-4	-8	16	32
11 -- 20	8	10	-3	-24	9	72
21 -- 30	15	25	-2	-30	4	60
31 -- 40	13	38	-1	-13	1	13
41 -- 50	4	42	0	0	0	0
51 -- 60	4	46	1	4	1	0
61 -- 70	2	48	2	8	4	8
71 -- 80	1	49	3	3	9	9
81 -- 90	0	50	4	4	16	0
Total	50			-56		194

$n = 50 \qquad \sum f d = -56 \qquad \sum f d^2 = 194 \qquad i = 10$

$$\sigma = \{\sqrt{[(\sum f d^2 / n) - (\sum f d)^2) / n^2]}\}\ (i) = \{\sqrt{[(194 / 50) - (-56)^2) / 50^2]}\}\ (10)$$

$$= \{\sqrt{[(3.88) - (1.2544)]}\}\ (10)$$

$$= \{\sqrt{2.6256}\}\ (10)$$

$$= \{1.6204\}\ (10) = 16.204$$

2.2.2.2.4 Quartiles

Quartile measures dispersion when median is used as average. It is useful, as a measure of dispersion, to study special collections of data like salaries of employees. Lower quartile is the value in the array below which there are one quarter of the observations. Upper quartile is the value in the array below which there are three quarters of the observations. Interquartile range is the difference between the quartiles. Unlike range, interquartile range is not sensitive to the sample size and not affected by extreme values. Since quartiles are, in strict sense, measures of location, interquartile range can be called a positional measure of variability. While range is overly sensitive to the number of observations, the interquartile range can either decrease or increase when further observations are added to the sample.

Example: 4 6 7 8 9 10 11 11 12 13

Lower quartile 7

Upper quartile 11

Interquartile range 11 – 7 = 4

Note: The quartiles & the median divide the array into four equal parts. Deciles divide the array into ten equal groups and Percentiles divide the array into one hundred equal groups.

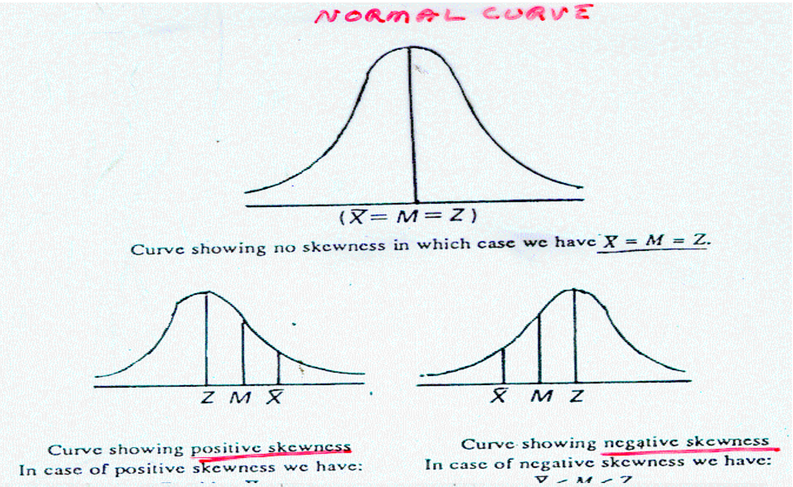
Home work: 5. Workout measures of central tendency and variance

7, 19, 8, 15, 9, 16, 12, 14, 9, 10, 8

2.2.2.3 Measure of Asymmetry (Skewness)

Normal Curve and Skewness

We have already learnt about normal distribution and it is a distribution of items in a series which is perfectly symmetrical. The curve drawn from normal distribution data is bell shaped and shows no asymmetry (i.e., skewness) as X = M = Z in a normal curve. Asymmetrical distribution which has skewness to the right (i.e., curve distorted on the right) is positive skewness (Z > M > X) and the curve distorted to the left is negative skewness (Z < M< X) (see figure).



Curve showing no skewness in which case we have $X = M = Z$.

Curve showing positive skewness
In case of positive skewness we have:

Curve showing negative skewness
In case of negative skewness we have:

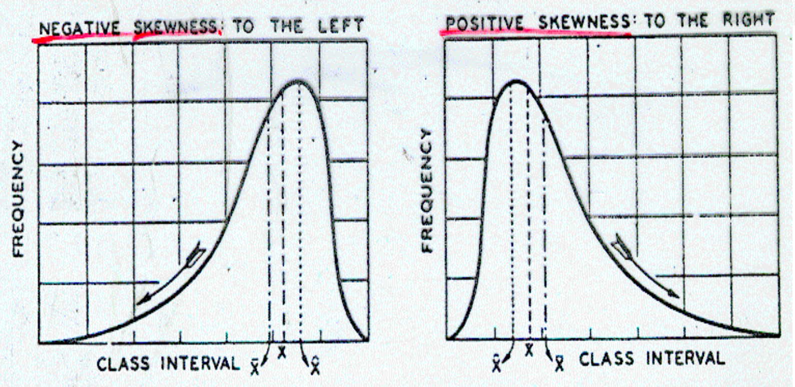


Figure Frequency curves showing negative and positive skewness.

Skewness is the difference between the mean and mode or median (i.e., X – Z or X – M). Skewness shows the manner in which the items are clustered around the average. It is useful in the study of formation of series and gives idea about the shape of the curve.

Coefficient of Skewness about the mode = $\frac{\text{Mean - Mode}}{\text{Standard deviation}}$

i.e., (J) = X – Z / σ

Coefficient of Skewness about the median = $\frac{\text{3 X (Mean - Median)}}{\text{Standard deviation}}$

i.e., (J) = 3 (X – M) / σ

Kurtosis is a measure of flat-topped ness of a curve, (i.e, humpedness). It indicates the nature of distribution of items in the middle of a series.

Mesokurtic is one having Kurtic in the center (i.e., normal curve).

Leptokurtic is more peaked than the normal curve.

Platykurtic is more flat than the normal curve.

Example: 4 6 7 8 9 10 11 11 11 12 13

Skewness = 9.27 - 11 = - 1.73 (using mode)

or 9.27 - 10 = - 0.73 (using median)

Coefficient of Skewness about mode, j = -1.73/2.64 = -0.66

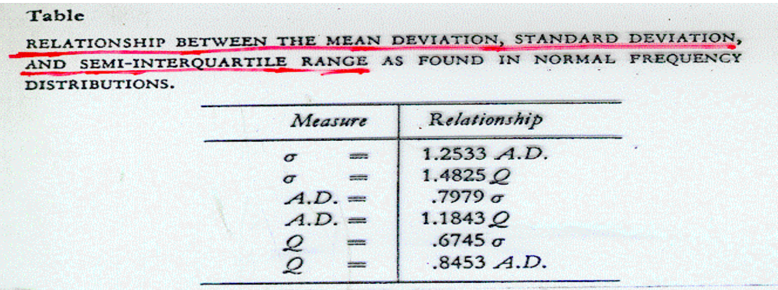
Coefficient of Skewness about median, j = (9-0.73) X 3 / 2.64 = - 0.83

Hence negatively skewed.

Home work: 6. Check the following for positive skewness;

7, 8, 8, 9, 9, 10, 12, 14, 15, 16, 18

Relationship between Measures of Variability (M D, S D and Semi-interquartile Range)



Table

RELATIONSHIP BETWEEN THE MEAN DEVIATION, STANDARD DEVIATION, AND SEMI-INTERQUARTILE RANGE AS FOUND IN NORMAL FREQUENCY DISTRIBUTIONS.

Measure		*Relationship*
σ	=	1.2533 *A.D.*
σ	=	1.4825 *Q*
A.D.	=	.7979 σ
A.D.	=	1.1843 *Q*
Q	=	.6745 σ
Q	=	.8453 *A.D.*

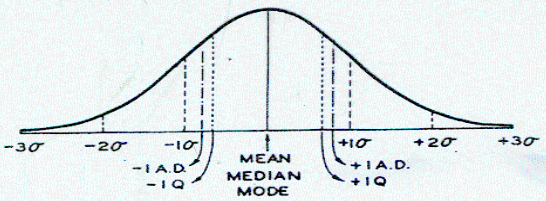


Figure Normal curve showing the relationship of measures of variability.

2.3 Some Statistical Tests

2.3.1. Chi-Square Test

As said before, bivariate and multivariate measures are beyond the scope of this unit. However a list of such measures are given below for information. An important non-parametric test for significance of association as well as for testing hypothesis regarding (i) goodness of fit and (ii) homogeneity or significance of population variance is chi-square test. As a non-parametric test it is used when responses are classified into two mutually exclusive classes like favour - not favour, like - dislike, etc. to find whether differences exist between observed and expected data. In other words, it is used to decide whether observed frequencies of occurrences of a qualitative attribute differ significantly from the expected or theoretical frequencies.

$$\chi^2 = \Sigma (O_{ij} - E_{ij})^2 / E_{ij}$$

Where, O_{ij} = Observed frequency of the cell in i th row & j th column

E_{ij} = Expected frequency of the cell in i th row & j th column

$$\text{Expected frequency of any cell} = \frac{\text{total for the row of that cell} \times \text{total for the column of that cell}}{\text{Grand total}}$$

Degrees of freedom, df = (c-1) (r-1)

If the calculated value of χ^2 is equal or more than that tabulated for the given df the association is significant.

Note: 1. χ^2 is not a measure of degree of relationship.

2. It assumes that random observations are random and items in the sample are independent.

3. Its constraints are that relation has to be linear and no cell contains less than five as frequency value and over all number of items must be reasonably large (Yale's correction can be applied to a 2x2 table if cell frequencies are smaller than five); Use Kolmogorov – Smirnov test.

Example : An opinion poll conducted by a library among its users about the duration for which books should be issued showed the following result: 15 days, 21 days & 30 days are respectively preferred by 18, 31 & 20 respondents. Test the data whether deference is significant in observed data at 0.5 significance level.

	Duration of issue of books		
	15 days	21 days	30 days
Observed preference(O_{ij})	18	31	20
Expected preference (E_{ij})	23	23	23

$$\chi^2 = \Sigma (O_{ij} - E_{ij})^2 / E_{ij} = (18 – 23)^2 /23 + (31 – 23)^2 /23 + (20 – 23)^2 /23$$

$$= 4.216$$

df = k – 1 = 3 – 1 =2

Tabulated value of χ^2 at 2 df and α = 0.05 is 5.991. (This can be had from the standard table provided in books). Hence the H_o (null hypothesi) is accepted or there is no significant difference between the observed and expected data.

Phi Coefficient, $\phi = \Sigma \chi^2 / N$, as a non-parametric measure of coefficient of correlation helps to estimate the magnitude of association.

Cramer's V-measure, $V = \phi^2 / \sqrt{}$min. (r-1), (c-1)

Coefficient of Contingency, C = $\sqrt{}\ \chi^2 / \chi^2 + N$, also known as coefficient of mean square contingency, is a non-parametric measure of relationship useful where contingency tables are higher order than 2x2 and combining classes is not possible for Yule's coefficient of association.

Example: Given below is the data regarding reference queries received by a library. Is there a significant association between gender of user and type of query ?

	L R	S R	Total
	query	query	
Male users	17	18	35
Female users	3	12	15

Total	20	30	50

Expected frequencies are worked out like E_{11} = 20X35 / 50 = 14

Expected frequencies are:

	L	S	Total
M	14	21	35
W	6	9	15
Total	20	30	50

Cells	O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2 / E_{ij}$
1,1	17	14	3	9/14 = 0.64
1,2	18	21	-3	9/21 = 0.43
2,1	3	6	-3	9/6 = 1.50
2,2	12	9	3	9/9 = 1.00
Total (Σ)				χ^2 = 3.57

df = (C-1) (r-1) = (2-1) (2-1) = 1

Table value of χ^2 for 1 df at 5 % significance is 3.841. Hence association is not significant.

Home work: 7. A library has incurred the following expenditure for two different years. Is the pattern of expenditure changed significantly between the years ? (α = 0.5)

Year	Expenditure in lakhs of Rupees			
	Journals	Books	Others	Total
1990-91	50	26	16	92
1994-95	86	30	19	135
Total	136	56	35	227

2.3.2 t-test, Z-test and F-test are used to compare two sets of quantitative data to determine whether they are similar or whether a significant difference exists.

t and Z tests are similar and the difference is only that t-test is used when the number of observations are less than 30 and Z-test is used when the number of observations are more than 30. Both are used to compare whether mean of a sample is significantly differ from the mean of the population or means of two different samples of the same population differ significantly. In either case the standard deviation should be known.

F-test uses variance (the square of standard deviation) to decide whether or not the variances of two samples are significantly different in order to decide that the two samples are drawn from the same population.

2.3.3 Finding Relationships among Variables

Explore to find relationship between two or more variables; If related, determine directly or inversely related & find degree of relation; Is it cause and effect relationship ? If so, degree and direction. Some of the techniques used are listed below:

1. Association (Attributes)
   (I) Cross tabulation

(ii) Yule's co-efficient of association
(iii) Chi- square test
(iv) Co-efficient of mean square contingency

2. Correlation (Quantitative)
   (I) Spearman's (Rank) coefficient of correlation (ordinal)
   (ii) Pearson's coefficient of correlation
   (iii) Cross tabulation and scatter diagram
3. Cause and Effect (Quantitative)
   (I) Simple (linear) & regression
   (ii) Multiple (complex correlation & regression
   (iii) Partial correlation

2.3.4 Other Measures and Statistical Techniques of importance to Research

1. Index number
2. Time series analysis
3. Anova
4. Anocova
5. Discriminant analysis
6. Factor analysis
7. Cluster analysis
8. Model building

2.4 Summary

This unit farming part of research methodology is set out with an objective of introducing you to basics of statistics in general and procedures and techniques required for processing and analysis of data in particular. You have learnt the importance of statistics in research. You have also noticed verities of techniques under descriptive statistics and inferential analysis. Having noted briefly the difficulties of processing and analysis of qualitative data, the rest of the unit exposes you to processing and analysis of quantitative data. You have understood the four basic types of data (namely, nominal, ordinal, interval and ratio data), variable, experiment, hypothesis and normal distribution.

Under processing of data, editing (both field editing and central editing), coding, classification, creation of class intervals, tabulation and graphical presentation of both discrete and grouped data are explained with examples. Under analysis of data, univarate measures of central tendency (mean, median, mode), dispersion (range, mean deviation, standard deviation, quartiles), asymmetry/ skewness of data are explained with illustrations.

The unit has also provided you with details and examples for using data analysis tools like chi-square test, t- test, Z – test and F – test.

Other tools for finding relationships among variables, and bivariate and multivariate measures are enlisted for the benefit of advanced processing and analysis of research data.

Summary of example: **4 6 7 8 9 10 11 11 11 12 13**

Univariate Measures:

A. Central Tendency	
1. Mean	9.27
2. Median M	10
3. Mode Z	11
B. Dispersion	
1. Range	9
2. Mean deviation	2.25
3. Coefficient of MD	0.24
4. Standard deviation	2.64
5. Coefficient of SD	0.28
6. Coefficient of variation	28
7. Variance	6.97
8. Lower quartile	7
9. Upper quartile	11
10. Inter quartile range	4
C. Asymmetry	
1. Skewness	
w.r.t. Mode	1.73
w.r.t. Median	0.73
2. Coefficient of Skewness	
w.r.t. Mode	0.66
w.r.t. .Median	0.8

2.5 Keywords

Absolute value: A number or computation where the sign is ignored, that is, seen to be positive.

Canonical analysis deals with simultaneously predicting a set of dependent variables (both measurable & non measurable).

Cell: One section of a percentage table.

Central tendency: Statistics that identify a category of a variable, either real or hypothetical, around which a distribution clusters, hangs together, or centers.

Coding: Relabeling raw data in some manner. (The process of changing observations into usable data).

Contingency table analysis: Synonym for percentage table analysis.

Cross-sectional survey: Survey in which units are measured at *one* point in time.

Cross-tabs: Synonym for percentage table analysis.

Cumulative frequency distribution: A frequency distribution where the frequencies of each category are successively added to the sum of the frequencies preceding that category.

Cumulative survey: Surveys where data are collected at different points in time but all the data are put together in one group.

Cutting points: The values that border particular categories of a variable and determine how many categories of a variable there are going to be.

Deviation: How far one score is from another.

Dispersion: Statistics that give an estimate of how spread out, scattered, or dispersed the scores of a distribution are.

Frequency polygon: A line graph representing a simple frequency distribution.

Geometric Mean is nth Root of the product of the values of n items.

$GM = n \sqrt{} \Pi x_i \quad X \quad \sqrt{}_n \; x_1 \, x_2 \ldots . x_n$

Example: 4 6 9

$GM = \sqrt{}_3 \; 4 \text{ x } 6 \text{ x } 9 = 6$

NOTE : 1. Logarithm is used to simplify; 2. GM is used in the preparation of indexes (i.e., determining average percent of change) and dealing with ratios.

Grouped data: Data are recorded in such a way that the categories of the variable include more than one distinct type of observation, or score, or value.

Harmonic Mean is reciprocal of the average of reciprocals of the values of items in series.

$$HM = \frac{n}{1/x_1 + 2/x_2 + \ldots . f_i/ x_n} = \frac{\sum f_i}{\sum f_i/x_i}$$

Example: 4 5 10

$$HM = \frac{3}{1/4 + 1/5 + 1/10} = 60/1 = 5.45$$

Harmonic Mean has limited application as it gives largest weight to the smallest item and smallest weight to the largest item. It is used in cases where time and rate are involved (ex: time and motion study)

Histogram: A type of graph using bars to represent a simple frequency distribution.

Kurtosis: How pointed the peaks are in a frequency polygon.

Longitudinal panel survey: Survey in which the same units are measured at different points in time on the same variables.

Mean: Provides a score that each case would have if the variable were distributed equally among all observations. The arithmetic average.

Mean deviation: A deviation score which tells how far away from the mean each observation would be if all the observations were the same distance away from the mean. (The average distance of each score from the mean.)

Median: A score that separates all of the observations (frequencies) into two equal-sized groups.

Midpoint: The center of a category obtained by subtracting the lower limit of a category from the upper limit, dividing by 2, and adding that figure to the lower limit.

Mode: Score (value or category) of the variable which is observed most frequently.

Multivariate analysis is concerned with simultaneously analysing more than two variables.

Multiple regression analysis is predicting dependent variable based on its covariance with all concerned independent variables.

Multiple Discriminate Analysis is predicting an entity's possibility of belonging to a particular group based on several predictors.

Multi-ANOVA is extension of two-way ANOVA (Analysis of Variance) and determines ratio of among group variance to within group variance.

Ogive: A line graph representing a cumulative distribution.

Parameter : A characteristic of a population

Percentage distributions: Distributions where frequencies have been converted into percentages.

Percentage table analysis: Using percentages to determine mutual patterned change (a relationship) between two or more variables.

Range: The difference between extreme scores calculated by subtracting the lower limit of the lowest score from the upper limit of the highest score.

Raw data: Types of observations as originally recorded. (There is no such thing as cooked data.)

Scan (scanning): Moving the eyes from left to right in order to "see" change.

Simple frequency distribution: A two-column table in which the categories of the variable are placed in the left-hand column and the sum of the tallies (frequency) is placed in the right-hand column.

Skewed distribution: A distribution where more scores occur at one end than the other.

Standard deviation (SD): A measure of dispersion, based on deviations from the mean, calculated by taking the square root of the average squared deviations of each score from the mean.

Statistic: A characteristic of a sample (estimation of a parameter from a statistic is the prime objective of sampling analysis)

Strength of a relationship: Combination of the extent and precision of relationships used in percentage table analysis and computed by determining percentage differences from column to column.

Tabular analysis: Synonym for percentage table analysis.

Tally: Counting how many cases one observes in each category of variable.

Univariate analysis: The description and/or summarization of the scores (categories, observations) of one variable.

2.6 Answers to home works

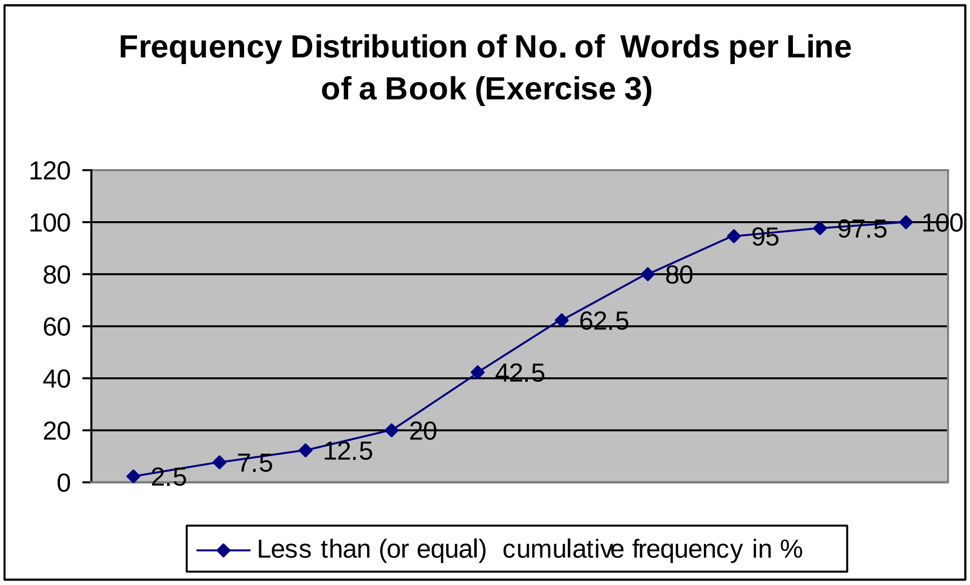
1. Work out a frequency table with less than cumulative and more than cumulative frequencies for the raw data of number of words per line in a book given below :

12	10	12	09	11	10	13	13
07	11	10	10	09	10	12	11
01	10	13	10	15	13	11	12
08	13	11	10	08	12	13	11
09	11	14	12	07	12	11	10

Frequency distribution of number of words per line in a book :

No. of Words per line	***Frequency (No. of Words)***		***Less than (or equal) cumulative frequency***	***More than (or equal) cumulative frequency***
	No.	***%***	***%***	***%***
05	1	2.5	2.5	100
07	2	5.0	7.5	97.5
08	2	5.0	12.5	92.5
09	3	7.5	20.0	87.5
10	9	22.5	42.5	80.0
11	8	20.0	62.5	57.5
12	7	17.5	80.0	37.5
13	6	15.0	95.0	20.0
14	1	2.5	97.5	5.0
15	1	2.5	100	2.5
TOTAL	40			

2. Prepare a less than or equal cumulative line graph for frequency table developed for Exercise 3 (use Excel of MS Office software).



3. Workout measures of central tendency and variance

7, 19, 8, 15, 9, 16, 12, 14, 9, 10, 8

(**Discrete data**): 7 19 8 15 9 16 12 14 9 10 8

Mean, M = 7 + 19 + 8 + 15 + 9 + 16 + 12 + 14 + 9 + 10 + 8 / 11 = 126/11 = 11.45

Median**,**

Data : 7 19 8 15 9 16 12 14 9 10 8

Ascending order : 7 8 8 9 9 10 12 14 15 16 18

Serially numbered frequencies : 1 2 3 4 5 6 7 8 9 10 11

M = Value of [N+1] / 2 th item = [11 + 1] / 2 = 6th item i.e., 10

Mode,

Ascending order : 7 8 8 9 9 10 12 14 15 16 18

^

As both 8 and 9 occur twice, the distribution is bi-modal and it has two modes.

Range,

7 8 8 9 9 10 12 14 15 16 18

Range = 18 - 7 = 11

Mean Deviation,

$$\delta_X = \frac{7- 11.45 + 8 -11.45 + 8 -11.45 +9 -11.45 + ...18 -11.45}{11} = \frac{34.45}{11} = 3.13$$

Coefficient of MD = δ_X / $\overline{X}$ = 3.13 / 11.45 = 0.27

Standard Deviation, $\sigma = \sqrt{[\Sigma (x_i - \overline{x})^2 / n]}$.

$\sigma = \sqrt{[(7- 11.45)^2 + (8 -11.45)^2 +\ldots\ldots+ (18 -11.45)^2 / 11]} = \sqrt{139.275/11} = 11.8/ 11 = 1.07$

Coefficient of S D , $\tilde{\sigma} \overline{X}$ = 1.07/ 11.45 = 0.0934

Variance, i.e., VAR $= \Sigma (x_i - \overline{x})^2 / n$

i.e., $(1.07)^2 = 1.1449$

Coefficient of Variation, i.e., 1.1449 x 100 = 114.49

Quartile of data : 7 8 8 9 9 10 12 14 15 16 18

Lower quartile 8

Upper quartile	16
Interquartile range	16 – 8 = 8

4. Check the following for positive skewness;

7 8 8 9 9 10 12 14 15 16 18

Example: 4 6 7 8 9 10 11 11 11 12 13

Skewness = 11.45 – 8 = 3.45 and 11.45 – 9 = 2.45 (using mode)

11.45 - 10 = 1.45 (using median)

Coefficient of Skewness about mode, j = X – Z / σ = 11.45 – 8/ 1.07 = 3.22

j = X – Z / σ = 11.45 – 9/ 1.07 = 2.29

Coefficient of Skewness about median, j = 3 (X – M) / σ = (11.45 - 10) X 3 / 1.07 = 4.06

Hence positively skewed

5. A library has incurred the following expenditure for two different years. Is the pattern of expenditure changed significantly between the years ? (α = 0.5)

Year	Expenditure in lakhs of Rupees			
	Journals	Books	Others	Total
1990-91	50	26	16	92
1994-95	86	30	19	135
Total	136	56	35	227

Expected frequencies are worked out like E_{11} = 136X92 / 227 = 55

Expected frequencies are:

	Journals	Books	Others	Total
1990-91	55	23	14	92
1994-95	81	33	21	135
Total	136	56	35	227

Cells	O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2 / E_{ij}$
1,1	50	55	-5	25/55 = 0.45
1,2	26	23	3	9/23 = 0.39
1,3	16	14	2	4/14 = 0.29
2,1	86	81	5	25/81 = 0.31
2,2	30	33	-3	9/33 = 0.27
2,3	19	21	-2	4/21 = 0.19
Total (Σ)				χ^2 = 1.90

df = (C-1) (r-1) = (3-1) (2-1) = 2

Table value of χ^2 for 2 df at 5 % significance is 5.991. Hence pattern of expenditure has not changed significantly between the years.

2.7 References

1. Anderson, T W and Sclove, Stanley L. An introduction to the statistical analysis of data. Boston : Houghton Miffin Company, 1978.
2. Best, Joel. Damned lies and statistics. California: University of California Press, 2001.
3. Best, Joel. More damned lies and statistics; how numbers confuse public issues. Berkeley: University of California Press, 2004
4. Koosis, Donald J. Baseness statistics. New York: John Wiley,1972.
5. Miller, Jane E. The Chicago guide to writing about numbers. Chicago: the University of Chicago Press, 2004.
6. Rodger, Leslie W. Statistics for marketing. London: Mc-Graw Hill, 1984.
7. Salvatoe, Dominick. Theory and problems of statistics and econometrics (Schaum's outline series). New York: McGraw-Hill, 1982.
8. Spiegel, Murray R. Schauim's outline of theory and problems of statistics in SI units. Singapore: Mc Graw Hill ,1981.
9. Simpson, I.S. How to interpret statistical data: a guide for librarians and information scientists. London: Library Association, 1990.
10. Walizer, Michael H and Wienir, Paul L. Research methods and analysis: searching for relationships. New york: Harper & Row, 1978.

About the Author

Dr. M. S. Sridhar is a post graduate in Mathematics and Business Management and a Doctorate in Library and Information Science. He is in the profession for last 36 years. Since 1978, he is heading the Library and Documentation Division of ISRO Satellite Centre, Bangalore. Earlier he has worked in the libraries of National Aeronautical Laboratory (Bangalore), Indian Institute of Management (Bangalore) and University of Mysore. Dr. Sridhar has published 4 books, 81 research articles, 22 conferences papers, written 19 course materials for BLIS and MLIS, made over 25 seminar presentations and contributed 5 chapters to books.

E-mail: sridharmirle@yahoo.com, mirlesridhar@gmail.com, sridhar@isac.gov.in
Phone: 91-80-25084451; Fax: 91-80-25084476.

NORMAL DISTRIBUTION

1. A special continuous distribution
2. Great many techniques used in applied statistics are based on this
3. Many populations encountered in the course of research in many fields seems to have a normal distribution to a good degree of approximation (I.o.w., nearly normal distributions are encountered quite frequently)
4. Sampling distributions based on a parent normal distributions are manageable analytically
5. The experimenter musts know, at least approximately, the general form of the distribution function which his data follow. If it is normal, he may use the methods directly; if it is not, he may transform his data so that the transformed observations follow a normal distribution. When experimenter does not know the form of his population distribution, then he must use other more general but usually less powerful methods of analysis called non-parametric methods