UNIT 11 ANALYSIS ASSOCIATION

Objectives

After studying this unit, you should be able to:

- define a contingency table
- interpret cross-tabulation tables
- discuss the application of chi-square test to examine the independence of two nominally scaled variables
- explain different measures to ascertain the strength of relationship between two nominal variables
- carry out appropriate tests of difference between two population parameters
- distinguish between correlation and regression and their use in measuring association between variables.

Structure

11.1 INTRODUCTION

We discussed the description and inference from statistical data by taking one variable at a time in unit number 10. Our concern in this unit is on bivariate analysis of data where we take the simultaneous relationship between two variables. We are concerned with the cross-tabulation of data and its interpretation, analysis of association by using chi-square test and correlation coefficient. The cause and effect relationship between two variables will be examined by using simple linear regression model. We would carry out tests of statistical significance on the samples drawn from bivariate population viz. tests for the differences of means & differences of proportions. The use of various techniques will be illustrated with the help of survey data.

11.2 CROSS TABULATION

A matrix display of the categories of two nominal scaled variables, containing frequency counts of number of subjects in each bivariate category is called cross-tabulation table or contingency table. For preparing cross tables to examine simultaneous relations between two variables, we will consider the data given in Table 1 of the last unit. For the sake of convenience, the data is reproduced below.

TABLE 1: SURVEY DATA ON PREFERENCE FOR NEW PAKAGING DESIGN

RESPONDENT NUMBER	PREFERENCE	PREFERENCE CODE	SEX	MARRITAL STATUS	AGE IN YEARS	EDUCATION CODE	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	[	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	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

Please note that the codes used in the table have the same meaning as explained in unit 10.

To interpret the results of cross tabulated data we will, as an illustration, try to answer the following questions:

1. Divide the sample into two groups: (a) those showing interesting the new design and (b) those who are either indifferent or not interested in the new design. Cross- tabulated these two groups along - with education

   (I) higher education- graduation and above and

   (II) lower education –below graduation.

2. Perform similar exercise to ascertain association between preference for new design and income level of respondents, taking the first income in the poor class; second and third taking levels in the middle class; and the fourth one in the upper class.

3. Perform similar exercise to ascertain association between geographical location and preference for new design as categorized in question.1.

4. Preparation a cross – tabulation for preference for new design versus age groups , when there are two age groups – older respondents (40 years and above) and younger respondents (below 40 years).

5. Prepare a cross-table for sex variable against preference for new design.

6. Also prepare cross-table for marital status variable against preference for new design.

Interpret all the above cross tabulations

Before we attempt to answer the above questions, it would be of help to identify dependent and independent variable. In the questions listed above the dependent variable is preference - indicated by interest shown by consumers in the new design. The independent variable varies from question to question. For example, it is education, income, geographical location, age, sex and marital status for question numbers 1, 2, 3, 4, 5 & 6 respectively.

Preference vs. Education: As indicated before, the sample is divided into two groups (a) those interested in the new design and (b) those who are either indifferent or not interested in the new design. This variable is cross tabulated alongwith education variable which is divided into two groups

(i) higher education – graduation and above &

(ii) lower education – below graduation.

The result of the cross – tabulation are presented in Table 2 below:

Table 2:CROSS TABULATION OF PREFERENCE WITH EDUCATION

Education preference	**Under graduation**	**Graduation & Above**	**Row Total**
Not Interested	20(79.6%)	13(38.2%)	33
Interested	6(20.4%)	21(61.8%)	27
Column Total	26(100%)	34(100%)	60

To interpret the results of the above table, a question which arises is - should we compute the percentages column wise or row wise? There is a general rule for computing percentages. The rule is that we should cast the percentages in the direction of causal variable. In Table 2 the causal variable is education whereas the dependent variable is preference. There are 26 consumers who are under-graduate and 20 of them are not interested in the new design. This means 76.9% of the undergraduates are not interested in the new design. We observe that there are 34 consumers whose qualification is graduation or above and 13 of them are not interested in the new design which accounts for 38.2% of the consumers with education - graduation and above. The results clearly indicates with the increase in education, the interest for the new design increases. Therefore we must cast the percentages column wise as shown in Table 2 above.

Preferences vs. Income: The income data is divided into three categories by taking the first level in the poor class; second and third level in the middle class; and the fourth level in the upper class. The preference data is categorized as already explained. The cross tabulated data for these variables is given in Table 3 below:

Table 3 : CROSS TABULATION OF PREFERENCE WITH INCOME

Preference / Income	Not Interested	Interested	Row Total
Poor Class	10(99.01%)	1(.09%)	11 (100%)
Middle Class	22(53.66%)	19(46.34%)	41 (100%)
Upper Class	1(12.5%)	7(87.5%)	8 (100%)
Column Total	33	27	60

There are 11 consumers in the income group categorized as poor class and out of which only 0.09% prefer the new design. In the middle income class, there are 41 consumers and 46.34% of them prefer the new design. There are 8 consumers in the upper income class and 87.5% of them have shown preference for the new design. The results clearly indicate that with increase in income, preference for the new design increases.

Preference vs. Geographical Location: In Table 1 we have data of consumers from northern, southern, eastern and western part of the country. When this data was cross-tabulated against the preference variable the results as indicated in Table 4 appears.

Table 4 : CROSS TABULATION OF PREFERENCE WITH GEOGRAPHICAL LOCATION

Preference / Region	Not Interested	Interested	Row Total
North	10 (66.67%)	5(33.33%)	15(100%)
South	l0 (62.5%)	6(37.5%)	16(100%)
East	7 (46.67%)	8(53.33%)	15(100%)
West	6 (42.86%a)	8(57.14%)	14(100%)
Column Total	33	27	60

There are 15, 16, 15 and 14 consumers from northern, southern, eastern and western regions respectively. There are 66.67% of the consumers from northern region, 62.5% from southern region, 46.67% from eastern region and 42.86% from western region are not interested in the new design. This shows that a higher proportion of customer from eastern and western part of the country prefer the new design than from northern and southern part of the country. However the difference in the preference does not seem to

be significantly very large. To test for the statistical significance of the association between these two variables, one needs to carry out chi-square test which will be explained in the next section.

Preference Vs. Age Group: The data on age of the respondents given in Table 1 is divided into two groups - older respondents (40 years and above) and younger respondents (below 40 years). This variable is cross-tabulated against preference variable, the results of which are given in Table 5 below:

Table 5: CROSS TABULATION OF PREFERENCE WITH AGE GROUP

Age Group / Preference	Younger Respondents	Older Respondents	Row Total
Not Interested	20 (52.63%)	13 (59%)	33
Interested	18 (47.37%)	9 (41%)	27
Column Total	38 (100%)	22 (100%)	60

The table indicates that there are 38 younger and 22 older respondents. Out of the 38 younger respondents, 47.37% are interested in the new design. However in case of olders respondents only 41 % have shown an interest in favour of new design. The results indicate that the preference for the new design decreases with the age of the consumers. A plausible reason for this could be that the old customers have got used to the old design and therefore are showing resistance towards the new design.

Preference vs. Sex: In Table 1 the respondents were classified according to their sex. This variable when cross tabulated against preference variable shows the following results.

Table 6: CROSS TABULATION OF PREFERENCE WITH SEX

Preference	Not Interested	Interested	Row Total
Female	15 (55.6%)	12 (44.4%)	27 (100%)
Male	18 (54.55%)	15 (45.45%)	33 (100%)
Column Total	33	27	60

There are 27 female and 33 male respondents in the sample. Out of the 27 female respondents, 44.4% have shown an inclination towards the new design whereas 45.45% of the male respondent are in favour of new design. The difference between the proportion of male and female respondents preferring the new design does not seem to he large and therefore may be treated as insignificant.

Preference vs. Marital Status: As indicated in Table 1 the respondents were divided into two categories - Married & Single. This variable when cross-tabulated against preference variable shows the following results.

Table 7: CROSS TABULATION OF PREFERENCE WITH MARITAL STATUS

Preference / Marital Status	Not Interested	Interested	Rosy Total
Married	18 (40%)	12 (60%)	30 (100%)
Single	15 (50%)	15 (50%)	30 (100%)
Column Total	33	27	60

The sample comprises of equal number of married and single respondents. There are 40% of married respondents and 50% of single respondents showing interest in favour of new design. This is evident from the percentages casted row wise in the above table.

Activity 1

Suppose you have prepared a cross-tabulation table between two variables viz. family size and food expenditure. How will you go about interpreting any numbers that appear in such a table?

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

Activity 2

A survey was conducted in a locality to estimate the consumption expenditure of house- holds. A sample of 150 households was taken and data on a number of variales was collected. It was found that 50, 30 and 20 percent of households spend less than Rs.2000, between Rs.2000 and 4500; and greater than Rs.4500 respectively on food items per month. One of the cross-tabulations of the survey data is shown below:

Monthly Income (Rs.)

	< 5000	**5000 - 9000**	**> 9000**	**Total**
< 2000	45%	30%	25%	**100%**
2000 - 4500	30%	40%	30%	**100%**
> 4500	10%	40%	50%	**100%**

Have the percentages been cast in the right direction? If not, why? Recast the percent-ages in right direction and interpret the results.

……
……
……
……
……

11.3 CHI-SQUARE TEST FOR ANALYSIS OF ASSOCIATION

In the previous section we discussed the association between two cross-tabulated variables by computing percentages in the direction of causal variable. However the same analysis can be carried out by using chi-square tests of independence between two variables which are categorized into two or more groups. This may be examined by testing the following hypothesis:

H_0 (Null hypothesis): The two variables are not related,

H_1 (Alternative hypothesis): The two variables are dependent.

To test the hypothesis, one applies chi-square test of independence The observed frequencies are obtained from the survey data whereas the corresponding expected frequencies are computed under the assumption that the null hypothesis . is true. Corresponding to a contingency table (e.g. Tables 2 to 6) the expected frequencies for

any (*ij*) cell of the tables is found as

Where Eij = expected frequency corresponding .to the cell in ith row and jth column.

Ri = Total of observed frequencies cor[r]esponding to ith row.

Cj = Total of observed frequencies corresponding to jth column.

G = Grand total of frequencies.

The chi-square statistic is computed by using the following formula: 2

$$x^2_{(r-1)(c-1)} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where Oij = Observed frequency of the cell in ith row and jth column.
Eij = Expected frequency of the cell in ith row and jth column.

and (r-i)(c-i) indicates the degrees of freedom where r stands for the number of rows and c for number of columns.

For any given level of significance , the computed chi-square value is compared with the tabulated chi-square value. In case computed chi-square is greater than the tabulated chi-square, we reject the null-hypothesis to conclude that the variables are dependent.

We will test the following sets of hypothesis.

1. HO. Preference for the new design is independent of education.
   Hl. Preference and education are related.
2. HO. Preference for the new design is not related to the income level.
   HI. Preference for the new design is dependent on income.
3. HO. Preference for new design is independent of geographical location.
   Hl. Preference for new design is related to geographical location.
4. HO. Preference for new design is not related to the age.
   Hl. Preference for new design is dependent on age.
5. HO. Preference for new design is independent of sex of the respondent.
   Hl. Preference for new design is &pendent on sex of the respondent.
6. HO. Preference for new design is independent of marital status of the respondent.
   H1. Preference for new design is dependent on marital status of the respondent.

The tables of observed and expected frequencies corresponding to the above sets of hypothesis are given below (see Tables 8 to 13).

Table 8: OBSERVED & EXPECTED FREQUENCIES

Preference	**Education**		
	Under - graduation	**Graduation & above**	**Row Total**
Not Interested	20(14.3)	13(18.7)	33
Interested	6(11.7)	21(15.3)	27
Column Total	26	34	60

Note :Figures in the parenthesis() represent expected frequencies.

Table 9: OBSERVED & EXPEC I ED FREQUENCIES			
Preference		**Preference**	
	Not Interested	**Interested**	**Row Total**
Poor Class	10 (6.05)	1 (4.95)	11
Middle Class	22 (22.55)	19 (18.45)	41
Upper Class	1 (4.40)	7(3.60)	8
Column Total	33	27	60

Note: Figures in the parenthesis ( ) represent expected frequencies.

Table 10: OBSERVED & EXPECTED FREQUENCIES			
Preference	**Preference**		
	Not Interested	**Interested**	**Row Total**
North	10(8.25)	5(6.75)	15
South	10(8.8)	6(7.2)	16
East	7(8.25)	8(6.75)	15
West	6(7.70)	8(6.30)	14
Column Total	33	27	60

Note: Figures in the parenthesis ( ) represent expected frequencies.

Table 11: OBSERVED & EXPECTED FREQUENCIES			
Preference	**Age Group**		
	Younger Respondents	**Older Respondents**	**Row Total**
Not Interested	20 (20.9)	13 (12.1)	33
Interested	18 (17.1)	9 (9.9)	27
Column Total	38	22	60

Note: Figures in the parenthesis ( ) represent expected frequencies.

Table 12: OBSERVED & EXPECTED FREQUENCIES			
Sex	**Preference**		
	Not Interested	**Interested**	**Row Total**
Female	15(14.85)	12(12.15)	27
Male	18(18.15)	15(14.85)	33
Column Total	33	27	60

Note: Figures in the parenthesis ( ) represent expected frequencies.

Table 13: OBSERVED & EXPECTED FREQUENCIES			
	Preference		
Marital Status	**Not Interested**	**Interested**	**Row Total**
Married	18 (16.5)	12 (13.5)	30
Single	15 (16.5)	15 (13.5)	30
Column Total	33	27	60

Note: Figures in the parenthesis ( ) represent expected frequencies.

The computed chi-square value, tabulated chi-square value, the degrees of freedom and the decision for acceptance or rejection of Hypotheses numbering 1 to 6 is summar summarized in the Table 14.

Table 14: CO MUTED & TABULATED VALUES OF CBI-SQUARE, DEGREES OF FREEDOM & DECISION REGARDING VARIOUS HYPOTHESES

Hypothesis Number	Chi-square computed	Degrees of freedom (r-1) (c-1)	Chi-square tabulated	Decision
1	08.910	1	3.8415	Reject H_o
2	11.600	2	5.9915	Reject H_o
3	02.440	3	7.8147	Do not reject H_o
5	00.006	1	3.8415	Do not reject H_o
6	00.006	1	3.8415	Do not reject H_0

By examining the results of above table we may conclude to reject Hypothesis number 1 and 2 whereas we do not have enough evidence to reject Hypotheses numbering 3, 4, 5 & 6. This shows that preference for new design is related to education and income level. However, preference for new design is independent of geographical location, age, sex and marital status. One precaution which may be taken while computing chi-square statistic is that the observed frequency in each cell should be atleast 5. However, we have ignored this, as the objective was merely to show the computation for chi-square test. Whenever the frequencies in any cell is less than 5, it is advised to re-define the groups in such a way that the frequency for each cell is atleast 5.

Activity 3

A marketing manager was given the cross-tabulated data in the following table to show the association between age and preference for foreign brand of shirts. Use chi-square test to examine an appropriate hypothesis. What conclusions should be drawn?

	Age		
Preference	**Under 40**	**40 and over**	**Total**
Interested	80	30	110
Not Interested	20	60	80
Total	100	90	190

11.4 STRENGTH OF ASSOCIATION BETWEEN TWO NOMINAL VARIABLES

Having found that there is a relationship between variables, it would be interesting to examine the strength of association between two variables. There are three methods available to determine the degree of association. They are outlined as below.

1. **Contingency Coefficient:** When the null hypothesis of independence between two nominal variables is rejected, we may desire to measure the degree of assocaition between them. One of such measure called contingency coefficient (C) is given as:

$$c = \sqrt{\frac{x^2}{n + x^2}}$$

Where x^2 is computed by the method discussed in previous section and n is the total size of the sample. We know that the computed value of x^2 cor^{r}esponding to hypothesis number 1 is 8.91. The corresponding contingency coefficient (C) is computed as:

$$c = \sqrt{\frac{8.91}{60 + 8.91}} = 0.36$$

The question which may be asked at this stage is whether a value of 0.36 for the contingency coefficient indicates a strong or weak association. It has to be compared against its limits. The lowest value of limit is zero (C = 0) when x2 = 0 and which is possible when there is absolutely no association between variables. However for a perfectly correlated situation, this value .can not be unity as is the case for linear correlation coefficient (see Unit 17 of MS-8 course). This makes the task of interpreting the value of C difficult.

However, for a contingency table, where the number of rows and columns are equal, the upper limit of C is given by $\sqrt{\frac{(r-1)}{r}}$, where r is the number of rows or columns. Since the contingency table corresponding to hypothesis number 1 has two rows and two columns, the maximum possible value of C is

$$\sqrt{\frac{1}{2}} = 0.707$$

This suggests there is a moderate association between two variables since the computed value of the contingency coefficient (0.36) is halfway between the limits of zero and the upper limit of 0.707.

As already explained, it is not possible to Measure meaningfully the contingency coefficient when the number of rows and columns in the table are different. Another measure called cramer's V-statistic has been proposed under such a situaion. The expression for such a statistic is given by

$$V = \sqrt{\frac{x^2}{n(f-1)}}$$

where x^2 and n are defined as before and f is the samller of the number of rows and columns in a contingency table, that is, f = min (r, c).

It may be noted that the maximum value that the computed x^2 can take is n (f-1). Therefore V = 0, when there is lack of any association between variables and V = 1 when there is a perfect relationship. The value of V corresponding to hypothesis number 2 (where there are three rows and two columns in the contingency table) is given as

$$V = \sqrt{\frac{x^2}{n(f-1)}} = \sqrt{\frac{11.60}{60 \times 1}} = 0.44$$

This shows a moderate relationship between preference and the income level of the respondents.

2. **Phi Correlation Coefficient**: This is used to measure strength of association for two nominal variables in a contingency table of order 2 x 2 and is termed as phi-coefficient (4)). This phi-coefficient like correlation coefficient -can assume any value between -1 and 1. Further, ^{42} (the square of phi-coefficient) measures the proportion of one variable that is explained by the other variable.

Let us consider 2x2 cross-tabulation of Table 8 corresponding to Hypothesis number 1. We wish to examine the strength of association between preference for new design and education level. For the sake of convenience, the data is reproduced below in Table 15.

Table 15: PREFERENCE VS. EDUCATION

Preference	Education		Row Total
	Under-graduation	Graduation & Above	
Not interested	20 (a)	13 (b)	33 (a + b)
Interested	6 (c)	21 (d)	27 (c + d)
Column total	26 (a + c)	34 (b+d)	60 (a+b+c+d)

Phi-coefficient (4)) may be computed by using the following formula:

$$\varnothing = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$$= \frac{(20)(21) - (13)(6)}{\sqrt{(33)(27)(26)(34)}}$$

$$= \frac{342}{\sqrt{787644}}$$

$$= \frac{342}{887.49} = 0.385$$

The following description of the strength of relationship for a given particular phi value may be used.

Value of ± ∅	Strength of Relationship
Greater than 0.80	Strong
0.40 to 0.80	Moderate
0.20 to 0.40	Weak
0.00 to 0.20	Negligible

Source: Marketing Research by Luck & Rubin p-503.

According to the above description, there is a weak association between preference and education. The value of $\varnothing^2 = 0.15$, indicates that 15% of the variations in preference is explained by education level.

The phi-coefficient may assume positive or negative value. However, sign of [j] does not have any particular meaning. If the responses were concentrated in cells c & b instead of a and d, the sign of phi would have been negative.

It may be noted that when total sample size increases, the computing formula for phi may become tedious. . In such a situation, the phi-coefficient may be computed as

$$\varnothing = \sqrt{\frac{x^2}{n}}$$

3. **Goodman and Kruskal's Lambda:** The two measures namely contingency coefficient and phi are the symmetrical measures of association in the sense that they do not predict which variable predicts the other. However, there can be instances where a marketing resarcher may desire to determine the direction of prediction. To meet such a requirement Goodman and Kruskal's Lambda - asymmetric coefficient may be used. This method measures the extent to which we may reduce the error in predicting categories of one variable from the knowledge of the categories of other variable.

   Let us use the cross-tabulation given in Table 2. Suppose we wish to predict preference from education level. We would need to rearrange Table 2 which may appear as:

Table 16: CROSS TABULATION OF PREFERENCE WITH EDUCATION

Preference / Education	Not Interested	Interested	Row Total
Undergraduation	20	6	26
Graduation	13	21	s4
Column Total	33	27	60

$$\lambda \text{a s y m} = \frac{\sum_{i=1}^{R} f_{ir} - F_e}{n - F_e}$$

where fir = the maximum, frequency found within each category of the row variable.

Fe = the maximum frequency among the marginal totals of the column variable

n = sample size

R = number of categories in row variable

$\sum_{i=1}^{R} f_{ir} = 20 + 21 = 41$

$F_e = 33$

$n = 60$

$$\lambda asym = \frac{41-33}{60-33} = 0.259$$

If we want to compute λa s y m for the case when we want to predict education from preference, we would need to use Table 2 as our data input. We could compute

$\sum_{i=1}^{R} f_{ir} = 20 + 21 = 41$

$F_e = 34$

$n = 60$

Therefore, λ asym for predicting education from the preference is

$$\lambda asym = \frac{41-34}{60-34} = 0.259$$

The result indicates that depending on which variable predicts another variable, we will have different values of lamda.

If we are only interested in measuring the association (and not direction) between two variables, we may compute 2 symmetrical as given by

$$\lambda sym = \frac{\sum_{i=1}^{R} f_{ir} + \sum_{i=1}^{c} f_{ie} - (F_r + F_c)}{2n - (F_r + F_c)}$$

Where fic = the maximum frequency found within each category of the column.

Fr = the maximum frequency among the marginal totals of the row variable.

C = number of categories in a column.

Applying the formula to the data in Table 2, we get

$\sum fir = 20+21 = 41$

$\sum fic = 20+21 = 41$

$F_r = 33$

$F_e = 34$

$$\lambda asym = \frac{41+41-(33+34)}{120-(33+34)} = \frac{15}{53} = 0.28$$

The term 2, sym is used to represent an average predictability between the two variables of interest. Both the forms of lambda can be tested for statistical significance. However their discussion is beyond the scope of this unit.

Activity 4

State whether the following statements are True (T) or False (F).

S.No.	Statement	True/False
l:	The contingency coefficient may be used to determine the direction of prediction.	
2.	The upper limit of contingency coefficient is unity.	
3.	Phi-coefficient is computed for contingency tables of order 2x2.	
4-	Phi-coefficient is a symmetrical measure of association between two vriables.	
5.	Goodman and Kruskal's lambda may also be used for measuring degree of association between variables.	

Activity 5

List out the limitations of contingency coefficient in interpreting the results of association between nominal variables.

……
……
……
……...

11.5 CORRELATION COEFFICIENT

The degree of association between two variables is also computed using correlation coefficient. A detailed treatment of correlation coefficient was taken up in Unit no. 18 of course (MS-8) Quantitative Analysis for Managerial Application. We know that the correlation coefficient may be positive, negative or zero. A measure of linear correlation coefficient between two variables X & Y is measured by correlation coefficient, the formula of which is given below:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 (Y - \overline{Y})^2}}$$

Where $\overline{X}$ = Sample mean for the variable X

$\overline{Y}$ = Sample mean for the variable Y

n = Number of observations in the sample

This measures can assume any value between -1 and +1 and is independent of units of measurements. The correlation coefficient can be computed for interval and ratio scale data. If we examine the data in Table 1, we find that preference data indicated in column 2, age data in column 6 and acutal income data in column 8 can be utilised to compute correlation coefficient. For the sake of illustration let us compute: the correlation coefficient between preference and actual income using the above mentioned formula. It is -found that the correlation coefficient between preference and income is equal to .4597 indicating a positive relationship between these two variables. Further the magnitude of correlation coefficient suggests that there is a moderate relationship between preference and income

To test the significance of population correlation coefficient (^{0}) we may use t test described below.

H_0(Null Hypothesis) : There is no correlation between preference and income (P = 0)

H_1(Alternative Hypothesis) : There is correlation between preference and income ($\rho \neq 0$).

The test statistic used to test the hypothesis is given by:

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where r = Sample correlation coefficient

n = Sample size

The significance of correlation coefficient was tested by using above formula and the computations are as given below.

$$t = \frac{0.46\sqrt{58}}{\sqrt{1-.2116}}$$

$$= \frac{0.46 \times 7.616}{0.8879}$$

$$= \frac{3.50336}{0.8879} = 3.9459$$

For a given level of' significance = .05 (say), the table value of t is obtained as 1.96. This value being less than the computed value of t (3.9459), we reject the hypothesis of no correlation between the preference and actual income.

Activity 6

What are the differences in two measures viz. phi-coefficient and correlation coefficient in measuring the strength of association between two variables.

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

Activity 7

Consider the data given in Table 1. Use the preference data in column 2 and age data in column 6 of the table to find if there is any correlation coefficient between these variables. Also test for its significance.

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

11.6 SIMPLE LINEAR REGRESSION

One of the limitations of correlation analysis is that it only measures the degree of association between two variables and does not deal with cause and effect relationship between variables. Regression analysis overcomes this problem by assuming a causal relationship between variables (regression analysis only assumes cause and effect relationship and does not establish it.) Assuming Y is a function of X, we are treating X as a causal variable and Y as a dependent variable. This means it is on' the variable X

which influences the variable Y and not the other way round. We have taken a very simplistic case of a situation where the dependent variable Y is explained only by one independent (causal variable) X. In real life situation there may be a host of variables influencing the dependent variable Y. The discussion of such a case comes under multiple regression model and will be taken up in Unit no. 12 of this course. Here our concern is with simple linear regression model where there is only one dependent and one independent variable. Assuming a linear relationship between the dependent and the independent variable, the relationship between Y and X may be written as

$$\mathbf{Y = b_0 + b_1X + u}$$

Where b_o = intercept term

b_1 = slope term

u = random error

The estimation of b_o and b_1 is carried out by using ordinary least squares method of estimation (for details on this method, please consult Unit no. 19 of MS - 8 Course.) The ordinary least squares method (OLS) 'aims at minimizing the error sum of squares to obtain the estimated value of b_o and b_1. The estimated values are obtained using the following formulae:

Where $\overline{X}$ = Sample mean of the variable X

Y = Sample mean of the variable

$\hat{b}_0, \hat{b}_1$ = Estimates of population parameters b_o and b_1

We have taken preference as the dependent variable and income as an independent variable and have run a simple regression of preference on income. The results are given below:

Y 1.79 +0.00027X

t (3.94) (4.68)

The above estimated regression equation indicates that as the income increases by one rupee the preference reading goes up by .00027 points. Further the positive sign of the slope coefficient indicates that the relationship is positive. The significance of the slope coefficient is tested with the help of t statistic. Here the computed value of t is 4.68 and if we assume 5% level of significance and two tailed alternative, the tabulated value of t would be 1.96. It is seen that the computed t is greater than the tabulated t indicating that the coefficient corresponding to income variable is significant at 5% level.

The goodness of fit of the regression equation is estimated by r^2, the coefficient of determination. The value of r^2 is always between 0 and 1 and higher the value of r^{-2} better is the goodness of fit. The value of r^2, also represents the explanatory power of the regression model. If the value of r^2 equals .84 (say), it means that 84% of the variations in the dependent variable are explained by the independent variable. In the case of our example the value of r^{-2} equals .21 which means 21% of the variation in the preference variable are explained by income variable. (We have very briefly touched upon this section. For details on regression analysis please refer to Unit no.19 of MS-8 Course.)

Activity 8

Consider the data on preference (column 2) and *age (column 6) variables as given in Table 1. Compute the simple regression of preference for new design on age and carry out appropriate statistical tests of significance. Also interpret your results.

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

11.7 ANALYSIS OF DIFFERENCES

In this section we would like to answer some of the questions such as:

1. Is there any difference between the monthly income of respondents who are interested in the new design from those who are not interested in the new design?

2. Is there any difference in the proportion of female and male respondents preferring the new design?

The above questions may be answered by testing the following hypotheses:

1.	H_0 (Null Hypothesis)	:	There is no difference between the monthly income of those who are interested in the new design and those not interested in the new design. ($\mu_1 = \mu_2$)
	H_1 (Alternative Hypothesis)	:	The monthly income of the respondents preferring the new design is not equal to those not preferring the new design ($\mu_1 \neq \mu_2$)
2.	H_0 (Null Hypothesis)	:	The proportion of females preferring the new design is equal to the proportion of males preferring the new design. ($\pi_1 = \pi_2$)
	H_1 (Alternative Hypothesis)	:	The proportion of females preferring the new design is not equal to the proportion of males preferring the new design. ($\pi_1 \neq \pi_2$)

where μ_1 = Population mean income of those preferring the new design

μ_2 = population mean income of those not preferring the new design

μ_3 = proportion of females in the population preferring the new design

μ_4 = proportion of males in the population preferring the new design

Tests of Difference between Means

To test the difference between two population means one needs to apply either t or z test. Appendix 1 of this unit discusses the situation where each of these tests are applicable. In our present case, there are 27 respondents preferring the new design (n_1=27) and 33 respondents (n_2=33) not preferring the new design. The variable of interest is income. For testing the Hypothesis number 1 we have computed the following:

$\overline{x_1}$	=	6,218.6665
$\overline{x_2}$	=	4,2774243
s_i	=	1,885.9080
s_2	=	1,830.6273
Standard Error of difference between Means $\left(\sigma_{x_1-x_2}\right)$	=	481.5306
Computed t		4.03
Critical t with 58 d.f. (assuming 1% level of significance)		2.576

Since computed t is greater than tabulated t, we reject the null hypothesis and conclude that the monthly income of respondents preferring the new design is different from those not preferring the new design.

Tests of Difference Between Proportion

The difference between two population proportions may be tested using z test, assuming normal approximation to binomial distribution. Appendix 1 of this unit discusses the test briefly. For testing Hypothesis 2, we have 27 females ($n_1 = 27$) and 33 male respondents ($n_2 = 33$). The variable of interest is the proportion of respondents preferring the new design.In our case we have the following Computed information

Proportion of females preferring the new design (p_1)	=	0.4444
Proportion of males preferring the new desing (p_2)	=	0.4545
Estimated standard error of difference between proportion ($_pa_{...p}$)	=	0.1314
Computed z value	=	0.077
Critical z value at 5% level of significance		1.96

Since computed z value is less than the critical z value we accept the null hypothesis of no difference between two population proportion. Therefore the proportion of females preferring the new designs is not significantly different from the proportion of males preferring the new design.

Activity 9

Consider the data on relevant variables to test the hypothesis that there is no difference in the average age of those who prefer the new design and the one who do not. YOU may use 5% level of significance.

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

11.8 SUMMARY

In this unit, our concern was with bivariate analysis of data. We examined the simultaneous relationship between two variables at a time. We discussed ,haw to interpret a cross-tabulation table by computing percentages. The rule devised was to cast the

percentages in the direction of causal variable while interpreting cross tabulation tables. We also discussed chi-square test for examining the independence between two nominally scaled variables. The strength of association between two variables was studied using three methods namely Contingency Coefficient, Phi Correlation Coefficient and Goodman & Kruskal's Lamda. The limitation of contingency coefficient is that it is not possible to measure meaningfully the contingency coefficient when the number of rows and columns in the contingency table are not equal. Another statistic called Cramer's V-Statistic is used in such a situation. The phi -coefficient is used for 2 x 2 table and can take values between -1 & +1. Goodman & Kruskal's Lamda are of two types viz. λ asym & λ sym. The first one is used to determine the direction of prediction whereas the second may be used for measuring association between two nominal variables.

The degree of two association between two variables is also computed with the help of simple correlation coefficient which may take any value between -1 & +1. However, the correlation coefficient does not deal with cause and effect relationship and to overcome this regression analysis is used. We have discussed the estimation and interpretation of regression results.

The analysis of difference between two population means can be tested with the help of t or z statistic. We have discussed the situations uncle- which these tests are applicable. The difference between two proportions may be tested with the help of z statistic.

All the above methods are illustrated with the help of survey data exhibited in Table 1 of this unit.

11.9 SELF-ASSESSMENT QUESTIONS

1. What are the ways in which percentages may be computed in a cross-tabulation table? which way is the best?
2. In a contingency table, what type of hypothesis is tested using chi-square test? What precautions may be taken while applying the test?
3. What are the various measures used to 'measure the strength of association between two nominal variables? Describe them and clearly mention their limitations, if any.
4. Discuss the problem of using the contingency coefficient in intepreting the results of correlation analysis?
5. Explain the difference between correlation & regression..
6. A marketing researcher interested in the business publication reading habits of purchasing agents has assembled the following data in the form of cross-tabulation alongwith the information whether each agent holds a degree or not.

	First Choice Mentions		
Business Publication	**Degree**	**No degree**	**Total**
W	20	15	35
X	15	15	30
Y	25	20	45
Z	30	25	55
Total	90	75	165

a) Is there an association between business-publication choice and type of college degree?

b) What is the appropriate null hypothesis for this illustration?

c) Use an appropriate method to measure the strength of association between two variables.

7. A marketing research analyst has one nominally scaled variable and one interval and Analysis scaled variable.The analyst wishes to use a chi-square test to examine the association between two variables.Give your suggestion as to what can be done.

8. What type of scalar data (Nominal, Ordinal, Interval or Ratio) is typically utilized in cross tabulation• analysis?

9. A new breakfast cereal is being test marketed in selected cities on the east and west coasts.Consumer panels are being used for the evaluation in each of the selected cities, and after four weeks of product use, the consumer reactions have been obtained as follows:

	East coast	**West coast**
Total responses	632	428
Preferred cereal or considered	468	327
cereal equivalent to others		

We are interested in finding out whether or not there' are regional differences in consumer acceptance of the product. Conduct an appropriate test on the results two samples at I% level of significance.

10. Two batches of the same product are tested for their mean life. Assuming that the lives of the product follow a normal distribution with an unknown variance, test the hypothesis that the mean life is the same for both the batches, given the following information:

Batch	**Sample Size** (X)	**Mean life in hrs.** (s)	**Standard deviation**
I	10	750	12
II	8	820	14

Choose a significance level of 5%.

11. An automobile manufacturing firm is bringing out a new model. In order to map out its advertising campaign, it wants to determine whether the model will appeal most to a particular age-group or equally to all age-groups.

The firm takes a random sample from persons attending a preview of the new model and obtained the results summarised below:

Age Groups

Persons	**Under 20**	**20 - 39**	**40 - 59**	**60 and over**
Liking the car	146	78	48	28
Disliking the	54	32	32	62

What conclusion would you draw from the above data? You may use 1% level of significance.

12. A study was conducted to know whether there is any difference between the monthly income of the probationary officers in public sector banks to that of private sector banks. Some sample results are given below:

Sample Statistic	**Public Sector Banks**	**Private Sector Banks**
Sample size	45	60
Sample mean	Rs.9,500	Rs.10,800
Sample variance	Rs.438	Rs.725

Test the hypothesis that the income of probationary officers in private sector banks is higher than that of public sector banks using 5% level of significance.

11.10 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

We would briefly review the statistical tests of differences between two means & proportions. It would be of help if you go through Unit no. 15 of the course on Quantitative Analysis for Managerial Applications (MS-8).

Test of two means: We want to test whether or not the observed differences between two sample means is significant. In other words - are population means really different? To examine this hypothesis we may have to make a choice between Z and t test as was done in the case of univariate analysis. We choose Z test when standard deviations are known for both population or when sample sizes are sufficiently large (i.e., $n_1 > 30$ and $n_2 > 30$). The use of t test is made when population standard deviation for either population is unknown or when one or both samples are small (i.e. $n_r < 30$ or $n_2 < 30$).

Z test for difference between two means

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ (Assuming a two-tailed alternative)

Test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

Where $\bar{X}1$ = sample mean for the first variable

$\bar{X}2$ = sample mean for the second variable

μ_1 & μ_2 = hypothesized population means for two variables

$\sigma_{\bar{X}_1 - \bar{X}_2}$ = standard error of the difference between means

The standard error of difference between means may be computed as

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where σ_1^2 *and* σ_2^2 are the population variances for the two variables and n_1 and n_2 are the respective sample sizes.

For a given level of significance α, the corresponding Z value is obtained from standard

normal tables and is compared with computed Z value. If computed Z value (in absolute term) exceeds corresponding tabulated Z value, we reject the null hypothesis (H0) of no difference between two population means.

In case σ_1 and σ_2 are unknown but both n_1 and n_2 are greater than 30, we may still use Z test. However s_1 and s_2 (sample standard deviaiton) are used to estimate σ_1 and σ_2 respectively and the estimate of standard error of difference between means is obtained as

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2$$

and $$s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2$$

The t test on difference between two means when one or both sample sizes are small (i.e. $n_1 < 30$ or $n_2 < 30$), a t-test would be appropriate. For using t-test following assumptions are required

i) Independent samples are drawn from two normal populations.

ii) Population variances are unknown but equal.

The hypothesis may be stated as:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ (Assuming a two-tailed alternative)

The test statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $n_1 + n_2 - 2$ degrees of freedom, where sp is an estimate of pooled standard deviation, given by

$$sp = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

For a given level of significance , the corresponding t value with n_1+n_2-2 degrees of freedom is obtained from t-distribution table. If computed t (absolute value) exceeds tabulated t, we reject the null hypothesis.

Test of two proportions: Suppose we are interested in testing the hypothesis involving the equality of two population proportions and the two sample sizes n_1 and n_2 from the populations are large enough so as to have each of the quantities n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ atleast 5, we may use normal approximation to binomial distribution to apply Z test. The hypothesis to be tested is:

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 \neq \pi_2$ (Assuming a two-tailed test)

The appropriate test statistic is

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_{p_1 - p_2}}$$

Where p_1, p_2 = sample proportions

π_1, π_2 = hypothesized population proportion

$\sigma_{p_1 - p_2}$ = population standard error for the difference between proportions

An estimate of the population standard error for the difference between proportions is given by

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\hat{q} = 1 - p$$

For a given level of significance α, the value of Z is obtained from standard normal table. If computed value of Z (absolute value) is greater than the corresponding tabulated value, we reject H_0.