Contents

Chapter 11

Association Between Variables

11.1 Introduction

In previous chapters, much of the discussion concerned a single variable, describing a distribution, calculating summary statistics, obtaining interval estimates for parameters and testing hypotheses concerning these parameters. Statistics that describe or make inferences about a single distribution are referred to as **univariate statistics**. While univariate statistics form the basis for many other types of statistics, none of the issues concerning relationships among variables can be answered by examining only a single variable. In order to examine relationships among variables, it is necessary to move to at least the level of **bivariate statistics**, examining two variables. Frequently the researcher wishes to move beyond this to **multivariate statistics**, where the relationships among several variables are simultaneously examined.

Cross classification tables, used to determine independence and dependence for events and for variables, are one type of bivariate statistics. A test for a difference between two proportions can also be considered a type of bivariate statistics. The only other example of bivariate methods used so far in this textbook is the test for the difference between two means, using either the normal or the t distribution. The latter is the only bivariate method which has been used to examine variables that have interval or ratio level scales.

An example of a relationship that a researcher might investigate is the relationship between political party supported and opinion concerning socioeconomic issues. In Chapters 9 and 10, the relationship between political party supported and opinion concerning various explanations for unemployment, among a sample of Edmonton adults, was examined. This type of relationship was examined using a cross classification table and the chi square statistic. Differences of proportions, or difference of mean opinion could have been used as a method of examining this relationship as well. In this chapter, various summary measures are used to describe these relationships. The chi square statistic from the cross classification table is modified to obtain a measure of association. Correlation coefficients and regression models are also used to examine the relationship among variables which have ordinal, interval or ratio level scales.

Bivariate and multivariate statistics are useful not only for statistical reasons, but they form a large part of social science research. The social sciences are concerned with explaining social phenomena and this necessarily involves searching for, and testing for, relationships among variables. Social phenomena do not just happen, but have causes. In looking for causal factors, attempting to determine which variables cause or influence other variables, the researcher examines the nature of relationships among variables. Variables that appear to have little relationship with the variable that the researcher is attempting to explaing may be ignored. Variables which appear to be related to the variable being explained must be closely examined. The researcher is concerned with whether a relationship among variables exists or not. If the relationship appears to exist, then the researcher wishes to know more concerning the nature of this relationship. The size and strength of the relationship are of concern, and there are various tests concerning these.

In this chapter, there is no examination of multivariate relationships, where several variables are involved. This chapter looks only at bivariate relationships, testing for the existence of such relationships, and attempting to describe the strength and nature of such relationships. The two variable methods of this chapter can be extended to the examination of multivariate relationships. But the latter methods are beyond the scope of an introductory textbook, and are left to more advanced courses in statistics.

11.1.1 Measure of Association

Measures of association provide a means of summarizing the size of the association between two variables. Most measures of association are scaled so that they reach a maximum numerical value of 1 when the two variables have a perfect relationship with each other. They are also scaled so that they have a value of 0 when there is no relationship between two variables. While there are exceptions to these rules, most measures of association are of this sort. Some measures of association are constructed to have a range of only 0 to 1, other measures have a range from -1 to +1. The latter provide a means of determining whether the two variables have a positive or negative association with each other.

Tests of significance are also provided for many of the measures of association. These tests begin by hypothesizing that there is no relationship between the two variables, and that the measure of association equals 0. The researcher calculates the observed value of the measure of association, and if the measure is different enough from 0, the test shows that there is a significant relationship between the two variables.

11.1.2 Chapter Summary

This chapter begins with measures of association based on the chi square statistic. It will be seen in Section 11.2 that the χ^2 statistic is a function not only of the size of the relationship between the two variables, but also of the sample size and the number of rows and columns in the table. This statistic can be adjusted in various ways, in order to produce a measure of association. Following this, in Section 11.3, a different approach to obtaining a measure of association is outlined. This is to consider how much the error of prediction for a variable can be reduced when the researcher has knowledge of a second variable. Section **??** examines various correlation coefficients, measures which summarize the relationship between two variables that have an ordinal or higher level of measurement. Finally, Section **??** presents the regression model for interval or ratio variables. The regression model allows the researcher to estimate the size of the relationship between two variables, where one variable is considered the independent variable, and the other variable depends on the first variable.

11.2 Chi Square Based Measures

One way to determine whether there is a statistical relationship between two variables is to use the chi square test for independence of Chapter 10. A cross classification table is used to obtain the expected number of cases under the assumption of no relationship between the two variables. Then the value of the chi square statistic provides a test whether or not there is a statistical relationship between the variables in the cross classification table.

While the chi square test is a very useful means of testing for a relationship, it suffers from several weakenesses. One difficulty with the test is that it does not indicate the nature of the relationship. From the chi square statistic itself, it is not possible to determine the extent to which one variable changes, as values of the other variable change. About the only way to do this is to closely examine the table in order to determine the pattern of the relationship between the two variables.

A second problem with the chi square test for independence is that the size of the chi square statistic may not provide a reliable guide to the strength of the statistical relationship between the two variables. When two different cross classification tables have the same sample size, the two variables in the table with the larger chi square value are more strongly related than are the two variables in the table with the smaller chi square value. But when the sample sizes for two tables differ, the size of the chi square statistic is a misleading indicator of the extent of the relationship between two variables. This will be seen in Example 11.2.1.

A further difficulty is that the value of the chi square statistic may change depending on the number of cells in the table. For example, a table with 2 columns and 3 rows may give a different chi square value than does a cross classification table with 4 columns and 5 rows, even when the relationship between the two variables and the sample sizes are the same. The number of rows and columns in a table are referred to as the **dimensions** of the table. Tables of different dimension give different degrees of freedom, partly correcting for this problem. But it may still be misleading to compare the chi square statistic for two tables of quite different dimensions.

In order to solve some of these problems, the chi square statistic can be adjusted to take account of differences in sample size and dimension of the table. Some of the measures which can be calculated are phi, the contingency coefficient, and Cramer's V. Before examining these measures, the following example shows how sample size affects the value of the chi square statistic.

Example 11.2.1 Effect of Sample Size on the Chi Square Statistic

The hypothetical examples of Section 6.2 of Chapter 6 will be used to illustrate the effect of sample size on the value of the chi square statistic. The data from Tables 6.9 and 6.10 will first be used to illustrate how a larger chi square value can be used to indicate a stronger relationship between two variables when two tables have the same sample size. Then the misleading nature of the chi square statistic when sample size differs will be shown.

Opinion	Male	Female	Total
Agree	65 (60.0)	25 (30.0)	90
Disagree	35 (40.0)	25 (20.0)	60
Total	100	50	150

$$\chi^2 = 0.417 + 0.833 + 0.625 + 1.250 = 3.125$$

$$df = 1$$

$$0.075 < \alpha < 0.10$$

Table 11.1: Weak Relationship between Sex and Opinion

Table 11.1 gives the chi square test for independence for the weak relationship between sex and opinion, originally given in Table 6.9. The first entry in each cell of the table is the count, or observed number of cases. The number in brackets in each cell of the table is the expected number of cases under the assumption of no relationship between sex and opinion. It can be seen that the value of the chi square statistic for the relationship shown in Table 11.1 is 3.125. With one degree of freedom, this value is statistically significant at the 0.10 level of significance, but not at the 0.075 level. This indicates a rather weak relationship, providing some evidence for a relationship between sex and opinion. But the null hypothesis of no relationship between the two variables can be rejected at only the 0.10 level of significance.

Table 11.2 gives much stronger evidence for a relationship between sex and opinion. In this table, the distribution of opinions for females is the same as in the earlier table, but more males are in agreement, and less in disagreement than in the earlier table. As a result, the chi square value for Table 11.2 gives a larger value, indicating a more significant relationship

Opinion	Male	Female	Total
Agree	75 (66.7)	25 (33.3)	100
Disagree	25 (33.3)	25 (16.7)	50
Total	100	50	150

$$\chi^2 = 1.042 + 2.083 + 2.083 + 4.167 = 9.375$$

$$df = 1$$

$$0.001 < \alpha < 0.005$$

Table 11.2: Strong Relationship between Sex and Opinion

than in Table 11.1. For Table 11.2, the chi square value is 9.375, and with one degree of freedom, this statistic provides evidence of a relationship at the 0.005 level of significance.

When comparing these two tables, the size of the chi square value provides a reliable guide to the strength of the relationship between sex and opinion in the two tables. The larger chi square value of Table 11.2 means a stronger relationship between sex and opinion than does the smaller chi square value of Table 11.1. In these two tables, the sample size is the same, with $n = 150$ *cases in each table.*

Now examine Table 11.3, which is based on the weak relationship of Table 11.1, but with the sample size increased from $n = 150$ *to* $n = 600$. *In order to preserve the nature of the relationship, each of the observed numbers of cases in the cells of Table 11.1 are multiplied by 4. The new table again shows that females are equally split between agree and disagree, but males are split* $260/140 = 65/35$ *between agree and disagree. The pattern of the relationship between sex and opinion is unchanged from Table 11.1. But now the chi square statistic is dramatically increased. In Table 11.3, the chi square statistic is 12.5, as opposed to only 3.125 in Table 11.1. The 12.5 of the new table is even larger than the chi square value of 9.375 of Table 11.2. The larger sample size in the new table has increased the value of the chi* *square statistic so that even the relatively weak relationship between sex and opinion becomes very significant statistically. Given the assumption of no relationship between sex and opinion, the probability of obtaining the data of Table 11.3 is less than 0.0005.*

Opinion	Male	Female	Total
Agree	260 (240.0)	100 (120.0)	360
Disagree	140 (160.0)	100 (80.0)0	240
Total	400	200	600

$$\chi^2 = 1.667 + 3.333 + 2.500 + 5.000 = 12.500$$

$$df = 1$$

$$\alpha < 0.0005$$

Table 11.3: Weak Relationship - Larger Sample Size

This example shows how the value of the chi square statistic is sensitive to the sample size. As can be seen by comparing Tables 11.1 and 11.3, multiplying all the observed numbers of cases by 4 also increases the chi square statistic by 4. The degrees of freedom stay unchanged, so that the larger chi square value appears to imply a much stronger statistical relationship between sex and opinion.

Considerable caution should be exercised when comparing the chi square statistic, and its significance, for two tables having different sample sizes. If the sample size for the two tables is the same, and the dimensions of the table are also identical, the table with the larger chi square value generally provides stronger evidence for a relationship between the two variables. But when the sample sizes, or the dimensions of the table differ, the chi square statistic and its significance may not provide an accurate idea of the extent of the relationship between the two variables.

One way to solve some of the problems associated with the chi square statistic is to adjust the chi square statistic for either the sample size or the dimension of the table, or for both of these. Phi, the contingency coefficient and Cramer's V, are measures of association that carry out this adjustment, using the chi square statistic. These are defined in the following sections, with examples of each being provided.

11.2.1 Phi

The measure of association, *phi*, is a measure which adjusts the chi square statistic by the sample size. The symbol for phi is the Greek letter phi, written ϕ, and usually pronounced 'fye' when used in statistics. Phi is most easily defined as

$$\phi = \sqrt{\frac{\chi^2}{n}}.$$

Sometimes phi squared is used as a measure of association, and phi squared is defined as

$$\phi^2 = \frac{\chi^2}{n}.$$

In order to calculate these measures, the chi square statistic for the table is first determined, and from this it is relatively easy to determine phi or phi squared. Since phi is usually less than one, and since the square of a number less than one is an even smaller number, ϕ^2 can be extremely small. This is one the reasons that phi is more commonly used than is phi squared.

Example 11.2.2 ϕ and ϕ^2 for Tables of Section 11.2

Table 11.4 gives the three two by two tables shown in the last section, without the row and column totals. The chi square statistic and sample size for each of the tables is given below the frequencies for each cell in the table. From these, ϕ^2 and ϕ are then determined. For the first table, with the strong relationship, having females equally divided on some issue, but with males split 75 agreeing and 25 disagreeing, $\chi^2 = 9.375$. The sample size for this table is $n = 150$ so that

$$\phi^2 = \frac{\chi^2}{n} = \frac{9.375}{150} = 0.0625$$

and

$$\phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{9.375}{150}} = \sqrt{0.0625} = 0.25.$$

	Nature of Relation and Sample Size					
	Strong, $n = 150$		Weak, $n = 150$		Weak, $n = 600$	
Opinion	Male	Female	Male	Female	Male	Female
Agree	75	25	65	25	260	100
Disagree	25	25	35	25	140	100
χ^2	9.375		3.125		12.500	
n	150		150		600	
ϕ^2	0.0625		0.02083		0.02083	
ϕ	0.25		0.144		0.144	

Table 11.4: ϕ^2 and ϕ for 2×2 Tables

The values of ϕ^2 and ϕ for the other tables are obtained in a similar manner. Note how small ϕ^2 is in each of the tables. Since a very small value might seem to indicate no relationship between the two variables, sex and opinion, it might be preferable to use ϕ rather than ϕ^2. Note that ϕ is 0.25 for the strong relationship, and only 0.144 for the weak relationship. By comparing the two values of ϕ, you can obtain some idea of the association between sex and opinion. This indicates that the relationship of Table 11.2, for which $\phi = 0.25$, is a stronger relationship than is the relationship of Table 11.1, where ϕ is only 0.144. Also note in the two right panels of Table 11.4 that ϕ for the weak relationship is the same, regardless of the sample size. As shown earlier, in Tables 11.1 and 11.3, the value of χ^2 is quite different for these two types of data, but ϕ is the same. That is, the nature of the relationship is the same in the two right panels of Table 11.4, but the sample size is four times greater on the right than in the middle panel. This dramatically increases the size of the chi square statistic, but leaves the values of ϕ^2 and ϕ unchanged.

Example 11.2.3 Relationship Between Age and Opinion Concerning Male and Female Job Roles

The Regina Labour Force Survey asked respondents the question

> *Do you strongly agree, somewhat agree, somewhat disagree or strongly disagree that a majority of jobs are best done by men?*

Responses to this question could be expected to vary considerably for different types of respondents. It might be expected, for example, that older people, with more traditional views concerning women's and men's roles, would agree with this view. On the other hand, younger respondents might be expected to have views which are more consistent with similar job roles for the two sexes. This expectation is examined in this example.

Table 11.5 gives a cross classification of age and opinion concerning whether a majority of jobs are best done by men. Age has been classified into three categories, 'less than 40,' '40 to 54,' and '55 and over.' The first entry in each cell of the table is the observed number of cases, and the second entry in each cell is the expected number of cases. The χ^2 statistic, along with ϕ^2 and ϕ are given at the bottom of the table.

Again the value of ϕ^2 is very small, so it might seem as if there is no relationship between the two variables, age and opinion. But $\phi = 0.2247$, indicating that there is some relationship between the two variables. While the relationship is not close to 1, it is considerably greater than 0, indicating that for different ages, there are considerably different opinions.

By examining the differences between the observed and expected numbers of cases, the pattern of the relationship between age and opinion can be determined. These differences are greatest for the youngest and the oldest age groups. For ages under 40, there are considerably fewer observed respondents than the number of respondents expected to agree under the hypothesis of no relationship between the two variables. In contrast, for the 55 and over age group, there are more observed than expected respondents who agree. What these results show is that younger respondents tend to disagree that the majority of jobs are best done by men, older respondents tend to agree, and the middle age group falls between these two. The χ^2 statistic and ϕ support the view that there is a relationship between age and opinion, with more egalitarian views concerning male and female job roles among younger than among older respondents.

The measures of association ϕ and ϕ^2 cannot have a value less than zero, since the minimum value for χ^2 is zero. If there is no relationship between the two variables, so that $\chi^2 = 0$, then $\phi = \phi^2 = 0$. If the chi square value is small, ϕ will also be relatively small. When the chi square statistic is large, indicating a strong relationship between the two variables, then ϕ will

	Attitude Response				
	Strongly	Somewhat	Somewhat	Strongly	
Age	Agree	Agree	Disagree	Disagree	Total
< 40	46 (64.5)	115 (129.1)	135 (123.9)	230 (208.6)	526
40-54	26 (24.7)	45 (49.3)	41 (47.3)	89 (79.7)	201
55+	40 (22.8)	64 (45.6)	39 (43.8)	43 (73.7)	186
Total	112	224	215	362	913

$$\chi^2 = 46.116; \quad \text{df} = 6; \quad \alpha < 0.00001$$

$$\phi^2 = \frac{46.116}{913} = 0.05051; \qquad \phi = \sqrt{0.05051} = 0.2247$$

Table 11.5: Relationship Between Age and Attitude

also be large. Exactly what is large and what is small depends on the type of data being compared. When examining several tables, those tables with larger values for ϕ indicate stronger relationships between variables than do those tables with smaller values for ϕ.

The maximum value for ϕ depends on the dimensions of the table. If a table has r rows and c columns, then the maximum possible value for ϕ is one less than the smaller of r or c. For example, if a table has 5 rows and 3 columns, then ϕ could be as large as $3 - 1 = 2$. This makes it difficult to compare values of ϕ between tables of different dimensions. However, if several tables, all of the same dimensions are being compared, ϕ is a useful measure.

11.2.2 Contingency coefficient

A slightly different measure of association is the **contingency coefficient**. This is another chi square based measure of association, and one that also adjusts for different sample sizes. The contingency coefficient can be defined as

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}.$$

Since $\phi^2 = \chi^2/n$, it is straightforward to show that

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}.$$

The contingency coefficient has much the same advantages and disadvantages as does ϕ. When there is no relationship between two variables, $C = 0$. The contingency coefficient cannot exceed the value $C = 1$, so that it is constrained more than is ϕ. But the contingency coefficient may be less than 1 even when two variables are perfectly related to each other. This means that it is not as desirable a measure of association as those which have the range 0 to 1.

Example 11.2.4 Contingency Coefficient for Two by Two Tables

The three tables examined in Example 11.2.2 can be used to show the size of the contingency coefficient. These are given in Table 11.6. For the first of the three relationships, the contingency coefficient is

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{9.375}{150 + 9.375}} = \sqrt{0.05882} = 0.2425$$

	Nature of Relation and Sample Size					
	Strong, $n = 150$		Weak, $n = 150$		Weak, $n = 600$	
Opinion	Male	Female	Male	Female	Male	Female
Agree	75	25	65	25	260	100
Disagree	25	25	35	25	140	100
χ^2	9.375		3.125		12.500	
n	150		150		600	
C	0.2425		0.1429		0.1429	

Table 11.6: Contingency Coefficient for 2 × 2 Tables

A similar calculation shows that $C = 0.1429$ for each of the two weaker relationships shown in the middle and the right of Table 11.6.

The middle and right panels show that when the relationship is the same, the contingency coefficient does not change in size as the sample size changes. Also note that the contingency coefficient is considerably larger in the case of the stronger relationship of the panel on the left of Table 11.6 than in the case of the centre and right panels. In general, the contingency coefficient has a similar value, and similar behaviour to that of ϕ, with the advantage that it cannot exceed 1.

Example 11.2.5 Measures of Association between Saskatchewan Provincial Political Preference and 4 Attitude Questions

In order to illustrate how measures of association can be used to summarize the strength of relationship between variables, Table 11.7 gives four sets of measures of association. These measures were obtained on a computer, using cross classification tables which are not presented here. (The cross classifications of provincial political preference by variable G12 and G14 are given below in Tables 11.8 and 11.9.) In each of the four cases, the relationship is between one of the attitude variables in the Regina Labour Force Survey, and respondents' provincial political preference. The four attitude variables are as follows:

G8 Do you strongly agree, somewhat agree, somewhat disagree or strongly disagree that government and politics are so complicated that it's hard to understand what's going on.

G18 Do you strongly agree, somewhat agree, somewhat disagree or strongly disagree that there are a majority of jobs that are best done by men?

G12 Do you strongly agree, somewhat agree, somewhat disagree or strongly disagree that government is more on the side of business than labour in labour relations.

G14 If a federal election were held in the next month, which party would you vote for?

Political preference at the provincial level and at the federal level is based on support for the three major political parties.

	Relationship Between Provincial Political Preference and:			
Statistic	G8	G18	G12	G14
χ^2	3.411	9.280	126.105	536.744
df	6	6	6	4
Significance	0.756	0.158	< 0.00001	< 0.00001
Sample Size	599	591	566	553
ϕ^2	0.006	0.016	0.223	0.971
ϕ	0.075	0.125	0.472	0.985
C	0.075	0.124	0.427	0.702
Cramer's V	0.053	0.089	0.334	0.697

Table 11.7: Measures of Association for Relationship between Provincial Political Preference and 4 Attitude Questions

The four relationships have been placed in order, from the smallest relationship to the largest relationship. The first three columns of these statistics are based on cross classification tables which have four rows representing the four different attitudes. Since there are only 3 political preferences, this means that there are $(4-1) \times (3-1) = 6$ *degrees of freedom for the first three columns of Table 11.7. The last column of this table is based on a* 3×3 *cross classification table, giving only 3 degrees of freedom.*

The first column gives summary statistics for the relationship between provincial political preference and G8, politics and government are too hard to understand. This relationship is almost nonexistent, with a small chi square value, so small that it is not siginificant, with 0.756 being the level of significance. The chi squared based measures of association, phi, phi squared, and the contingency coefficient are also very low, indicating a very small relationship between political preference and question G8. (Cramer's V is examined in the next section). The lack of association implies that the levels of agreement and disagreement concerning whether politics and government are too hard to understand are very similar among supporters of the 3 major political parties.

The second column of Table 11.7 gives the relationship between provincial political preference and the attitude question concerning whether or not a majority of jobs are best done by men. In Example 11.2.3, there was a strong relationship between age and this attitude question. But Table 11.7 shows that supporters of different political have fairly similar views on this attitude question. The chi square value is larger than in the case of variable G8, with a smaller significance level. The significance is still over 0.15, considerably higher than the level usually regarded as sufficient to conclude that there is a relationship. The strength of the association between political preference and G18 is fairly small, as indicated by a low $\phi = 0.125$ *and a similar low contingency coefficient of 0.124. These small values for the measures of association indicate a relatively small association between the variables.*

For the third column, there is a much larger relationship, with a very statistically significant chi square value, and larger values for the measures of association. For the relationship between political preference and attitude concerning whether government is more on the side of business than of labour, $\phi = 0.472$ *and* $C = 0.427$*. These are considerably larger than the corresponding measures for the first two columns, indicating that political preference is related to this attitude question. Supporters of the different political parties do have quite different views on this question. By examining*

Table 11.8 below, the patterns of the relationship can be observed.

For the final column of Table 11.7, the relationship between political preferences at the provincial and federal levels is very strong. The chi square statistic is extremely large, and the values for phi and the contingency coefficient are much closer to 1 than in the case of either of the other columns. These results imply that respondents tend to support the same political party at both the federal and provincial level. While the relationship is not perfect, it is a very strong relationship. This can clearly be seen in Table 11.9 below.

The last example shows how the measures of association can be used to summarize the relationship between two variables. Rather than presenting four cross classification tables, only the summary measures of association need be presented. Not all of these measures need be presented either, but the researcher might present only ϕ or C. By comparing these across the 4 relationships, the strength of association can be shown in summary form.

11.2.3 Cramer's V

One final chi square based measure of association that can be used is Cramer's V. This measure is defined as

$$V = \sqrt{\frac{\phi^2}{t}} = \sqrt{\frac{\chi^2}{nt}}$$

where t is the smaller of the number of rows minus one or the number of columns minus one. If r is the number of rows, and c is the number of columns, then

$$t = \text{Minimum } (r-1, c-1).$$

By using the information concerning the dimensions of the table, Cramer's V corrects for the problem that measures of association for tables of different dimension may be difficult to compare directly. Cramer's V equals 0 when there is no relationship between the two variables, and generally has a maximum value of 1, regardless of the dimension of the table or the sample size. This makes it possible to use Cramer's V to compare the strength of association between any two cross classification tables. Tables which have a larger value for Cramer's V can be considered to have a strong relationship between the variables, with a smaller value for V indicating a weaker relationship.

Example 11.2.6 Cramer's V for Two by Two Tables

For the three two by two tables examined earlier, Cramer's V has the same value as ϕ. With a cross classification table having two rows and two columns, $r = c = 2$ so that $r - 1 = c - 1 = t$ and

$$V = \sqrt{\frac{\phi^2}{t}} = \sqrt{\frac{\phi^2}{1}} = \phi.$$

When working with a table which has only two rows and two columns, it makes no difference whether ϕ or V is used.

Example 11.2.7 Cramer's V for the relationship between Provincial Political Preference and Attitude Concerning Government and Labour

One of the cross classifications reported in Example 11.2.5 is given in this example. In the earlier example, the measures of association were reported, but the cross classifications themselves, from which these measures were obtained, were not reported. Table 11.8 shows the relationship between political preference at the provincial level in Saskatchewan, and variable G12. Variable G12 asks respondents' atttitudes concerning whether they view government being more on the side of business than labour in labour relations. This data was obtained from the Regina Labour Force Survey.

In Table 11.8, $\chi^2 = 126.105$ and the sample size is $n = 566$. There are $r = 4$ rows and $c = 3$ columns, so that $r - 1 = 4 - 1 = 3$ and $c - 1 = 3 - 1 = 2$. In determining V, t is the smaller of these, so that $t = c - 1 = 2$. Cramer's V is

$$V = \sqrt{\frac{\chi^2}{nt}} = \sqrt{\frac{126.105}{566 \times 2}} = \sqrt{0.1114} = 0.334.$$

Once the χ^2 value has been calculated, the determination of V is relatively straightforward.

The value of V for this table is not real large, but it is larger than V in the case of the first two relationships in Table 11.7. This means that there is some relationship between political preference and attitude toward government and labour. By examining the patterns in Table 11.8, the nature of the relationship between political preference and attitude can be determined. Note first that the majority of respondents either strongly or somewhat agree that government is more on the side of business than of

Attitude	Provincial Political Preference			
G12	Liberal	NDP	Conservative	Total
Strongly Agree	31	227	31	289
	(44.9)	(168.5)	(75.6)	
Somewhat Agree	31	77	54	162
	(25.2)	(94.5)	(42.4)	
Somewhat Disagree	18	19	37	74
	(11.5)	(43.1)	(19.3)	
Strongly Disagree	8	7	26	41
	(6.4)	(23.9)	(10.7)	
Total	88	330	148	566

Table 11.8: Cramer's V for the Relationship between G12 and Provincial Political Preference

labour in labour relations. In addition, note that the NDP is the party which received by far the most support, followed by the Conservatives and Liberals.

The most dramatic differences are between the responses of the NDP and the Conservatives. Assuming there is no relationship between the two variables, there are many more NDP supporters than expected who strongly agree that government is more on the side of business than of labour. In contrast, relatively few Conservative supporters strongly agree, and the same can be said for the Liberals. For the Conservatives, there are many more observed than expected on the disagree side. In summary then, NDP supporters are heavily in agreement, Conservatives more in disagreement, with Liberals falling somewhere in between.

11.2.4 Summary of Chi Square Based Measures

Each of ϕ, ϕ^2, C and V are useful measures of association.

When there is no relationship between the two variables, each of these measures has a value of 0. As the extent of the relationship between the variables increases, each of these measures also increases, although by different amounts. Where these measures differ is in their maximum value. If there is a perfect relationship between the two variables of a cross classification table, it would be preferable to have the measure of association have a value of 1. ϕ and ϕ^2 can have values exceeding 1 in the case of tables with more than 2 rows and 2 columns. In contrast, the contingency coefficient C sometimes has a maximum value which is less than 1.

Cramer's V is the preferred measure among these χ^2 based measures. It generally has a maximum value of 1 when there is a very strong relationship between two variables. When it is computed, Cramer's V takes account of the dimensions of the table, implying that V for tables of different dimensions can be meaningfully compared. When comparing several tables of the same dimension, and similar sample size, it makes little difference which of these measures is used. Different researchers prefer different measures, and the choice of measure is largely a matter of personal preference.

You may have noted that no tests of significance have been given for these measures. In general, the test of significance is the same for each of these measures as it is for the chi square test of independence. The SPSS computer program gives the significance level for these measures of association when these measures are requested. Iin Example 11.2.5 for the relationship between provincial political preference and variable G18, $V = 0.089$. SPSS reports the significance of V for this relationship is $\alpha = 0.158$. From Table 11.7, it can be seen that this is the same level of significance as for the χ^2 statistic for this relationship.

The null hypothesis for the hypothesis test concerning each measure is that there is no association between the two variables in the cross classification table. The alternative hypothesis is that there is some association between the two variables. The significance level reported is the probability of obtaining a statistic of that value, assuming no relationship between the two variables. In the case of the relationship between provincial political preference and G18, this probability is 0.158. That is, assuming that there is no relationship between provincial political preference and attitude concerning whether or not a majority of jobs are best done by men, the probability of obtaining a V of 0.089 or larger is 0.158. This is not a real small probability, and the researcher might not reject the null hypothesis of no association on the basis of a probability of 0.089. While there may be some association between these two variables, it is not a large one, and there is not strong evidence for an association. In contrast, for the relationship between political preference and each of variables G12 and G14, Table 11.7 shows that the probability of obtaining measures of association as large as reported, is less than 0.00001. This is very strong evidence of an association between each pair of these variables.

The chi square based measures of association are often used to determine the strength of relationships where at least one of the variables is nominal. When both variables are measured at the ordinal or higher level, other measures of association, such as correlation coefficients, are more commonly used. The following section shows another approach to determining the nature of association, and correlation coefficients are discussed in Section **??**.

11.3 Reduction in Error Measures

A different approach to measuring association is to attempt to predict the values of one variable in two different ways. First, the values for one variable are predicted without knowing the values of a second variable. Then the values of the first variable are predicted again, after taking into account the values of the second variable. The extent to which the error of prediction for values of one variable can be reduced by knowing the value of a second variable forms the basis for the **reduction in error** approach to measuring association.

The most common measure of association that is based on this approach is *lambda*. The measure of association lambda has the symbol λ, the Greek letter lambda. Rather than attempt to discuss this measure in the abstract, the following examples are used to introduce and explain the measure. Once you see how this measure is calculated, it becomes a relatively simple matter to calculate. A general formula for λ is given after the following examples, on pages 791 and 793.

Example 11.3.1 Lambda for Relationship between Provincial and Federal Political Preference

Another cross classification for which summary statistics were reported in Table 11.7 is presented here. This example looks at the relationship between provincial and federal political preference in Saskatchewan. Table 11.9 gives provincial political preference of respondents in the columns of the table. Political preference for the same political parties at the federal level is shown in the various rows. A quick glance at the table confirms the suspicion that supporters of each political party at the provincial level generally *support the same party at the federal level. The summary statistics shown in Table 11.7 showed a large association between political preference at the two levels, with* $\phi = 0.985$, $V = 0.697$. *All of these measures are significant statistically at less than the 0.00001 level. At the same time, the relationship is not a perfect one, with some respondents supporting a party at one level but not at the other. The most notable switches are that considerable numbers of both the provincial NDP (58) and Conservative (24) supporters switch their support to the Liberals at the federal level.*

Federal Political	Provincial Political Preference			
Preference	Liberal	NDP	Conservative	Total
Liberal	80	58	24	162
NDP	6	244	4	254
Conservative	5	14	118	137
Total	91	316	146	553

Table 11.9: Cross Classification of Provincial and Federal Political Preference, Regina

The reduction in error approach to measuring association for this table is to begin by asking how many errors of prediction there would be concerning values of one of the variables, if values of the other variable are **not** *known. Suppose that the researcher is interested in predicting the federal vote in Regina. If one of the 553 respondents in the table is picked, and nothing is known about what the political preference of this respondent is, what would be the best guess concerning which party this respondent supports at the federal level? Suppose first that the researcher predicts that the respondent will vote Liberal at the federal level. Of the 553 respondents, 162 say they vote Liberal federally, so the researcher is correct in only 162 out of 553 cases. This means that there are* $553 - 162 = 391$ *errors of prediction.*

Suppose next that the researcher changes the prediction, and predicts that any respondent selected will vote NDP at the federal level. In this case, the researcher is correct in 254 of the 553 cases, since 254 of the *respondents say they will vote this way. The number of errors of prediction, if NDP is predicted, is* $553 - 254 = 299$. *Finally, if the researcher predicts a Conservative vote at the federal level for each respondent, the researcher will make* $553 - 137 = 416$ *errors of prediction.*

In this example, if provincial political preference of respondents is not known, the best guess is to predict the value of the variable with the largest row total. In this case, this prediction is that a respondent will vote NDP. The number of errors of prediction is then $553 - 254 = 299$ *errors, and this prediction results in fewer errors than if either a Liberal or Conservative vote is predicted.*

The next step in determining λ *is to ask whether these errors of prediction can be reduced by knowing the political preference of the respondent at the provincial level. The above exercise began with no knowledge of political preference at the provincial level. In order to determine the reduction in the number of errors of prediction, it is necessary to examine each of the three values of provincial political preference.*

Begin first with those who support the Liberals at the provincial level. If the researcher knows that a respondent supports the Liberals at the provincial level, what is the best prediction concerning how the respondent will vote at the federal level. Of the 91 respondents who vote Liberal at the provincial level, the largest single number, 80, also vote Liberal at the federal level. If the researcher predicts that any voter who votes Liberal provincially will also vote Liberal federally, he or she would be correct 80 times and incorrect in $91 - 80 = 11$ *cases. There are thus 11 errors of prediction in this case. This is fewer errors of prediction than if the researcher had predicted a federal NDP or Conservative vote for provincial Liberal supporters. The general method of approach then is that for each column of the table, the row with the largest entry is the row whose value is predicted.*

Next look at the NDP. Of the 316 respondents who vote NDP provincially, 244 also vote NDP federally. If the researcher predicts that each of these 316 provincial NDP supporters will support the NDP federally, there are $316 - 244 = 72$ *errors of prediction. Finally, for the Conservatives, if the researcher predicts that each of the 146 provincial supporters will vote Conservative federally, there are* $146 - 118 = 28$ *errors of prediction.*

Using the method of the last two paragraphs, including the knowledge of the respondent's provincial political preference, there are $11 + 72 + 28 = 111$ *errors of prediction. This is considerably fewer errors of prediction than when provincial political preference was not known. Earlier, when provincial political preference was not taken into account, there were 299 errors of* *prediction.*

Lambda is defined as the **proportional reduction in error** *of prediction as a result of using the information concerning the second variable. In this case, the number of errors was originally 299 errors of prediction, and this was reduced by* $299 - 111 = 188$ *errors of prediction. The proportional reduction in the number of errors of prediction is thus*

$$\lambda = \frac{299 - 111}{299} = \frac{188}{299} = 0.629.$$

The value 0.629 means that the number of errors of prediction of federal political preference can be reduced by 0.629, or by 62.9%, if provincial political preference is known. This is quite a considerable reduction in errors of prediction. If knowledge of provincial political preference of a respondent in Saskatchewan is available, then a considerable improvement in the prediction of their federal political preference can be made, compared to the situation where provincial political preference is not known.

From the above example, it can be seen that the maximum value of λ is 1 and the minimum value is 0. When there is no association between the two variables, and knowledge of one variable does not help reduce the number of errors of prediction for the second variable, the number of errors of prediction is the same in each of the two cases, and $\lambda = 0$. On the other hand, if the two variables are perfectly associated, so that there are no errors of prediction when the second variable is known. In that case the reduction in the number of errors is 100%, producing a value of λ of 1. An example of each of these situations follows.

Example 11.3.2 λ for No and Perfect Association

Table 11.10 gives an hypothetical example of no association between the two variables sex and opinion. This example gives the same data as Table 6.6, where the data were used to demonstrate independence between sex and opinion. By examining the data in this table, it can be seen that there is no relationship between sex and opinion. Since both males and females have the same distribution of opinion, 60% in agreement, and 40% in disagreement, knowledge of the sex of the respondent is of no help in predicting the opinion of the respondent, and it will be seen that $\lambda = 0$.

If the opinion of the respondent is being predicted, and the respondent's sex is not known, then the best prediction is that the respondent agrees. This results in $150 - 90 = 60$ *errors of prediction. While this is considerable,*

	Sex		
Opinion	Male	Female	Total
Agree	60	30	90
Disagree	40	20	60
Total	100	50	150

Table 11.10: No Association

this is a better prediction than 'Disagree' where there would be $150-60=90$ *errors of prediction.*

Now if the sex of the respondent is known, for the males the best prediction is also 'Agree'. Of the 100 males, 60 agree, so that there are $100-60=40$ *errors of prediction. For the 50 females, the best prediction would again be 'Agree', and there are* $50-30=20$ *errors of prediction. The total number of errors of prediction for both males and females is* $40+20=60$.

Knowing the respondent's sex, there is no improvement in the number of errors of prediction. Without knowing sex, there were 60 errors of prediction, and even when the respondent's sex is known, there are 60 errors of prediction. For this example,

$$\lambda = \frac{60-60}{60} = \frac{0}{60} = 0,$$

and for this table, $\lambda = 0$ *shows there is no association between sex and opinion.*

Now examine Table 11.11 where there is an hypothetical example of perfect association between the two variables sex and opinion. In this example, all the males agree, with no males disagreeing. In contrast, all the females disagree on the opinion issue, and none agree. In this case, knowing the sex of the respondent is essential to predicting the opinion of the respondent, and it will be seen that $\lambda = 1$.

As before, if the opinion of the respondent is being predicted, and the respondent's sex is not known, then the best prediction is that the respondent agrees. This is because there are more respondents who agree than who

	Sex		
Opinion	Male	Female	Total
Agree	100	0	100
Disagree	0	50	50
Total	100	50	150

Table 11.11: Perfect Association

disagree. Predicting that any randomly selected respondent agrees results in $150 - 100 = 50$ *errors of prediction. While this is considerable, this is a better prediction than 'Disagree' where there would be* $150 - 50 = 100$ *errors of prediction.*

Now suppose that the sex of the respondent is known. For the males, the best prediction is also 'Agree'. Of the 100 males, all 100 agree, so that there are $100 - 100 = 0$ *errors of prediction. For the 50 females, the best prediction would be 'Disagree', and this would result in* $50 - 50 = 0$ *errors of prediction. For each of males and females, there are no errors of prediction.*

Using these errors of prediction, the value of λ *is*

$$\lambda = \frac{100 - 0}{100} = \frac{100}{100} = 1.$$

That is, when the sex of the respondent was not known, there were 100 errors of prediction. When the sex of respondent is known, there are 0 errors of prediction. This means a reduction from 100 to 0 errors, or a complete reduction in the number of errors of prediction. λ *is 1 in this case, and a situation such as this is often defined as* **perfect association** *between the variables.*

General Formula for λ. Let the number of errors of prediction when the column totals are not known be E_{TR}. For each column j, where $j = 1, 2, \ldots, c$, let the number of errors of prediction be E_{C_j} where there are c columns. Then let E_C be the sum of these E_{C_j}. That is, E_C is the total number of errors of prediction when the columns are used to predict the likely row results. Then

$$\lambda_R = \frac{E_{TR} - E_C}{E_{TR}}$$

where λ_R is the value of λ when the row values are being predicted. It will be seen a little later in this section that there can be a different value for λ when the column values are being predicted.

A formula for each of E_{TR} and E_C can also be given. For E_{TR}, the method was to take the total number of cases, determine which row has the largest frequency of occurrence, and predict this value. This means that

$$E_{TR} = \text{Grand total} - \text{Maximum (Row total)}.$$

For each of the columns of the table,

$$E_{C_j} = \text{Column total} - \text{Maximum count in that column}.$$

Then

$$E_C = \sum_{j=1}^{c} E_{C_j}.$$

Using these values for the number of errors, λ can be computed.

As noted above, there is a different λ if the rows are being used to predict the column values. Table 11.9 of Example 11.3.1 is used in the following example to determine λ when column values are being predicted.

Example 11.3.3 Prediction of Provincial Political Preference from Federal Political Preference

Using the data in Table 11.9, suppose that the respondent's provincial political preference is being predicted. Using only the column totals, the best prediction would be that the respondent is an NDP supporter. Of the 553 total respondents, more say they will vote NDP in a provincial election than say they will support any other party. If there is no knowledge of the respondent's federal political preference, then there are $553 - 316 = 237$ *errors of prediction if it is predicted that any randomly selected respondent will vote NDP.*

Now suppose that the respondent's federal political preference is known. Take the first row of federal Liberal supporters. Among these 162 respondents, there are 80 provincial Liberal supporters, more than for any other party. If it is predicted that all the 162 federal Liberal supporters will also *support the provincial Liberal party, then there are* $162 - 80 - 82$ *errors of prediction, and this is the prediction that results in the least prediction errors. In the second row, for the 254 federal NDP supporters, the best prediction would be that all of these also support the NDP at the provincial level. Since 244 of these 254 do support the provincial NDP, there would be only* $254 - 244 = 10$ *errors of prediction in this row. Finally, for the 137 federal Conservative party supporters, the best prediction would be that each of these supports the Conservative party provincially. Then there would be* $137 - 118 = 19$ *errors of prediction, fewer errors than if support for any other provincial party is predicted.*

Using the information concerning federal political preference, the number of errors of prediction is $82 + 10 + 19 = 111$. *This is a reduction from 237 errors of prediction when federal political preference is not known. Since* λ *is the proportional reduction in error,*

$$\lambda = \frac{237 - 111}{237} = \frac{126}{237} = 0.532.$$

That is, there has been a reduction in the number of errors of prediction of 0.532 or 53.2%.

The last example shows that λ may be different when predicting column values than when predicting row values. While the method is generally the same, when predicting column values the number of errors of prediction for each row is used. The general formula for λ when predicting the column values is

$$\lambda_C = \frac{E_{TC} - E_R}{E_{TC}}$$

where E_{TC} is the number of errors of prediction for the column values, when the row values are not taken into account. For each of the r rows, let the number of errors of prediction be E_{R_i} where $i = 1, 2, \ldots, r$. Then let E_R be the sum of these E_{R_i}. That is, E_R is the total number of errors of prediction when the rows are used to predict the likely column results. Then

$$\lambda_C = \frac{E_{TC} - E_R}{E_{TC}}$$

where λ_C is the value of λ when the column values are being predicted.

The formula for each of E_{TC} and E_R can also be given. For E_{TC}, the method was to take the total number of cases, determine which column has the largest frequency of occurrence, and predict this value. This means that

$$E_{TC} = \text{Grand total} - \text{Maximum (Column total)}.$$

For each of the rows of the table,

$$E_{R_i} = \text{Row total} - \text{Maximum count in that row.}$$

Then

$$E_R = \sum_{i=1}^{r} E_{R_i}.$$

Using these values for the number of errors, λ can be computed.

Asymmetric and Symmetric Lambda. Measures of association like λ, which differ depending on which of the two variables is being predicted, are termed asymmetric measures. All of the chi square based measures of association are symmetric, since the value of the chi square statistic does not change if the rows and columns are interchanged. There is also a symmetric λ which is sometimes given by computer programs. The formula for the symmetric λ is not given here. It lies between the values of λ_R and λ_C, although it is not a simple mean of the two asymmetric lambdas. For the relationship between provincial and federal political preference in Saskatchewan, the SPSS computer program gives the value $\lambda = 0.586$ for the symmetric lambda. This is between $\lambda_r = 0.629$ and $\lambda_c = 0.532$.

Dependent and Independent Variables. The asymmetric lambdas introduced in this section can be used to illustrate the difference between **dependent** and **independent** variables. This distinction has not been used so far in this textbook, but is important in the regression model later in this chapter.

The variable which is being predicted is often termed the **dependent** variable, and the variable being used to predict the dependent variable is called the **independent** variable. In Example 11.3.1, the column variable provincial political preference was being used to predict the row variable, federal political preference. In this example, federal political preference was the dependent variable, and the column variable, provincial political preference was the independent variable. This does not necessarily mean that the independent variable causes, or determines values of, the dependent variable, although that is sometimes the case. Rather, the distinction between dependent and independent variables is a statistical one. The independent variable is the variable which is being used to explain one or more other variables. Each of the variables that is being explained is called a dependent variable.

In Example 11.3.3 the variables are reversed, and federal political preference is being used as the independent variable. This independent variable is being used to predict provincial political preference. This meant that provincial political preference is being treated as the dependent variable for this example. In Example 11.3.2, sex was treated as the independent variable, with opinion as the dependent variable. Sex may be a cause of different opinions, and in this case it makes no sense to reverse the direction of influence, making sex depend on opinion.

Summary. The proportional reduction in error approach to determining association between two variables is quite a different approach from that used for the chi square based measures of association. As shown in the examples here, λ can be computed for variables which are measured at no more than the nomimal level. While lambda has some major weaknesses as a measure, it does help to understand one way in which variables relate to each other. That is, λ shows how the values of one variable can be used to assist in reducing the number of prediction errors when predicting values of another variable.