UNIT 10 DESCRIPTION & INFERENCE FROM SAMPLE DATA

Objectives

After studying this unit, you should be able to:

- distinguish between descriptive and inferential statistic

- explain the usefulness of univariate tables

- interpret the results of descriptive measures related to univariate analysis

- discuss various inferential techniques as applicable to univariate analysis

Structure

10.1 INTRODUCTION

The unit on data processing covered coding, tabulation and presentation of data collected from both primary and secondary sources. Once the data has been collected, it needs to be analysed for making sound marketing decisions. Analysis can be done on the basis of one variable, two variables & more than two variables at a time. When we are analysing one variable at a time, the analysis is-named as Univariate Analysis. When the association between two variables is of concern, the type of analysis is termed as Bivariate Analysis. However, when simultaneous relationship between three or more variables is of concern to the researcher, a Multivariate Analysis is taken. In this unit we will be concerned with description and inference from the sample data considering one variable at a time. The association between two variables will be covered in Unit 11, whereas blocks 4 & 5 covering multivariate analysis will deal with three units each describing the simultaneous relation between variables.

Activity 1

Define the following concepts with the help of examples:

1.   Univariate Analysis

……………………………………………………………………………………

2.   Bivariate Analysis Multivariate Analysis

……………………………………………………………………………………

3.   Multivariate Analysis

……………………………………………………………………………………

10.2 DISTINCTION BETWEEN DESCRIPTION & INFERENCE

Before we proceed further, it is essential to distinguish between descriptive and inferential statistic. The descriptive statistic deals with the summary measures relating to the sample data. The examples are as under:

1. *What is the average income of the sampled respondents?*
2. *What is the dispersion of ages of the sampled respondents?*
3. *What is the distribution of respondents by sex?*
4. *Which income group has maximum number of users of a particular brand of product?*
5. *What percentage of respondents are married?*
6. *Which newspaper is read by majority of the sampled respondents?*
7. *Is there any association between the income and willingness to buy of the product?*
8. *What proportion of consumers of a particular brand of product are professional?*
9. *Is there any relationship between age and the frequency of purchase of a particular brand of a product?*

The inferential statistic deals with drawing inferences about the population parameters on the basis of sample results. The concern, here, is to generalize the results obtained from sample. Here the analysis is based on probability theory and a necessary condition for the use of inferential statistic is the randomization of the sample. The following are the examples of some questions which are addressed under inferential statistic.

1. *Is the average income of population greater than Rs. 9,000 per month?*
2. *Does the difference in the frequency of purchase of a particular brand of a product vary with age significantly?*
3. *Does the income of users and non-users of the product vary significantly?*
4. *Is the decline in sales of the product significant?*

Both descriptive and inferential statistic are important in making sound marketing decisions. The researcher must be aware of the type of analysis to be carried out since both descriptive and inferential techniques vary with the levels of measurement. Appendix 1 and 2 of this unit review briefly such techniques for carrying out univariate analysis.

Activity 2

Distinguish between descriptive and inferential analysis of data.

....
....
....

10.3   FRAMEWORK FOR ANALYSIS

Once the raw data is collected it needs to be brought into array and presented in the form of tables and graphs so that further analysis could be carried out with ease. Having done this, one would be interested in finding out the summary measures of sample data viz. measures of central tendency, dispersion etc. The marketing researcher would also like to use inferential techniques to draw inferences about population on the basis of sample results to arrive at generalization. As mentioned in Unit 7 on Attitude Measurement & Scaling, the method of analysis depends upon the levels of measurements used in measuring the data. It was mentioned that there are four types of measurements namely Nominal Scale, Ordinal Scale, Interval Scale & Ratio Scale. The statistical methods applicable under each of these scales of measurement have already been summarized in Table 1 of Unit 7.

We will take up an example to illustrate the use of descriptive and inferential analysis for our purpose.

Example: XYZ company is considering the introduction of a new packaging design for its product. To date, the product was packed in a glass jar. The company wishes to replace it

with plastic containers. However, before doing so it wants to ascertain the opinion from its customers regarding their preference for new packaging or the inclination to stick to the old one.

The product of the company is sold in all parts of the country. It decides to take a random sample of 60 covering customers from northern, southern, eastern and western part of the country. The question asked to the customers is

How interested are you in buying the product if the packaging for the product is switched to the plastic container?

Extremely interested	Interested	Indifferent	Not interested	Not at all interested
5	4	3	2	1
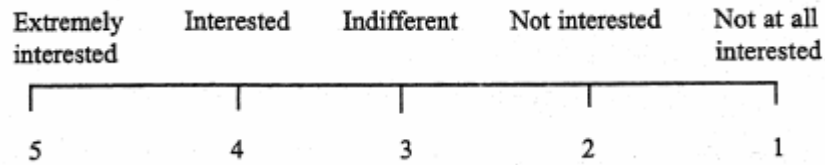
The other variables on which data collected from respondents included sex, marital status age, education and monthly income.

The collected data is presented below in Table 1.

Table 1: SURVEY DATA ON PREFERENCE FOR NEW PACKAGING DESIGN

RESPONDEN NUMBER	PREFERENCE	PREFERENCE CODE	SEX	MARITAL STATUS	AGE IN YEARS	EDUCATION CO	ACTUAL INCOME (RS.)	INCOME CODE	REGION CODE
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	3	Not interested	M	M	25	11	2800	21	31
2	2	Not interested	M	M	27	13	3782	22	33
3	2	Not interested	F	S	28	13	6072	23	32
4	5	Interested	M	S	24	14	9050	24	34
5	3	Not interested	M	M	30	12	4982	22	32
6	1	Not interested	M	S	35	12	5375	23	31
7	2	Not interested	F	M	39	12	4210	22	34
8	4	Interested	F	S	37	14	7563	23	31
9	4	Interested	M	M	36	13	5384	23	33
10	2	Not interested	F	M	29	13	3218	22	32
11	2	Not interested	M	S	41	11	1976	21	31
12	1	Not interested	M	S	43	11	3723	22	32
13	2	Not interested	M	M	40	12	2187	21	34

14	5	Interested	F	M	31	13	9572	24	34
15	4	Interested	M	S	35	13	5689	23	33
16	3	Not interested	F	S	45	12	5791	23	32
17	2	Not interested	F	M	46	11	2500	21	31
18	4	Interested	M	S	38	13	3780	22	33
19	5	Interested	F	M	40	14	5004	23	34
20	4	Interested	M	S	39	14	8730	24	32
21	2	Not interested	M	S	27	13	4611	22	31
22	3	Not interested	M	M	27	13	6666	23	34
23	4	Interested	F	S	31	13	6129	23	33
24	5	Interested	M	M	24	14	8289	24	32
25	2	Not interested	F	S	32	13	7270	23	33
26	4	Interested	M	M	38	14	6235	23	33
27	2	Not interested	F	M	25	13	4219	22	31
28	3	Not interested	F	M	29	12	4136	22	32
29	4	Interested	M	S	40	13	3784	22	34
30	3	Not interested	F	S	45	14	10000	24	33
31	4	Interested	F	M	35	13	6782	23	32
32	5	Interested	F	S	32	14	6115	23	31
33	1	Not interested	M	S	29	12	7068	23	32
34	2	Not interested	M	M	42	12	2991	21	34
35	1	Not interested	F	M	27	11	3787	22	33
36	3	Not interested	M	M	29	11	2736	21	32
37	4	Interested	M	S	28	11	5436	23	34
38	3	Not interested	M	M	41	12	1985	21	31
39	4	Interested	M	S	43	13	4803	22	33

40	3	Not interested	F	S	50	12	1999	21	31
41	5	Interested	M	M	52	13	5072	23	32
42	3	Not interested	F	M	47	14	5815	23	33
43	4	Interested	F	S	30	13	3925	22	34
44	2	Not interested	M	M	53	11	2673	21	32
45	1	Not interested	M	S	39	12	4271	22	31
46	4	Interested	F	M	55	13	3791	22	34
47	3	Not interested	M	S	49	11	4213	22	33
48	4	Interested	F	M	38	12	5824	23	32
49	3	Not interested	M	S	27	13	3270	22	34
50	4	Interested	F	S	46	13	6184	23	31
51	3	Not interested	M	S	47	14	4634	22	31
52	2	Not interested	F	M	38	14	6224	23	33
53	1	Not interested	F	M	28	13	3182	22	32
54	5	Interested	M	M	43	11	8467	24	31
55	3	Not interested	F	S	39	12	2789	21	34
56	4	Interested	M	S	44	12	6972	23	31
57	5	Interested	F	M	29	13	8131	24	33
58	4	Interested	M	S	41	11	2835	21	32
59	5	Interested	F	S	26	12	5138	23	34
60	4	Interested	M	M	32	13	9220	24	33

From the above table, the following points may be noted:

(i)   Column one indicates the respondent number.

(ii)   Column two indicates the rating given by the respondents for their preference or dislike for the new packaging on a 5-point scale. As discussed earlier the scale value of 5 indicates that the respondent is extremely interested, whereas the scale value of 1 represents that respondent is not at all interested in the new packaging.

(iii)   In column three, the respondents are divided into two-categories - Interested & Not interested. All those respondents whose score were 4 and 5 are grouped under "Interested" and those whose score were 1, 2, 3 are grouped as "Not-interested".

(iv) Column four indicates the sex of the respondent, `M' stands for male whereas stands for Female respondent.

(v) Column five exhibits the marital status of the respondent. If a respondent is married he/she is coded as 'M', where as if the respondent is single he/she is coded as `S'.

(vi) Column six gives the age of the respondent measured as a ratio scale variable.

(vii) Column seven indicates the education category to which the respondent belongs. There are four categories for this variable namely below higher-secondary, higher secondary, graduation and post graduation. The code for below higher secondary is 11, for higher secondary 12, for graduation 13 and for post graduation 14. This is an ordinal scale data and therefore may also be used as a nominal variable.

(viii) Column eight gives the monthly income of the respondent. Again this is a ratio scale data.

(ix) In column nine, the data on monthly income as given in column eight has been classified into four categories viz. less than Rs. 3000 coded as 21, Rs. 3000 - Rs. 4999 coded as 22, Rs. 5000 - Rs. 7999 coded as 23 and Rs. 8000 and above coded as 24. Again this is the case of ordinal scale data which can also be converted into nominal scale data.

(x) Column ten indicates the region from where the respondent is selected. Four regions namely, north, south, east and west were considered with respective codes given by 31, 32, 33 and 34.

10.4 ANALYSIS OF SUMMARY MEASURES OF SAMPLE DATA

The concern is to analyse sample data with measures of central tendency and dispersion.

Preference: The preference data as given in column two is an interval scale data mitt therefore the measures of central tendency and dispersion applicable in this case are mean and standard deviation. Of course, all other measures applicable in the case of ordinal and nominal scale data can also be used in this case. The summary results corresponding to preference data (column 2) is given below in Table 2

Table 2 : SUMMARY RESULTS OF PREFERENCE DATA			
Measure	Value	Measure	Value
Mean	3.183	Std. Deviation	1.228
Median	3.000	Skewness	-.192

The above table indicates a mean value of 3.183 & median value of 3.0. The standard deviation & skewness are 1.228 & -.192 respectively. The results indicates that the distribution is skewed to the left, has a mean value of 3.183 which is slightly above "indifferent" and is on the side of "interested" category. Median falls in the neutral category. Mode should not be applied to ordinal or interval data unless these have been grouped first.

The same preference data when categorized as "interested" and "not interested' as given in column number three gives the following results. Out of 60 respondents, 27 prefer the new packaging design whereas 33 are either neutral or do not want the

41

change in design. This being nomina¹ scale data, we can only compute mode of the distribution which falls in the category of "not interested". The measure of dispersion applicable in this case is relative frequencies which indicate that 55% of the respondents are 'not interested in the new design.

2. Sex: The Table 3 below gives the distribution of respondent with respect to sex

Table 3 : FREQUENCY DISTRIBUTION OF RESPONDENT BY SEX

Sex	Frequency	Percent
Female (F)	27'	45.0
Male (M)	33	55.0
Total	**60**	**100.0**

The above table indicates that there are 33 male respondents comprising 55% of the sample, whereas there are only 27 female respondents. This being a nominally scaled data, the mode of the distribution corresponds to " male category.

3. Marital Status: The Table 4 below gives the marital status of the sample respondents.

Table 4 : FREQUENCY DISTRIBUTION OF RESPONDENTS BY MARITAL STATUS

Marital Status	Frequency	Per cent
Married (M)	30	50.0
Single(S)	30	50.0
Total °	**60**	**100.0**

The above table indicates that there is an equal number of married and single respondents.

4. Age: The Table 5 below gives-the results corresponding to age of the respondents.

Table 5: SUMMARY RESULTS CORRESPONDING TO AGE VARIABLE

Measure	Value	Measure	Value
Mean	36.417	Std. Deviation	8.164
Median	37.500	Skewness	0.309

The results indicates that the average age of the respondent is 36.417 years. The median value of 37.5 indicates that 30 respondents are below the age of 37.5 whereas the remaining 30 are above the age of 37.5. The distribution is positively skewed.

5. Education: The data on education measured in nominal scale is given in Table 6 below:

Table 6 : FREQUENCY DISTRIBUTION OF RESPONDENTS BY EDUCATION			
Education	**Frequency**	**Per cent**	**Cum Percent**
Below higher secondary	11	183	18.3
Higher secondary	15	25.0	43.0
Graduation	23	38.0	81.0 '
.Post-graduation	11	18.3	100.0
TOTAL	**60**	**100.0**	

It may be observed that about 57% of the respondent have an education level' of Graduation and above. The mode of the distribution is the category "Graduation" Here we are treating this as a nominally scaled data. It can however be considered as ordinal scale data under the assumption that for doing post- graduation one has to spend more number of years than for a graduation and so on.

6.   **Income:** Income data expressed in ratio scale is given in column 8 of Table:-l. The summary measures computed for this is given in the following Table 7.

Table 7: SUMMARY MEASURES FOR INCOME. VARIABLE			
Measure	**Value**	**Measure**	**Value**
Mean	5150.983	Std. Deviation	2081.689
Median -	4993.000	Skewness	0.500

The mean income is 5150.983, the median being Rs. 4993 expressing that 50% of the respondent have income below it and 50% have income above it, The distribution is skewed to the right. The results for income variable when expressed in coded form using coding scheme already explained is given below in Table 8.

Table 8 : FREQUENCY DISTRIBUTION OF RESPONDENTS .BY INCOME			
Value	**Frequency**	**Per cent**	**Cum Per cent**
21	11	18.3	18.3
22	19	31.7	50.0
23	22	3 6.7	86.7
24	8	13.3	100.0
TOTAL	**60**	**100.0**	

It can be noted that the mode in the above table is the income category Rs. 5000 .-
7999. The data may be treated as ordinal scale data and the median of the distribution
lies in the income group Rs. 3000 - Rs. 4999.

7. Region: The regionwise distribution of respondents is given in Table 9 below:

Table 9: FREQUENCY DISTRIBUTION OF RESPONDENT		
BY REGION		
Region	**Frequency**	**Percent**
North	15	25.0
South	16	26.7
East	15	25.0
West	14	23.3
Total	60	100.0

It may be noted that the maximum number. of respondents are from south whereas
'the minimum number is from west. 'However, there does not seem to be much
Variation among the number of respondents from the four regions.

The analysis carried out above by forming one-way table is useful for the following
purposes:

1. These tables provide valuable insights through comparison with other relevant
    distributions. For example our sample results' as, given in Table 3 indicates that
    there were 55% male and 45% female respondents. However, if this distribution
    is not in tune with the population distribution, we may-infer that the sample is
    not representative of the population. .

2. One way table can also suggest useful variable transformation. Consider, e.g. the
    results indicated in Table 7 for the income variable. It is seen that the
    distribution is skewed. Therefore one needs to transfer the data into appropriate
    transformation so as to enable, the analyst to carry out appropriate statistical
    tests which are otherwise not possible in the case of skewed distribution.

Activity 3

How will you use the results of univariate analysis to judge the representatives of a
sample?

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………...

Activity 4

How are (1) measures of central tendency and (2) measures of dispersion both
necessary to describe a sample?

....
....
....
....

10.5   RESULTS FROM INFERENTIAL STATISTIC

As mentioned earlier, inferential statistic attempts to generalize the results from sample to population. We will try to answer following questions:

1.   *Is the distribution of respondents across various regions uniform?*

2.   *Are the respondents uniformly distributed across 4 education level categories?*

3.   *Is the proportion of respondent pre / erring the new scheme different from .50?*

4.   *Is the average income of all respondents significantly different from Rs. 4500?*

To answer first question we test the following hypothesis.

H$_o$ (Null-hypothesis):   The distribution of respondents preferring the new packaging design is uniform across all the four regions.

H$_1$ (Alternative-hypothesis):   The distribution of respondents across the four regions is not uniform.

We are given the observed number of respondents belonging to various regions in Table 10. If the null-hypothesis is true, there should be 15 respondents from each of the four regions. Further to test the hypothesis we need to carry out the test of goodness of fit by computing Chi-square statistic. The results are given below in Table 10:

Table 10: TEST OF GOODNESS OF FIT FOR REGIONS		
S.No.	**Observed Frequency (Oi)**	**Expected Frequency (Ei)**
1	15	15
2	16	15
3	15	15
4	14	15
Total	**60**	**60**

Computed chi-square: $\sum_{i=1}^{k} \dfrac{(O_i - E_i)^2}{E_i}$ =   0.1333

Critical chi-square   =   7.8147
(assuming level of significance = 0.05)

As computed chi-square value is less then the critical chi-square value, we do not reject null-hypothesis. This shows that the distribution of respondents across various region is uniform. The results may be shown diagramatically as follows:

```
                        C H I  -  S Q U A R E
                      .  .
                   .        .
                 .            .
               .                .
              .                   .
             .                      .
            .                         .           Do Not Reject                    Reject
           .                            .
          .                               .
         .                                  .
        .                                     .
       .                                       . .
      .                                          . .    . :
     O                                              : ........ ...............

    0.133                                                                    7.81473
    Computed value of chi-square                                 Table value of chi-square
```
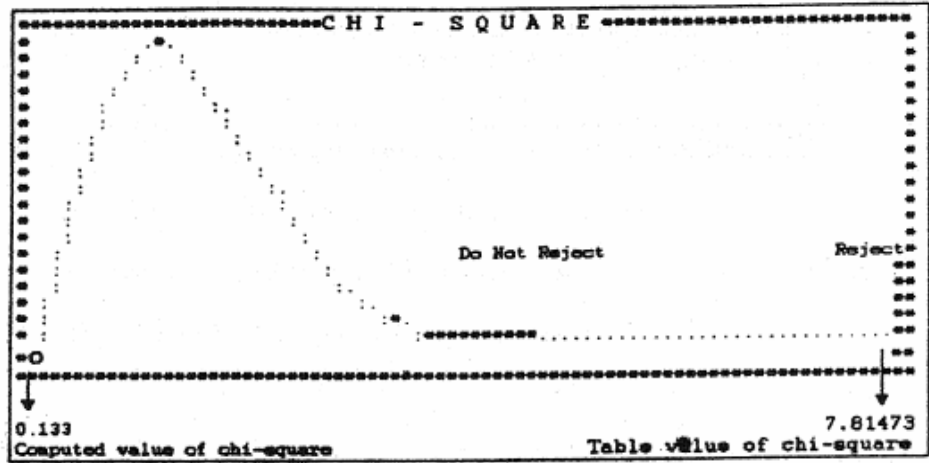
To answer second question the following hypothesis needs to be tested.

H$_o$ (Null-hypothesis):	The distribution of respondents preferring the new packaging design is uniform across all the four education categories.
H$_I$ (Alternative-hypothesis):	The distribution of respondents across the four education categories is not uniform.

The hypothesis can be tested using chi-square test for goodness of fit. Since there are 60 respondents, therefore, under the assumption that the null-hypothesis is true, it may be expected that each of the categories has 15 respondents.
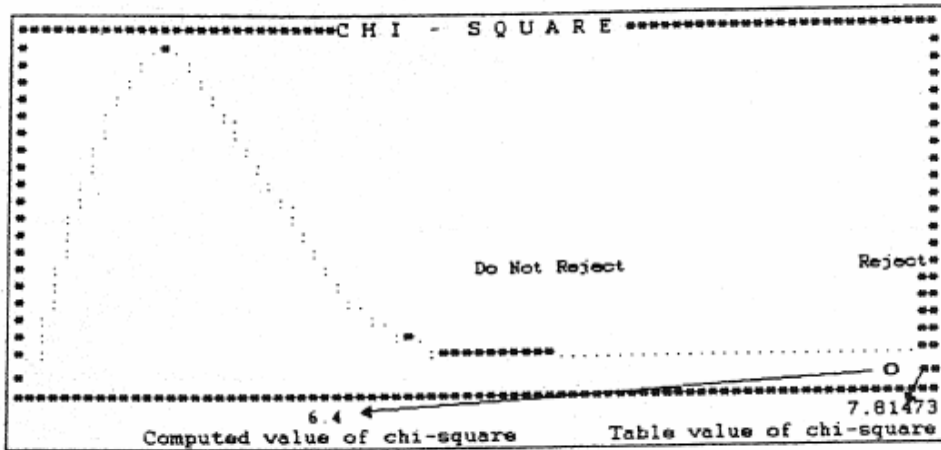
Table 11 : TEST OF GOODNESS OF FIT FOR EDUCATIONAL CATEGORIES

Education Level	Observed Frequency (00	Expected Frequency (Ei)
Below higher secondary	11	15
Higher secondary	15	15
Graduation	23	15
Post graduation	11	15
Total	**60**	**60**

Computed chi-square: $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ = 6.4000

Critical chi-square = 7.8147
(assuming level of significance = 0.05)

As computed chi-square is less than the critical value of chi square we do not reject the null-hypothesis. Therefore, this shows that the distribution of respondents across various educational categories is uniform. The results may also be shown diagrammatically as:

```
**************************C H I - S Q U A R E *****************************
*                                                                        *
*      . .                                                               *
*    :                                                                   *
*  :   :                                                                 *
* :     :                                                                *
* :      :                                                               *
* :       :                                                             **
* :        :                                                            **
* :        :             Do Not Reject                      Reject*     **
* :         :                                                          **
* :          :                                                         **
* :           :                                                        **
* :            :.                                            .**       **
* .*.             : :**********.............................**         **
*                 :                                                O **..**
*                                                              .****
**************************************************************************
              6.4                                           7.81473
         Computed value of chi-square          Table value of chi-square
```

To answer the third question, we test the following hypothesis.

H_0 (Null-hypothesis): The proportion of respondents preferring the new packaging design equals .50. ($\pi = 0.50$)

H_1 (Alternative-hypothesis): The proportion is different from .50. ($\pi \neq 0.50$)

The sample proportion $p = \dfrac{27}{60} = .45$.

As sample is large (n = 60) and both np & n (1-p) are greater than 5, we may use normal approximation to binomial distribution and apply Z test.

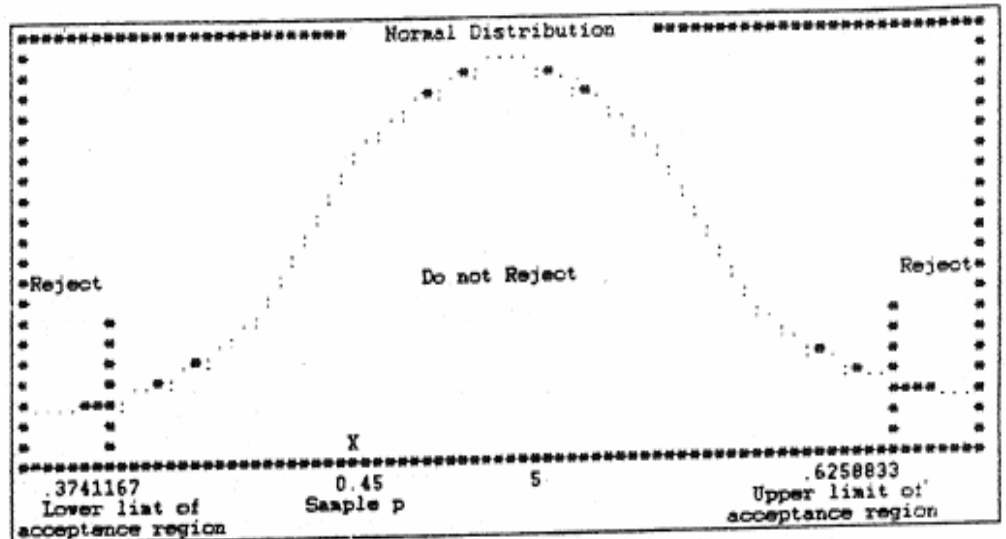Let us assume 5% level of significance.

Critical z value $=$ 1.96.

Lower limit of acceptance region $= \pi_{Ho} - Z_{\alpha/2}\,\sigma_p$

$=$ .50 - 1.96 (.0642)

$=$ .3741

Upper limit of acceptance region $= \pi_{Ho} + Z_{\alpha/2}\,\sigma_p$

$=$ .50 + 1.96 (.0642)

$=$ .6259

Since p lies in the upper and lower limit of acceptance region, we do not reject H_0 .This is shown graphically as follows.

47

The fourth question may be answered by testing the following hypothesis.

H₀(Null-hypothesis):	Average income of all the respondents equals Rs. 4500 ($\mu = 4500$).
H₁(Alternative - hypothesis):	The average income of all the respondents is different from Rs.4500 ($\mu \neq 4500$).

We use data given in column eight of Table 1 to test this hypothesis at 5% level of significance. The sample size being large (n = 60), we may use z test for the purpose.

$$M = l_i + \frac{\left(\dfrac{n}{2}\right) - CF}{f} \times (i)$$

where M = Median

† = Lower limit of the median class

n = Total number of frequencies in the distribution

CF=Cumulative frequencies for the class immediately below the class containing the medi

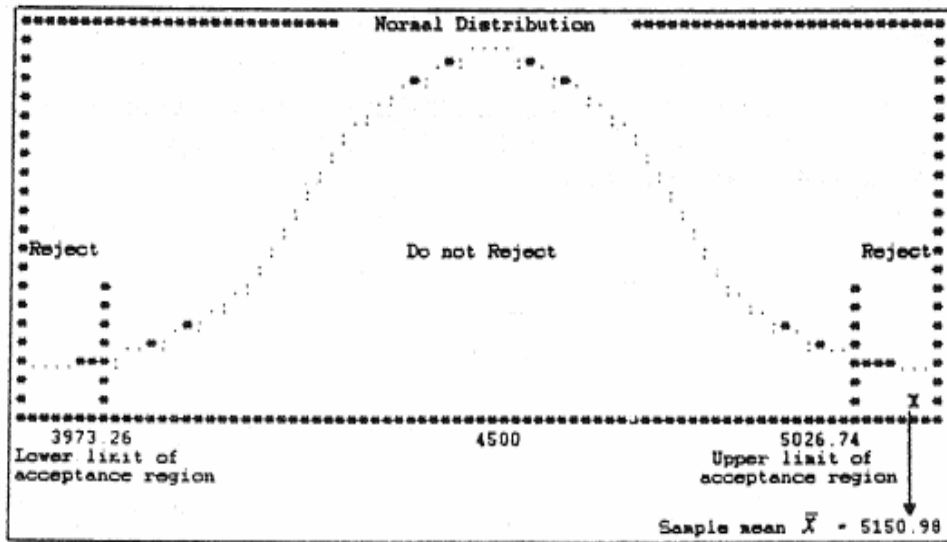Size of the interval of the median   f   Frequency of median class

Since $\overline{x}$ lies in the acceptance region we accept the null-hypothesis. This is shown graphically as follows:

48

```
********************************    Normal Distribution    ********************************
*                                        .  . . .                                        *
*                                       .*:     :*.                                       *
*                                      .*:         :*.                                    *
*                                     .:               :.                                 *
*                                    .:                   :.                              *
*                                   .:                      :.                            *
*Reject                            .:        Do not Reject    :.                   Reject*
*      .                          .:                            :.                       *
*     .                          .:                               :.                   . *
*    .              .           .:                                 :.  .*. .            .*
*  . . . .***  :  .*. .*:                                            : .*. .*           .*
*             .                                                       :.                .*
*            .                                                          ****...         .*
*           .                                                              .            .X*
**************************************************************************************************
    3973.26                              4500                        5026.74
Lower limit of                                                  Upper limit of
acceptance region                                              acceptance region

                                                      Sample mean X̄ = 5150.98
```

Activity 5

From past records it is known that the average life of a battery used in a digital clock is 305 days. The lives of the batteries are normally distributed. The battery was recently modified to last longer. A sample of 20 of the modified batteries was tested. It was discovered that the mean life was 311 days, and the sample standard deviation was 12 days. At the 0.05 level of significance, did the modification increase the mean life of the battery?

....
....
....
....
....

Activity 6

A company markets car tyres. Their lives are normally distributed with a mean of 40,000 kilometres and standard deviation of 3,000 kilometers. A change in the production process is believed to result in a better product. A test sample of 64 new tyres has a mean life of 41,200 km. Can you conclude that there is no significant difference between the new product and current one. You may use a 5% level of significance.

....
....
....
....
....
....

Activity 7

A high-speed automatic machine mass produces a small washer. Past experience reveals that 70 per cent of each day's production is perfect. Most of the remaining washers have a rough burr, which must be filed off before the washers can be inserted in the assembly. In an attempt to increase the per cent of production that is perfect, the machine was modified somewhat. A sample of 100 washers was then checked, and it was found that 72 percent were perfect. The boss thinks that there has been no change. The plant manager, however, believes that the production of the modified machine has definitely improved product quality.

49

That is, the per cent of perfect washers is greater than 70 percent. Is the plant manager correct? Test at the 0.02 level.

…………………………………………………………………………………
…………………………………………………………………………………
…………………………………………………………………………………
…………………………………………………………………………………

Activity 8

A product is classified as 1st, 2nd, 3rd, 4th grade. The past performance of the plant shows that the respective proportions are 8: 4: 2: 1. As a check on the run of the plant, 600 parts were examined and classified as follows:

Grade	:	1	2	3	4	Tota
Frequency	:	340	130	100	30	600

Is there any evidence of a change in the production standard? You may use 1% level of significance.

…………………………………………………………………………………
…………………………………………………………………………………
…………………………………………………………………………………

10.6 SUMMARY

In this unit we have distinguished between univariate, bivariate and multi-variate analysis of data. An attempt is also made to explain the difference between descriptive and inferential statistic. The various summary measures of descriptive statistics and a number of technique pertaining to inferential statistic concerning one population are discussed with the help of survey data. The descriptive measures are classified into two groups viz. measures of central tendency and measures of dispersion. The measures of central tendency include arithmetic mean, median, and mode. The measures of dispersion include range, standard deviation, coefficient of variation etc. Under inferential techniques, we have discussed the test procedures for inferring a specified value of population mean and proportion. A discussion is also made about when to apply t or z test while testing the population mean. The uses of univariate tables are also discussed.

10.7 SELF-ASSESSMENT QUESTIONS

1. Determine a statistical hypothesis and perform a chi-square test on the following survey data.

    a) Magazines are more interesting than television.

Agree	38
Neutral	46
Disagree	36

    2) Demographic characteristics of a group indicate

Education	Number
Post-graduate	42
Graduate	30
Below graduation	60

2. The answers to a researcher's question will be nominally scaled. What statistical test is appropriate to compare the sample data with hypothesized population data?

3. In a survey, respondents were asked to respond to a statement "Computer is very important for my work". Interpret the frequency distribution given below:

Category Label	Code	Absolute Frequency	Relative Frequency (Per cent)	Cumulative Frequency (Per cent)
Strongly disagree	1	110	24.44	24.44
Neutral	2	90	20.00	44.44
Strongly agree	3	250	55.56	100.00
Total		450		

4. The manager of a departmental store hypothesized the number of customers entering the store alone is 1.5 times the number entering in groups. A sample of 240 customers yielded the following results.

        Entering Alone 140

        Entering in Group   100

Is the manager's hypothesis about population of customers correct?

5. What is the basic use of a chi-square goodness of fit test'? How is the value of the test statistic calculated? How are the expected frequencies determined?

6. What is the appropriate test statistic for making inferences about a population mean when the variance is known? When the population variance is unknown? Suppose the population variance is unknown, but the sample size is large, what is the appropriate procedure then?

7. The number of demands for a particular spare part in a shop was found to vary from day to day. In a sample study the following information was obtained:

Days	:	Mon	The	Wed	Thu	Fri	Sat
No. of parts demanded	:	124	125	110	120	126	115

Test the hypothesis that the number of parts demanded doesn't depend upon the day of the week. You may use 5% level of significance.

8. The records of the owner of a fleet of trucks revealed that the average life of a set of spark plugs is 22,100 miles, The distribution of the life of the plugs is near normally distributed. A spark plug manufacturer claimed that its plugs has an average life in excess of 22,100 miles. The fleet owner purchased a large number of sets. A sample of the life of 18 sets revealed that the sample average life was 23,400 miles; the sample standard deviation was 1,500 miles. Is there enough evidence to accept the manufacturer's claim at the 0.05 level?

9. A food processor is concerned that the 16-ounce can of sliced pineapple is being overfilled. The quality control department took a random sample of 50 cans and found that the arithmetic mean weight was 16.05 ounces, with a sample standard deviation of 0.03 ounces. At the 5 per cent level of significance, can the hypothesis that the mean weight is equal to 16 ounces be rejected?

10. In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers? You may use 5% level of significance.

10.8 FURTHER READINGS

Beri, G. C. "Marketing Research - Text and Cases " Tata McGraw-Hill Publishing Co. Ltd. (1st Edition).

Green, Paul E. and Donald S. Tull "Research for Marketing Decisions" Prentice-Hall of India Pvt. Ltd. (4th Edition).

Kinnear, Thomas C. and James R. Taylog, 'Marketing Research - An Applied *Approach"* McGraw-Hill International Editions (3rd Edition).

Luck, David J. and Ronald S. Rubin, "Marketing Research" Prentice-Hall of India Pvt. Ltd. (7th Edition).

Majumdar, Ramanuj "Marketing Research - Text, Applications and Case Studies" Wiley Eastern Ltd. (1st Edition).

Westfall, Boyd and Stasch. 'Marketing Research - Text and Cases" Richard D. Irwin. Inc. (6th Edition).

APPENDIX - 1

This appendix gives a brief review of measures of central tendency and dispersion. They have already been discussed in detail in Unit numbering 7 & 8 of the course on Quantitative Analysis for Managerial Applications (MS-8). You are therefore advised to go through it in case you have forgotten it.

Measures Of Central Tendency:

The principal measures of central tendency are arithmetic mean, the median and the mode.

- **Arithmetic Mean:** This measure of central tendency is used in case Of interval or ratio scale data. In case of grouped data, the arithmetic mean may be obtained as:

- $$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n}$$

  Where $\overline{X}$ = the sample mean

  f_i = frequency of the ith class

  X_i = Midpoint of the ith class

  K = Number of classes

  n = Total number of observations in the sample.

  The population mean (μ) may be computed as:

  $$\mu = \frac{\sum f_i X_i}{N}$$

  Where μ = The mean of population

  N = Number of observations in the population

- **Median:** It is a measure of central tendency of a distribution. It is that value in the distribution such that 50 per cent of the observations are below it and 50 per cent are above it. Assuming odd number of observations in the distribution, it is the middle value of the distribution when the data is arranged in the ascending order of magnitude. When the number of observations are even, median is the mean of two middle values. In case of grouped data, the median is obtained by using the following formula :

$$M = l_i + \frac{\left(\frac{n}{2}\right) - CF}{f} \times (i)$$

Where M = Median

† = Lower limit of the median class.

n = Total number of frequencies in the distribution

CF = Cumulative frequency for the class immediately below the class containing the median.

i = Size of the interval of the median

f = Frequency of median class

- **Mode:** It is usually determined when data is categorised into frequency distribution. Mode represents the category having the highest number of observations. The mode may be computed from the grouped data as follows:

$$M_o = l_i + \frac{d_1}{d_1 + d_2}(i)$$

Where M_o = Mode

l_1 = Lower limit of the modal class.

d_1 = Difference between the frequency of the modal class and the frequency of the class immediately preceding it in the distribution.

d_2 = Difference between the frequency of the modal class and frequency of the class immediately following it in the distribution.

I = size of the interval of the modal class.

Measures of Dispersion: Measures of central tendency alone are not sufficient as they give a very inadequate description of the sample data. This is because these measures are not concerned with the spread of the distributio a. To measure spread or dispersion of the distribution some other measures like range, mean deviation, standard deviation, variance and co-efficient of variation are used.

- **Range:** This is the simplest measure of dispersion used for interval or ratio scaled data. It is measured as the difference between the highest and the lowest value in the distribution.

- **Variance and Standard Deviation:** The variance may be defined as the average of squared deviations of the observation values from the mean of the distribution. Standard deviation is the positive square root of the variance The population standard deviation (σ) be computed as:

53

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{2}}$$

Where X_i = the value of ith observation

μ = population mean

N = Number of observations in the population

For computing standard deviation from the sampled data, following formula may be used.

$$S = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{n-1}}$$

Where $\overline{X}$ = Sample mean

N = Number of observations in the samples

- Co-efficient of Variation: It is a relative measure of dispersion to compare the variability of more than one distribution. This is computed for ratio - scaled data. The formula for its computation is as given below:

$$CV = \frac{\sigma}{\mu}(100) or \frac{s}{\overline{X}}(100)$$

	Where	Coefficient of variation
σ	=	Standard deviation of the population.
μ	=	Mean of the population
s	=	Standard deviation of the sample.
$\overline{X}$	=	Mean of the sample.

To compare variability of various distribution, the coefficient of variation corresponding to each distribution is computed and the one having the minimum value is considered to be the more homogeneous.

APPENDIX - 2

This appendix gives a brief review of the inferential techniques for inferring the parameters of a single population (for details refer Unit 15 of MS-8 Course).

1. **Test Of Mean (Single population):** On the basis of sample drawn from the population one needs to infer about the population parameter. The parameter of interest could be population mean. One needs to carry out an appropriate statistical test of significance for testing the hypothesis concerning population mean. One can consider two cases - sample size being large (n > 30) and small sample (n 30). If n is large the sampling distribution of sample means follows a normal distribution as per central limit theorem. Therefore one can go for Z test. However when sample size is small, assuming that the samples are drawn from a normal population one may have to choose between Z and t test depending upon whether population standard deviation (6) is known or not. In case a is known one should go for Z test otherwise t test is carried out.

Let us assume that we want to test the hypothesis that the population mean (p) takes a specified value against an appropriate alternative hypothesis. For the sake of simplicity let us assume a two tailed alternative. Therefore the null and alternative hypothesis can be written as given below.

| H_0 (Null hypothesis) | μ | $=$ | μ_0 |
| H_1 (Alternative hypothesis) | μ | $\neq$ | μ_0 |

Let us assume a large sample and the level of significance being equal to a.. The test procedure involves the following steps:

1. Finding the sample mean $\bar{x}$

2. Computing the standard error of mean $\left(\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \right)$

3. Finding the value of $Z_{\frac{\alpha}{2}}$ from the standard normal table.

4. Finding the upper and lower limit of the acceptance region as given below:

   Upper limit of acceptance region $= \mu_{Ho} + Z_{\alpha/2}\sigma_{\bar{x}}$

   Lower limit of acceptance region $= \mu_{Ho} - Z_{\alpha/2}\sigma_{\bar{x}}$

   where μ_{Ho} represents the value of 1.1 under the assumption that the null hypothesis is true.

5. Examining whether sample mean $\bar{x}$ lies in the upper and lower limit of the acceptance region. If it does, we accept the null hypothesis otherwise reject it.

In case of small sample drawn from normal population with unknown standard deviation s the same hypothesis can be tested with the help oft statistic. The procedure is outlined below.

1. Find the sample mean $\bar{x}$

2. Find an estimate of population variance. The sample variance can be used to estimate population variance. The formula for computing sample variance is given as

$$s^2 = \frac{1}{n-1}\sum(Xi - \overline{X})^2$$

   S^2 is used as an unbiased estimator of σ^2

3. The standard error of mean is estimated $\hat{\sigma}_{\bar{x}} = \dfrac{S}{\sqrt{n}}$ where ^{A} stands for estimates.

4. Find the value of $t_{\frac{\alpha}{2}}$ with n - 1 degrees of freedom from t distribution table.

5. Find the upper and lower limit of the acceptance region as given
    below:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$$

where μ_{Ho} represents the value of μ under the assumption that the
null hypothesis is true.

6. Examine whether sample mean lies in the upper and lower limit of
    the acceptance region. If it does, we accept the null hypothesis
    otherwise reject it.

In case of one tail test, the procedure remains same except that one computes either
upper or lower limit of the acceptance region depending upon whether alternative
hypothesis is given by $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$. For example, in case of Z test the
upper limit will be computed as

$$\mu_{H0} + Z_\alpha \sigma_{\bar{x}}$$

2. **Test of Proportion:** In order to test the hypothesis that population proportion
    (p) takes a specified value p_a against a two sided alternative, we test the
    following hypothesis.

    H_o (Null Hypothesis)   : $\pi = \pi_0$

    H_1 (Alternative Hypothesis)   : $\pi \neq \pi$

    Assuming large sample such that both NP & nq > 5 (q =1-p), we may use
    normal approximation to binomial distribution to test the above hypothesis. It
    can be shown that

    1. E(P) $= \pi$

    2. σ_p $= \sqrt{\frac{\pi(1-\pi)}{n}}$

    where p denotes sample proportion proportion and σ_p denotes standard error
    of sample proportion

    To test the null hypothesis we compute upper and lower limit of acceptance
    region as given below

    Upper limit of acceptance region   = $\pi H_o + Z_{\alpha/2} \sigma_p$
    Lower limit of acceptance region   = $\pi H_o - Z_{\alpha/2} \sigma_p$

    If sample proportion p lies in the lower and upper limit of the acceptance
    region we accept H_o otherwise reject.

    For a one tail test the procedure is similar to as explained in the, case of tests
    of mean.

3. **Test of Goodness of Fit:** We will explain how chi-square statistic can be

used to test whether samples are drawn from a uniform population. One can compute the expected frequencies under the assumption that the null hypothesis is true. The hypothesis may be stated as:

H_o (Null Hypothesis):   The sample is drawn from a population which follows a normal distribution.

H_l (Alternate Hypothesis):   The data does not belong to the uniform distribution.

One can compute chi-square statistic by using the following formula:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$$

Here chi-square has k-1 degrees of freedom since we are assuming that the data is categorized into k groups. For a given level of significance one compares the table value of chi square with k-1 degrees of freedom to the computed value of chi-square and if tabulated chi-square is less than the computed chi-square we accept the null-hypothesis of uniformity of distribution.