UNIT 4 CHI-SQUARE AND KENDALL RANK CORRELATION

Structure

4.0 INTRODUCTION

In this unit, we will be discussing about the issues relating to the association and relationship between two or more variables. Generally when we want to measure the linear relationship between two variables, we apply Product Moment Coefficient of Correlation to the data and compute the 'r' value and check for its significance. This again we would do so if the data is normally distributed and the measurement of scores etc. are atleast in interval scale and there is a large sample. However if the sample size is small, and the distribution of the data is not known and the measurement is in nominal or ordinal scale, then we use non-parametric statistics related correlation, as for example the Rho or the Kendall Tau or where we need to know the association between two variables we may use the chi square test. In this unit we will be presenting first the measures of correlation both in parametric and non-parametric statistics, followed by Kendall rank order correlation, the Spearman Rank order correlation and the Chi Square test.

4.1 OBJECTIVES

On completing this unit, you will be able to:

- Define parametric and non-parametric tests of correlation;
- Explain the concepts underlying the non-parametric correlations;
- Describe the different non-parametric correlation techniques;
- Enumerate the step by step calculation of Kendall Tau; and
- Enumerate the step by step calculation of Chi Square test.

4.2 CONCEPT OF CORRELATION

The term “correlation” refers to a process for establishing whether or not relationships exist between two variables. Correlation quantifies the extent to which two quantitative variables, X and Y, “go together.” When high values of X are associated with high values of Y, a positive correlation exists. When high values of X are associated with low values of Y, a negative correlation exists. If values of X increases bringing about an increase in the values of Y simultaneously, X and Y are said to be positively correlated. If increases in X values bring about comparative decreases in the values of Y, then X and Y are said to be negatively correlated. If there is no typical trend in the increase or decrease of the variables then it is said to be not correlated or having zero correlation. Correlation ranges from -1 to 0 to +1. Correlation of +1.00 will indicate a perfect positive correlation and -1 will indicate a perfect negative correlation. Between these two extremes there could be many other degrees of correlation indicating positive or negative relationship between the variables. The correlation cannot exceed 1 in either direction. But it can have 0.54, 0.82, or 0.24, or 0.63 and so on at the positive level and at the negative level , it can have -0.55, -0.98, -0.67, -0.27 etc. All the latter are negative correlations and will not go beyond -1.00. Similarly the correlations that were mentioned as positive , will not exceed +1.00.

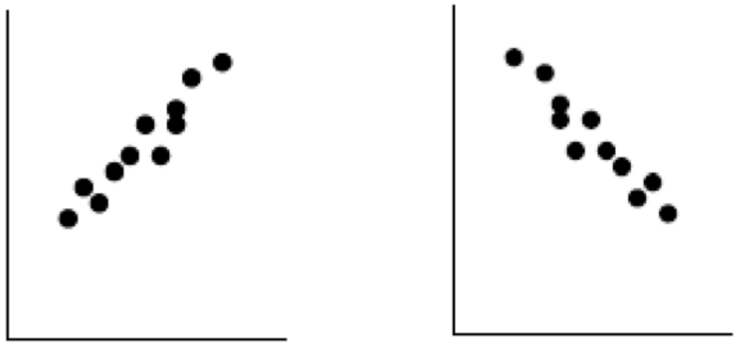
4.2.1 Scatter Plot

The first step is creating a scatter plot of the data. “There is no excuse for failing to plot and look.”

In general, scatter plots may reveal a

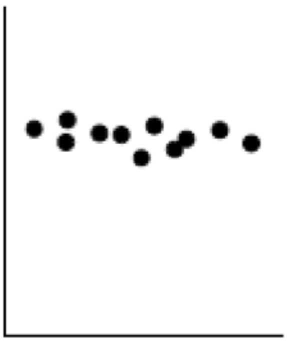
- positive correlation (high values of X associated with high values of Y)
- negative correlation (high values of X associated with low values of Y)
- no correlation (values of X are not at all predictive of values of Y).

These patterns are demonstrated in the figure below

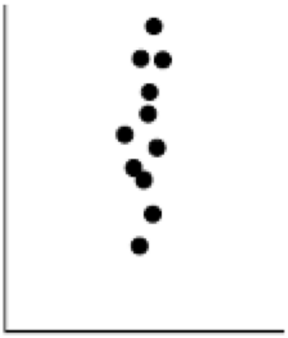


(A) Positive Correlation

(B) Negative Correlation



(A) No Correlation



(B) No Correlation

Correlation Coefficient

A single summary number that gives you a good idea about how closely one variable is related to another variable

This summary answers the following questions:

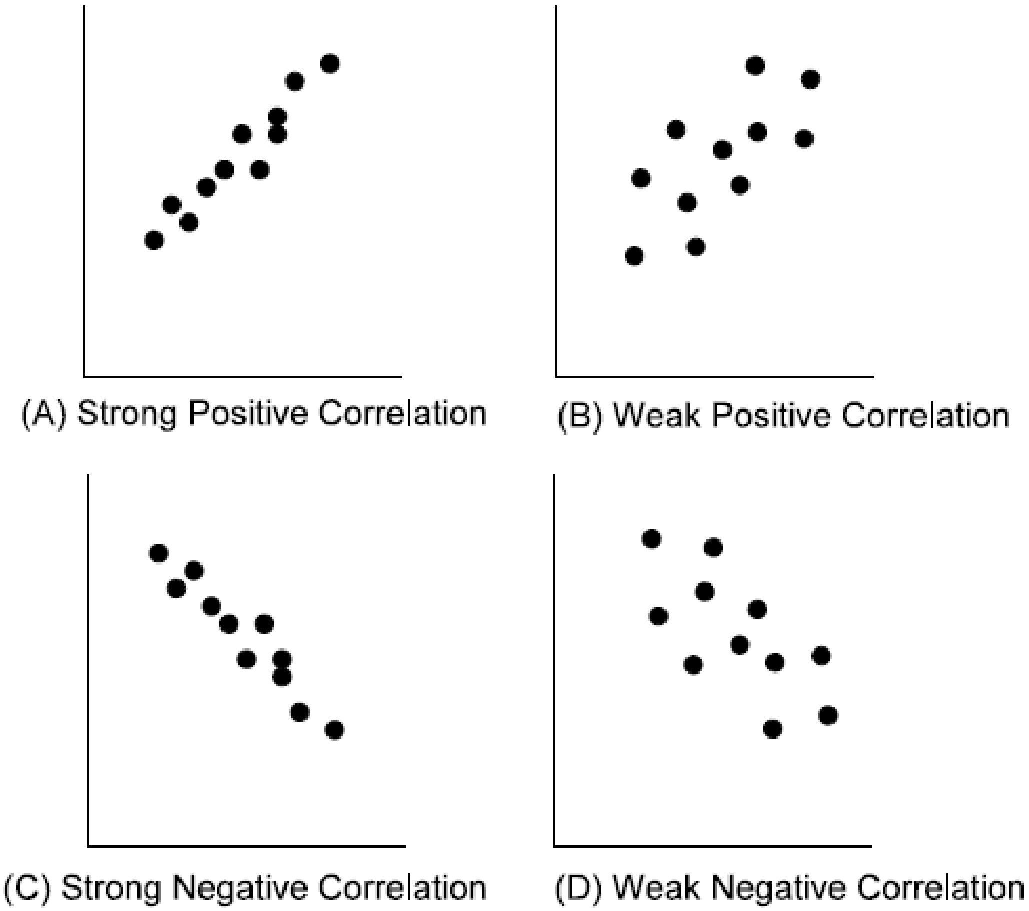
a) Does a relationship exist?

b) If so, is it a positive or a negative relationship? and

c) Is it a strong or a weak relationship?

Additionally, the same summary number would allow us to make accurate predictions about one variable when we have knowledge about the other variable.

Correlation coefficients (denoted by *r*) are statistics that quantify the relationship between X and Y in unit free terms. When all points of a scatter plot fall directly on a line with an upward incline, *r* = +1.00, but when all points fall directly on a downward incline, *r* =



(A) Strong Positive Correlation

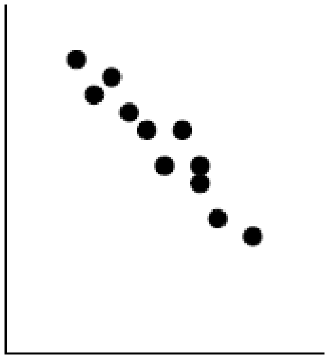
(B) Weak Positive Correlation

(C) Strong Negative Correlation

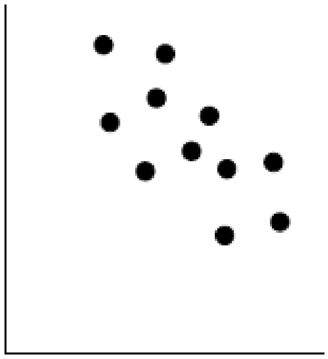
(D) Weak Negative Correlation

(A) Strong Positive Correlation

(B) Weak Positive Correlation



(A) Strong Negative Correlation



(B) Weak Negative Correlation

It is seen from the above that the strong correlation at both positive and negative directions is almost in a line with all the dots are placed very close to each other. On the other hand, the weak positive or negative correlation (refer to the graph above on the right hand side) that the points are placed far away from each other though the direction is somewhat clear. Thus there is a correlation but it appears rather weak.

4.2.2 Characteristics of Correlation

1) They tell you the direction of the relationship between two variables.

If your correlation coefficient is a negative number you can tell, just by looking at it, that there is a negative relationship between the two variables. As you may recall from the last chapter, a negative relationship means that as values on one variable increases (go up) the values on the other variable tend to decrease (go down) in a predictable manner.

If your correlation coefficient is a positive number, then you know that you have a positive relationship. This means that as one variable increases (or decreases) the values of the other variable tend to go in the same direction. If one increases, so does the other. If one decreases, so does the other in a predictable manner.

2) Correlation Coefficients always fall Between -1.00 and +1.00

All correlation coefficients range from -1.00 to +1.00. A correlation coefficient of -1.00 tells you that there is a *perfect negative relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *decrease* in values on the other variable. In other words, as one variable goes up, the other goes in the opposite direction (it goes down).

A correlation coefficient of +1.00 tells you that there is a *perfect positive relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *increase* in values on the other variable. In other words, as one variable goes up, so does the other.

A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can't really say anything about what happens to the other variable. Sometimes the other variable goes up and sometimes it goes down. However, these changes are not predictable.

3) Larger Correlation Coefficients Mean Stronger Relationships

Most correlation coefficients (assuming there really is a relationship between the two variables you are examining) tend to be somewhat lower than plus or minus 1.00 (meaning that they are not perfect relationships) but are somewhat above 0.00. Remember that a correlation coefficient of 0.00 means that there is no relationship between the two variables based on the data given .

The closer a correlation coefficient is to 0.00, the weaker is the relationship and the less able one is to tell exactly what happens to one variable based on the knowledge of the other variable. The closer a correlation coefficient approaches plus or minus 1.00 the stronger the relationship is and the more accurately you are able to predict what happens to one variable based on the knowledge you have of the other variable.

4.3 MEASURES OF CORRELATION

4.3.1 Parametric Statistics

a) Pearson product moment correlation coefficient (Most widely accepted as a single appropriate statistics for correlation)

4.3.2 Non-parametric Statistics

a) Spearman's rank order correlation coefficient: Better known as "Spearman Rho" (Siegel & Castellan, 1988) assumes that the variables under consideration were measured on at least an ordinal (rank order) scale, that is, that the individual observations can be ranked into two ordered series. Spearman R can be thought of as the regular Pearson product moment correlation coefficient, that is, in terms of proportion of variability accounted for, except that Spearman R is computed from ranks.

b) Kendall's Tau: Explained in section 4.3.

c) Chi Square (Categorical Variables): Explained in section 4.7

Self Assessment Questions

1) Fill in the blanks:

i) Scatter plots may reveal a ___________correlation (high values of X associated with high values of Y)

ii) Scatter plots may reveal a ___________correlation (high values of X associated with low values of Y)

iii) Scatter plots may reveal ________correlation (values of X are not at all predictive of values of Y).

iv) Correlation coefficients range from ______ to ______

v) A correlation coefficient of _____ tells you that there is a *perfect positive relationship* between the two variables.

vi) The closer a correlation coefficient is to 0.00, the __________ the relationship

vii) Correlation coefficient is a single summary number that gives you a good idea about *how closely one variable is* ___________ *to another variable*

2) What questions does correlation coefficient answers?

....

....

....

....

3) Name any two methods for calculating correlation?

....

....

....

....

4.4 KENDALL'S RANK ORDER CORRELATION (KENDALL'S TAU): (ð)

Kendall's tau (**ð**) is one of a number of measures of correlation or association. Measures of correlation are not inferential statistical tests, but are, instead, descriptive statistical measures which represent the degree of relationship between two or more variables. Upon computing a measure of correlation, it is a common practice to employ one or more inferential statistical tests in order to evaluate one or more hypotheses concerning the correlation coefficient. The hypothesis stated below is the most commonly evaluated hypothesis for Kendall's tau.

Null Hypothesis

H_o: **ð** $= 0$

(In the underlying population the sample represents, the correlation between the ranks of subjects on Variable X and Variable Y equals 0.)

4.4.1 Relevant Background Information on Test

Prior to reading the material in this section the reader should review the general discussion of correlation, of the Pearson product moment correlation coefficient and Spearman's rank order correlation coefficient (which also evaluates whether a monotonic relationship exists between two sets of ranks). Developed by Kendall (1938), tau is a bivariate measure of correlation/association that is employed with rank-order data. The population parameter estimated by the correlation coefficient will be represented by the notation **ð** (which is the lower case Greek letter tau). As is the case with Spearman's rank-order correlation coefficient Rho (***ñ***), Kendall's tau can be employed to evaluate data in which a researcher has scores for n subjects/objects on two variables (designated as the X and Y variables), both of which are rank-ordered.

Kendall's tau is also commonly employed to evaluate the degree of agreement between the rankings of m = 2 judges for n subjects/objects. As is the case with Spearman's rho, the range of possible values Kendall's tau can assume is defined by the limits - 1 to +1 (i.e., - 1 < r > +1). Although Kendall's tau and Spearman's rho share certain properties in common with one another, they employ a different logic with respect to how they evaluate the degree of association between two variables.

Kendall's tau measures the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of subject/objects.

One set of ranks represents the ranks on the X variable, and the other set represents the ranks on the Y variable.

Specifically, The data are in the form of the following two pairs of observations expressed in a rank-order format:

a) $(R_x, R_y,)$ (which, respectively, represent the ranks on Variables X and Y for the 1st subject/object); and

b) $(R_{xj}, R_{yj},)$ (which, respectively, represent the ranks on Variables X and Y for the jth subject/object).

If the sign/direction of the difference $(R\ x_i – R\ _{yj})$, that is a pair of ranks is said to be concordant (i.e. in agreement).

If the sign/direction of the difference $(R_{xi} – R_{xj})$, a pair of ranks is said to be discordant (i.e., disagree).

If $(R_{yi} – R_{yj})$ and/or $(R_{xi} – R_{xj})$ result in the value of zero, a pair of ranks is neither the concordant nor discordant.

Kendall’s tau is a proportion which represents the difference between the proportions of concordant pairs of ranks less the proportion of discordant pairs of ranks.

The computed value of tau will equal + 1 when there is complete agreement among the rankings (i.e., all of the pairs of ranks are concordant), and will equal -1 when there is complete disagreement among the rankings (i.e., all of the pairs of ranks are discordant).

As a result of the different logic involved in computing Kendall’s tau and Spearman’s rho, the two measures have different underlying scales, and, because of this, it is not possible to determine the exact value of one measure if the value of the other measure is known.

In spite of the differences between Kendall’s tau and Spearman’s rho, the two statistics employ the same amount of information, and, because of this, it is equally likely to detect a significant effect in a population.

In contrast to Kendall’s tau, Spearman’s rho is more commonly discussed in statistics books as a bivariate measure of correlation for ranked data. Two reasons for this are as follows:

a) The computations required for computing tau are more tedious than those required for computing rho; and

b) When a sample is derived from a bivariate normal distribution.

4.5 STEP BY STEP PROCEDURE FOR KENDALL RANK-ORDER CORRELATION

These are the steps in use of the Kendall rank order Correlation coefficient **ð(tau)**:

Rank the observations on the X variable from 1 to N. Rank the observations on the Y variable from 1 to N.

Arrange the list of N subjects so that the rank of the subjects on variable X are in their natural order, that is, 1, 2, 3,….N.

Observe the Y ranks in the order in which they occur when X ranks are in natural order. Determine the value of S, the number of agreements in order minus the number of disagreements in order, for the observed order of the Y ranks.

If there are no ties among either the X or the Y observations then we use the formula:

$T = 2S / (N (N -1))$

Where:

S = (score of agreement – score of disagreement on X and Y)

N = Number of objects or individuals ranked on both X and Y

If there are ties then the formula would be:

$T= 2S / [\quad N (N-1) – T_x \qquad T= 2S / [\quad N (N-1) – Ty$

Where:

S and N are as above

$T_x = \Sigma\, t (t – 1)$, t being the number of tied observations in each group of the ties on the X variable

$T_y = \Sigma\, t (t – 1)$, t being the number of tied observation in each group of the ties on the Y variable

If the N subjects constitute a random sample from some population, one may test the hypothesis that the variable X and Y are independent in that population. The method for doing so depends on the size of N:

For $N \leq 10$, Table — Upper tail probabilities for T, the Kendall rank order correlation coefficient

For N > 10, but less than 30, Table – Critical value for T, the Kendall rank order correlation coefficient

For N < 30 (or for intermediate significance levels for $10 < N \leq 30$) compute the value of z associated with T by using formula given below and use the z table

$z = 3T \quad N (N – 1) \ / \ \ 2 (2N+5)$

If the probability yielded by the appropriate method is equal to or less than the critical value, null hypothesis may be rejected in the favour of alternative hypothesis.

Worked up Example:

Without Ties:

Suppose we ask X and Y to rate their preference for four objects and give points out of 10. Now to see whether their preferences are related to each other we may use the following steps:

Data:

	A	B	C	D
X	6	8	5	2
Y	8	4	9	6

Step 1: Ranking the data of X and Y

	A	B	C	D
X	3	4	2	1
Y	3	1	4	2

Step 2: Rearrange the data of X in order of 1 to N (4 in this case)

	D	C	A	B
X	1	2	3	4

Step 3: Put the corresponding score of Y in order of X and Determine number of agreements and disagreements

	D	C	A	B
X	1	2	3	4
Y	2	4	3	1

To calculate S we need number of agreements and disagreements. This can be calculated by

Using the Y scores, starting from left and counting the number of ranks to its right that are larger, these are agreements in order. We subtract from this the number of ranks to its right that are smaller- these are the disagreements in order. If we do this for all the ranks and then sum the results we obtain S:

Y	2	4	3	1	Total
	2	+	+	-	+1
		4	-	-	-2
			3	-	-1
				1	0
				Grand Total= S	- 2

Step 4: Calculate T

T = 2S / (N (N -1))

T = 2 (– 2) / (4 (4 – 1))

T = – 4 / 12

T= – 0.33

Thus, T = – 0.33 is a measure of the agreement between the preferences of X and Y.

With Ties:

The two set of ranks to be correlated are:

Subject	A	B	C	D	E	F	G	H	I	J	K	L
Status striving rank	3	4	2	1	8	11	10	6	7	12	5	9
Yielding rank	1.5	1.5	3.5	3.5	5	6	7	8	9	10.5	10.5	12

As usual we would first rearrange X and observe the scores of corresponding Y scores to calculate S

Subject	D	C	A	B	K	H	I	E	L	G	F	J	
Status striving rank	1	2	3	4	5	6	7	8	9	10	11	12	
Yielding rank	3.5	3.5	1.5	1.5	10.5	8	9	5	12	7	6	10.5	Total
	3.5	0	-	-	+	+	+	+	+	+	+	+	8
		3.5	-	-	+	+	+	+	+	+	+	+	8
			1.5	0	+	+	+	+	+	+	+	+	8
				1.5	+	+	+	+	+	+	+	+	8
					10.5	-	-	-	+	-	-	0	-4
						8	+	-	+	-	-	+	0
							9	-	+	-	-	+	-1
								5	+	+	+	+	4
									12	-	-	-	-3
										7	-	+	0
											6	+	1
												10.5	0
											S=	Grand Total	25

We compute the value of S in usual way

S = (8-2) + (8-2) + (8-0) + (8-0) + (1-5) +

(3-3) + (2-3) + (4-0) + (0-3) + (1-1) + (1-0) = 25

It should be noted that, when there are tied observations, the ranks will be tied and neither rank in comparison pair precedes the other, so a value of 0 is assigned in the computation of S.

Having determined that S = 25, we now determine the value of T_x and T_Y. There are no ties among the scores on social status striving, i.e. in the X ranks and thus $T_x = 0$

On Y scores there are three sets of tied ranks. Two subjects are tied at 1.5, two subjects at 3.5, and two subjects' at 10.5 ranks. In each of these cases T = 2, the number of tied observations. Thus may be computed as:

$T_Y = \Sigma\, t\,(t - 1)$

$=2\,(2–1) + 2(2–1) + 2(2–1)$

$= 6$

With $T_x = 0$, $T_Y = 6$, $S = 25$, and $N = 12$, we may determine the value of T by using formula:

$T= 2S / [\sqrt{N\,(N-1) - T_x}\ \sqrt{N\,(N-1) - T_y}\ ]$

$T= (2 \times 25) / \sqrt{12\left(12-1\right)-0}\ \sqrt{12\left(12-1\right)-6}$

$= 0.39$

If we had not corrected the above coefficient for ties, i.e. we had used the previous formula for computing T we would have found T = 0.38. Observe that the effect of correcting for ties is relatively small unless the proportion of tied ranks is large or the number of ties in a group of ties is large.

4.6 FURTHER CONSIDERATIONS ABOUT KENDALL'S TAU

4.6.1 Comparison of Rho and Tau

For the example of tied observation if one calculates r it will be 0.62, whereas the T is 0.39. This example illustrates the fact that T and r have different underlying scales, and numerically they are not directly comparable to each other. That is if we measure the degree of correlation between A and B by using r and then do the same for A and C by using T, we cannot then say whether A is more closely related to B or to C because we have used noncomparable measures of correlation. It should be noted, however that there is a relation between the two measures which is best expressed in the following inequality:

$-1 \leq 3\,T - 2r \leq 1$

There are also differences in interpretation of the two measures. The spearman rank order correlation coefficient rho ***(ñ)*** is the same as a Pearson product moment correlation coefficient computed between variables the values of which consists of ranks. On the other hand, the Kendall rank-order correlation coefficient**(ð=tau)** has a different interpretation. It is the difference between the probability that, in the observed data, X and Y are in the same order and the probability that the X and Y data are in different orders. T_{XY} is different in the relative frequencies in the sample.

However, both coefficients utilise the same amount of information in the data, and thus both have the same sensitivity to detect the existence of association in the population. That is, the sampling distributions of T and r are such that for a given set of data both will lead to rejection of the null hypothesis at the same level of significance. However it should be remembered that the measures are different and measure association in different ways.

4.6.2 Efficiency of Rho

The Spearman Rho ***(ñ)*** and The Kendall (tau=ð) are similar in their ability to reject H_o, inasmuch as they make similar use of the information in the data.

When used on data to which the Pearson product moment correlation coefficient r is properly applicable, both Rho (***(ñ)*** and tau (**ð**) have efficiency of 91 percent. That is, Rho is approximately as sensitive a test of independence of two variables in a bivariate normal population with a sample of 100 cases as the Pearson r with 91 cases (Moran,1).

Self Assessment Questions

1) Fill in the blanks:

   i) Rho and tau have different underlying scales, and numerically they are not __________ to each other.

   ii) Developed by _____________ in year _________, tau is a __________measure of correlation/association that is employed with rank-order data.

2) State true or false:

   i) Kendall's tau measures the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of subject/objects.

   ii) Kendall's tau and Spearman's rho, the two measures have different underlying scales, and, because of this, it is not possible to determine the exact value of one measure if the value of the other measure is known.

   iii) Kendall's tau and Pearson's r both are rank order correlation, therefore both can be compared.

4.7 CHI-SQUARE TEST

The chi-square (X^2) test measures the alignment between two sets of frequency measures. These must be categorical counts and *not* percentages *or* ratios measures (for these, use another correlation test).

Note that the frequency numbers should be significant and be at least above 5 (although an occasional lower figure may be possible, as long as they are not a part of a pattern of low figures).

Chi Square performs two types of functions:

1) **Goodness of fit**

A common use is to assess whether a measured/observed set of measures follows an expected pattern. The expected frequency may be determined from prior knowledge (such as a previous year's exam results) or by calculation of an average from the given data.

The null hypothesis, H_0 is that the two sets of measures are not significantly different.

2) **Measure of Independence**

The chi-square test can be used in the reverse manner to goodness of fit. If the two sets of measures are compared, then just as you can show they align, you can also determine if they do *not* align.

The null hypothesis here is that the two sets of measures are similar.

The main difference in goodness-of-fit vs. independence assessments is in the use of the Chi Square table. For goodness of fit, attention is on 0.05, 0.01 or 0.001 figures. For independence, it is on 0.95 or 0.99 figures (this is why the table has two ends to it).

4.8 RELEVANT BACKGROUND INFORMATION ON TEST

The chi-square goodness-of-fit test, also referred to as the chi-square test for a single sample, is employed in a hypothesis testing situation involving a single sample. Based on some pre existing characteristic or measure of performance, each of *n* observations (subjects/objects) that is randomly selected from a population consisting of N observations (subjects/objects) is assigned to one of k mutually exclusive categories.' The data are summarized in the form of a table consisting of k cells, each cell representing one of the k categories.

The experimental hypothesis evaluated with the chi-square goodness-of-fit test is whether or not there is a difference between the observed frequencies of the k cells and their expected frequencies (also referred to as the theoretical frequencies). The expected frequency of a cell is determined through the use of probability theory or is based on some pre existing empirical information about the variable under study. If the result of the chi-square goodness-of-fit test is significant, the researcher can conclude that in the underlying population represented by the sample there is a high likelihood that the observed frequency for at least one of the k cells is not equal to the expected frequency of the cell. It should be noted that, in actuality, the test statistic for the chi-square goodness-of-fit test provides an approximation of a binomially distributed variable (when $k = 2$) and a multinomially distributed variable (when $k > 2$). The larger the value of *n*, the more accurate the chi-square approximation of the binomial and multinomial distributions.

The chi-square goodness-of-fit test is based on the following assumptions:

a) Categorical nominal data are employed in the analysis. This assumption reflects the fact that the test data should represent frequencies for k mutually exclusive categories;

b) The data that are evaluated consists of a random sample of n independent observations. This assumption reflects the fact that each observation can only be represented once in the data; and

c) The expected frequency of each cell is 5 or greater.

When this assumption is violated, it is recommended that if $k = 2$, the binomial sign test for a single sample be employed to evaluate the data. When the expected frequency of one or more cells is less than 5 and $k > 2$, the multinomial distribution should be employed to evaluate the data. The reader should be aware of the fact that sources are not in agreement with respect to the minimum acceptable value for an expected frequency.

Many sources employ criteria suggested by Cochran (1952), who stated that none of the expected frequencies should be less than 1 and that no more than 20% of the expected frequencies should be less than 5. However, many sources suggest the latter criteria may be overly conservative. In the event that a researcher believes that one or more expected cell frequencies are too small, two or more cells can be combined with one another to increase the values of the expected frequencies.

Zar (1999) provides an interesting discussion on the issue of the lowest acceptable value for an expected frequency. Within the framework of his discussion, Zar (1999) cites studies indicating that when the chi-square goodness-of-fit test is employed to evaluate a hypothesis regarding a uniform distribution, the test is extremely robust.

A robust test is one that still provides reliable information, in spite of the fact that one or more of its assumptions have been violated. A uniform distribution (also referred to as a rectangular distribution) is one in which each of the possible values a variable can assume has an equal likelihood of occurring. In the case of an analysis involving the chi-square goodness-of-fit test, a distribution is uniform if each of the cells has the same expected frequency.

4.9 STEP BY STEP PROCEDURE FOR CHI-SQUARE TEST

1) Write the observed frequencies in column *O*

2) Figure the expected frequencies and write them in column *E*.

Expected Frequencies:

When you find the value for chi square, you determine whether the observed frequencies differ significantly from the expected frequencies. You find the expected frequencies for chi square in three ways:

1) You hypothesize that all the frequencies are equal in each category. For example, you might expect that half of the entering freshmen class of 200 at Tech College will be identified as women and half as men. You figure the expected frequency by dividing the number in the sample by the number of categories. In this exam pie, where there are 200 entering freshmen and two categories, male and female, you divide your sample of 200 by 2, the number of categories, to get 100 (expected frequencies) in each category.

2) You determine the expected frequencies on the basis of some prior knowledge. Let us use the Tech College example again, but this time pretend we have prior knowledge of the frequencies of men and women in each category from last year's entering class, when 60% of the freshmen were men and 40% were women. This year you might expect that 60% of the total would be men and 40% would be women. You find the expected frequencies by multiplying the sample size by each of the hypothesized population proportions. If the freshmen total were 200, you would expect 120 to be men (60% x 200) and 80 to be women (40% x 200).

3) Use the formula to find the chi-square value:

   Chi Square = $\Sigma\ [(O - E)^2 / E]$

   Where:

   O is the Observed Frequency in each category

   E is the Expected Frequency in the corresponding category

4) Find the *df*. (*N*-1)

5) Find the table value (consult the Chi Square Table.)

6) If your chi-square value is *equal to or greater than* the table value, reject the null hypothesis: *differences in your data are not due to chance alone*

Worked Up Example:

Situation: Mr. X., the manager of a car dealership, did not want to stock cars that were bought less frequently because of their unpopular color. The five colors that he ordered were red, yellow, green, blue, and white. According to Mr. X, the expected frequencies or number of customers choosing each color should follow the percentages of last year. She felt 20% would choose yellow, 30% would choose red, 10% would choose green, 10% would choose blue, and 30% would choose white. She now took a random sample of 150 customers and asked them their colour preferences. The results of this poll are shown in Table below under the column labelled as observed frequencies."

Category Color	Observed Frequencies	Expected Frequencies
Yellow	35	30
Red	50	45
Green	30	15
Blue	10	15
White	25	45

The expected frequencies in Table are figured from last year's percentages. Based on the percentages for last year, we would expect 20% to choose yellow. Figure the expected frequencies for yellow by taking 20% of the 150 customers, getting an expected frequency of 30 people for this category. For the colour red we would expect 30% out of 150 or 45 people to fall in this category.

Using this method, Thai figured out the expected frequencies 30, 45, *15*, *15*, and 45. Obviously, there are discrepancies between the colours preferred by customers in the poll taken by Mr.X. and the colours preferred by the customers who bought their cars *last* year. Most striking is the difference in the green and white colours. If Thai were to follow the results of her poll, she would stock twice as many green cars than if she were to follow the customer colour preference for green based on last year's sales. In the case of white cars, she would stock half as many this year. What to do? Mr. X. needs to know whether or not the discrepancies between last year's choices (expected frequencies) and this year's preferences on the basis of his poll (observed frequencies) demonstrate a *real* change in customer colour preferences. It could be that the differences are simply a result of the random sample she *chanced to* select. If so, then the population of customers really has not changed from last year as far as colour preferences go.

The *null hypothesis* states that there is no significant difference between the expected and observed frequencies.

The *alternative hypothesis* states they *are* different. The level of significance (the point at which you can say with 95% confidence that the difference is NOT due to chance alone) is set at .05 (the standard for most science experiments.) The chi-square formula used on these data is

Chi Square = $\Sigma\ [(O - E)^2 / E]$

Where:

O is the Observed Frequency in each category

E is the Expected Frequency in the corresponding category

df is the "degree of freedom" (n-1)

We are now ready to use our formula for X^2 and find out if there *is* a significant difference between the observed and expected frequencies for the customers in choosing cars. We will set up a worksheet; then you will follow the directions to form the columns and solve the formula.

1) *Directions for Setting up Worksheet for Chi Square*

Category	O	E	O-E	$(O-E)^2$	$(O-E)^2$ / E
Yellow	35	30	5	25	0.83
Red	50	45	5	25	0.56
Green	30	15	15	225	15
Blue	10	15	-5	25	1.67
White	25	45	-20	400	8.89
				Total=	26.95

This Total is the Chi Square value. After calculating the Chi Square value, find the *"Degrees of Freedom."*

(Remember: DO *NOT* SQUARE THE NUMBER YOU GET, NOR FIND THE SQUARE ROOT - THE NUMBER YOU GET FROM COMPLETING THE CALCULATIONS AS ABOVE IS CHI SQUARE.)

2) *Degrees of freedom (df)* refers to the number of values that are free to vary after restriction has been placed on the data. For instance, if you have four numbers with the restriction that their sum has to be 50, then three of these numbers can be anything, they are free to vary, but the fourth number *definitely* is restricted. For example, the first three numbers could be 15, 20, and 5, adding up to 40; then the fourth number has to be 10 in order that they sum to 50. The degrees of freedom for these values are then three. The degrees of freedom here is defined as *N* - 1, the number in the group minus one restriction (4 - 1).

3) Find the table value for Chi Square. Begin by finding the *df* found in step 2 along the left hand side of the table. Run your fingers across the proper row until you reach the predetermined level of significance (.05) at the column heading on the top of the table. The table value for Chi Square in the correct box of *4 df* and *P=.05* level of significance is 9.49.

4) If the calculated chi-square value for the set of data you are analysing (26.95) is equal to or greater than the table value (9.49), reject the null hypothesis. *There is a significant difference between the data sets that cannot be due to chance alone.* If the number you calculate is LESS than the number you find on the table, then you can probably say that any differences are due to chance alone.

In this situation, the rejection of the null hypothesis means that the differences between the expected frequencies (based upon last year's car sales) and the observed frequencies (based upon this year's poll taken by Mr.X) are not due to chance. That is, they are not due to chance variation in the sample Mr.X took. There is a real difference between them. Therefore, in deciding what colour autos to stock, it would be to Mr.X's advantage to pay careful attention to the results of her poll!

Another Example:

Let us take an example of Males and Females in two different categories, full stop and rolling stop and no stop. Now to see whether they are different from each other or more similar to each other we will follow the following steps

Step 1: Add numbers across columns and rows. Calculate total number in chart.

Unobtrusive Male Versus Female

	Male	Female	
Full Stop	6	6	= 12
Rolling Stop	16	15	= 31
No Stop	4	3	= 7
	= 26	= 24	= 50

Step 2: Calculate the expected numbers for each individual cell. Do this by multiplying row sum by column sum and dividing by total number. For example: using 1st cell in table (Male/Full Stop);

12 x 26 / 50 = 6.24

2nd cell in table (Female/Full Stop):

12 x 24 / 50 = 5.76

Step 3: Now you should have an observed number and expected number for each cell. The observed number is the number already in 1st chart. The expected number is the number found in the last step (step 2). Sometimes writing both numbers in the chart can be helpful

	Male	Female	
Full Stop	6 (observed) 6.24 (expected)	6 (observed) 5.76 (expected)	= 12
Rolling Stop	16 (observed) 16.12 (expected)	15 (observed) 14.88 (expected)	= 31
No Stop	4 (observed) 3.64 (expected)	3 (observed) 3.36 (expected)	= 7
	= 26	= 24	= 50

Step 4:

Chi Square = Sum of (Observed - Expected)^2 / Expected

Calculate this formula for each cell, one at a time. For example, cell #1 (Male/Full Stop):

Observed number is: 6 Expected number is: 6.24

Plugging this into the formula, you have:

$(6 – 6.24)^2 /6.24 = .0092$

Continue doing this for the rest of the cells, and add the final numbers for each cell together for the final Chi Square number. There are 6 total cells, so at the end you should be adding six numbers together for you final Chi Square number.

Step 5: Calculate degrees of freedom (*df*):

(Number of Rows – 1) x (Number of Columns – 1)

(3 – 1) x (2 – 1)

2 x 1 =

2 *df* (degrees of freedom)

Step 6: Look up the number in the chart at end of handout. At .05 significance level, with 2 *df*, the number in chart should be 5.99. Therefore, in order to reject the null hypothesis, the final answer to the Chi Square must be greater or equal to 5.99. The Chi Square/final answer found was .0952. This number is less than 5.99, so you fail to reject the null hypothesis, thus there is no difference in these groups.

4.10 FURTHER CONSIDERATIONS ABOUT CHI SQUARE

Observations must appear in one cell only. For instance, if we looked at male and female swimmers and hurdlers, one person could appear in both the swimmers *and* the hurdlers category if they enjoyed both sports. This would make use of Chi square invalid. Actual frequencies must appear in the cells, not percentages, proportions or numbers which do anything other than count. For instance, the mean of an interval scale variable cannot appear.

LOW expected frequencies

One limitation is that one should not proceed with a chi square test where expected frequency cells fall below 5. The rule of thumb which most statisticians inherited, and which comes from Cochran (1954) which was that *no more than 20% of expected cells should fall below* 5. This would rule out any 2 X 2 in which at least one expected cell was less than 5.

Hypothetical table:

Age	**Conversed**	**Did not converse**	**Total**
5 years	2	6	8
7 years	6	2	8
Total	8	8	16

For total sample sizes less than 20 and two expected cells below *5,* the risk of a type I error is too high. For instance, the data shown in hypothetical table above give a *chi square* of 4.0 (which is 'significant' for one *df)* yet it's easy to see, again, without much formal statistical training, that the result was relatively likely to occur - only two children in each age group needed to move away, in opposite directions, from the expected frequencies of four in each cell for these results to occur. From first principles (working out all the possible combinations) the probability of these results occurring comes out substantially higher than 0.05. If you have these sort of data it doesn't take too long to work from first principles but it's far better to make sure your analysis will be valid by taking a large enough sample, with a sensible design. Even with tables larger than 2X2, if several expected frequencies fall below 5 and the row or column total are quite severely skewed, the possibility of a type I error increases.

Self Assessment Questions

1) What are the assumptions of chi-square goodness-of-fit test?

....

....

....

....

2) Chi square performs two major functions, what are these?

....

....

....

....

3) State true or false:

i) The expected frequency of a cell is determined through the use of probability theory or is based on some pre existing empirical information about the variable under study.

ii) If several expected frequencies fall below 5, the possibility of a type II error increases.

iii) The chi-square (c^2) test measures the alignment between two sets of frequency measures.

iv) “The data that are evaluated consists of a random sample of n independent observations.” Is not a cardinal assumptions of chi square?

4.11 LET US SUM UP

In this unit we learnt about the concept of correlation and how parametric test is used to compute the product moment coefficient of correlation. We thedn learnt about the non-parametric tests for corrleation and leatrnt about the Rho and Tau. The Rho was by Spearman and was known as Spearman Rank Correlation while Kendall’s Tau was known as ð (tau). We also learnt about how to calculate Kendall’s tau and learnt about the importance of Chi-Square test. We also learnt as to how to calculate chi-square.

4.12 UNIT END QUESTIONS

1) Compute correlation coefficient for each of the following pairs of sample observations:

a)

x	33	61	20	19	40
y	26	36	65	25	35

b)

x	89	102	120	137	41
y	81	94	75	52	136

c)

x	2	15	4	10
y	11	2	15	21

d)

x	5	20	15	10	3
y	80	83	91	82	87

2) Compare T and r in terms of correlation and state your views?

3) Should a chi-Square test be carried out on the following data?

7	1
2	7

4) A (fictitious) Survey shows that. in a sample of 100.9 I people are against the privatisation of health services, whereas 9 support the idea.

   a) What test of significance can be performed on this data?

   b) Calculate the chi square value and check it for significance.

   c) Could this test be one-tailed?

   If for a large sample, we knew *on/y* that 87% of people were against the idea and were for could we carry out the same test to see whether this split is significant

5) What is the difference between chi square goodness of fit test and measure of independence test?

6) What do you understand by efficiency of T?

4.13 SUGGESTED READINGS

Daniel, W. W. (1990) *Applied Non-parametric Statistics*, 2d ed. Boston: PWS-Kent.

Johnson, Morrell, and Schick (1992), Two-Sample Non-parametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

Siegel S. and Castellan N.J. (1988) *Non-parametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw Hill.

Wampold BE & Drew CJ. (1990) *Theory and Application of Statistics*. New York: McGraw-Hill.