Chapter 13

Analysis of Nominal and Ordinal Variables

13.1 Introduction

This chapter presents association measures of nominal and ordinal variables. In this context, these measures are used to explore the relationship between the variables. They are also used in the variable reduction process. Throughout the presentation, we follow the notation and formulations given in Agresti (2002).

We first present the details of the main tool used in the analysis of nominal and ordinal variables: contingency tables. Then we present the different measures of association between the variables.

13.2 Contingency Tables

A fundamental tool in the analysis of nominal and ordinal variables is what is known as contingency tables. The result of the analysis of contingency tables is a set of measures of the *association* between variables. Therefore, in the context of data preparation procedures, we can use these results in the following two areas.

- To reduce the number of independent variables by removing those that do not show reasonable association with the dependent variable.
- To compare the distribution of variables in two or more samples to make sure that the model training and validation partitions are not biased with respect to any of their variables. In other words, to make sure that the variables are *not* associated with the sampling process itself.

Table 13.1 Gender distribution of mailing campaign results.

	Response status		
Gender	*Yes*	*No*	*Total*
Female	587	18,540	19,127
Male	987	22,545	23,532
Total	1,574	411,085	42,659

Contingency tables are simply the *counts* of cross-tabulation of two or more nominal or ordinal variables. Therefore, when similar analysis is to be performed on continuous variables, binning may be necessary.

Table 13.1 shows the gender distribution of the response to a mail campaign. The purpose of the analysis is to investigate the level of association between the response behavior and gender. If the analysis reveals that gender does not play a significant role in deciding the response, then it may make sense to remove this variable from the mining view.

In Table 13.1, instead of the Gender variable, we could have used the two partitions used for training and validating a model. In this case, we could rephrase the question as: Do the two partitions represent the same population with respect to response rate? If the result of the analysis is No, then we have to reconsider our sampling approach.

We begin the analysis by setting the mathematical notation and the definition of the measures of association.

13.3 Notation and Definitions

We denote the variable with the categories spanning the columns as the Y variable and that spanning the rows of the table as the X variable. Both variables are discrete (i.e., nominal or ordinal) and may have more than two categories. Table 13.2 shows the notation used for the number of records in each cell of the table. The table contains I rows (levels of variable X) and J columns (levels of variable Y).

As shown in Table 13.2, the total number of records in row i is denoted n_{i*}, and the total number of records in column j is n_{*j}. The total number of records in the dataset is n, which is also equal to the sum of the row totals, the column totals, or the cell totals, that is,

$$n = \sum_{i=1}^{I} n_{i*} = \sum_{j=1}^{J} n_{*j} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}. \tag{13.1}$$

The cell proportion for the cell ij is defined as

$$p_{ij} = \frac{n_{ij}}{n}. \tag{13.2}$$

Table 13.2 Contingency table notation.

	Y					
X	y_1	$\cdots$	y_j	$\cdots$	y_J	*Total*
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1*}
$\vdots$						$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i*}
$\vdots$						$\vdots$
x_I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I*}
Total	n_{*1}	$\cdots$	n_{*j}	$\cdots$	n_{*J}	n

Next, we discuss a set of metrics for the association between the X and Y variables. We denote Y the *response* variable and X the *independent* variable. In addition, since the response variable Y, in most cases, has only two categories, we label one of these two categories *success* and the other *failure*. In cases where the term *success* has a clear business meaning, this definition will be most convenient. In all other cases, as will be shown later, it is immaterial which label is assigned to which category.

In addition to defining measures of association, sometimes it is important to make statistical inferences about that measure. This means that we would like to calculate the confidence intervals. The procedure in these cases is always the same. It consists of the following three steps.

1. Calculate the variance of the measure.
2. Determine the probability distribution that this measure should follow. In most cases, the cell counts (n_{ij}) follow either a binomial or Poisson distribution. When the response variable is binary and the total sample size is fixed, then the cell count follows a binomial distribution. On the other hand, when the X variable represents the count of an event, such as the number of times a customer uses a credit card per month, the cell count follows a Poisson distribution. In either case, when the sample size is large (> 30), both distributions approach that of the normal distribution. When examining measures that represent the ratio of two cell counts, the measure usually follows the χ^2 distribution.
3. The size of the confidence interval is calculated using the properties of the variable distribution and the standard deviation, which is sometimes called *standard error.*

Before introducing the different measures, we present a SAS macro that extracts the contingency table for any two variables in a dataset (see Table 13.3). The macro that follows is based on `PROC FREQ`, which is designed to do exactly that.

Table 13.3 Parameters of the macro `ContinMat()`.

Header	`ContinMat(DSin, Xvar, Yvar, ContTable);`
Parameter	*Description*
`DSin`	Input dataset
`XVar`	X variable
`YVar`	Y variable
`ContTable`	The output contingency table of X versus Y

```
%macro ContinMat(DSin, Xvar, Yvar, ContTable);
   proc freq data=&DSin noprint;
     tables &Xvar * &Yvar / out=&ContTable;
   run;
   Data &ContTable;
    set &ContTable;
    keep &Xvar &Yvar Count;
   run;
   %mend;
```

The code uses the output of `PROC FREQ` and stores it in the dataset `ContTable` the X and Y variables, as well as the count of their cells.

13.4 Contingency Tables for Binary Variables

We first present the analysis of contingency tables when both the X and Y variables have two levels, that is, binary variables, which is called *two-way contingency tables.*

The following are typical examples of cases when both X and Y are binary.

- Response behavior (Yes/No) in two samples or partitions
- Response behavior(Yes/No) with respect to a possible binary predictor, such as the presence or absence of some attribute (having a credit card, purchased a particular product, or responded to a certain campaign)
- The possible association between two independent binary variables, with the intention of removing one of them if they are strongly associated

In the list we use the generic term *response* to denote either actual response or any similar binary status, such as the credit status (Good/Bad), and profit level (High/Low).

The following subsections present several measures of association and their SAS implementation. We should note that `PROC FREQ` calculates all these measures.

However, in our presentation, we calculate some of the values from their original expressions to clarify the concepts.

13.4.1 Difference in Proportion

Table 13.1 provides the response behavior versus gender for a marketing mailing campaign. In this case, it is easy to label the observations for which the response to the mailing campaign was positive as the success event. The ratio of success rate for both males and females can then be calculated, as shown in Table 13.4.

Table 13.4 Success rate for each gender.

Gender	*Success rate*
Female	0.0307
Male	0.0419

The table shows that the difference in success rate between males and females is $0.0419 - 0.0307 = 0.0113$. The question now is: How significant is this difference? Can one infer that the campaign success rate with males is different from that with females?

To answer this question, we need to calculate the standard deviation of the difference in proportion of success. We denote the proportion of success in each of the two categories of the variable X by p_1 and p_2, such that

$$p_1 = n_{11}/n_{1*}, \tag{13.3}$$

and

$$p_2 = n_{12}/n_{2*}. \tag{13.4}$$

Note that we use the index 1 in the Y variable for the success event. When the cell count in the two rows is two independent binomial samples, then the standard deviation of the difference in proportion $p1 - p$, denoted $\hat{\sigma}(p_1 - p_2)$, is given by

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_{1*}} + \frac{p_2(1-p_2)}{n_{2*}}}. \tag{13.5}$$

Equation 13.5 shows that the standard error (standard deviation) decreases as n_{1*} and n_{2*} increase. When both n_{1*} and n_{2*} are large, then the distribution of the standard error $\hat{\sigma}(p_1 - p_2)$ follows the normal distribution, and we calculate the $100(1-\alpha)\%$ confidence interval as

$$(p_1 - p_2) \pm z_{\alpha/2}\hat{\sigma}(p_1 - p_2), \tag{13.6}$$

where $z_{\alpha/2}$ denotes the standard normal percentile, having right-tail probability of $\alpha/2$. For example, a 95% confidence interval has an $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$.

In the mailing campaign example, we did not really have two independent samples for each gender. However, it is reasonable to assume that we knew in advance the gender of each customer who received mail. Therefore, we can assume that when examining the results of the campaign, we have drawn two independent samples of 19,127 females and 23,532 males. By substitution in Expression 13.6, we obtain the 95% confidence interval for the difference in proportion of response rate in the two genders as

$$(0.0438 - 0.0317) \pm (1.96)\sqrt{\frac{0.0438(1-0.0438)}{23532} + \frac{0.0317(1-0.0317)}{19127}}.$$

The calculation gives the confidence interval between (0.0094) and (0.0131). Because both the upper and lower limits of the confidence interval are positive, we can infer that there *is* a significant difference between the rate of response in men and women. It is interesting to observe that the result is independent of our choice of the category of the variable Y as the success event. When we define the success event as `Response=No`, we obtain exactly the same result.

The SAS implementation of the preceding procedure assumes that we already have the contingency table, which could be obtained using the macro `ContinMat()` implemented in Section 13.3. The following macro, `PropDiff()`, calculates the upper and lower limit of the confidence interval for the difference in proportion (see Table 13.5).

Table 13.5 Parameters of macro `PropDiff()`.

Header	`PropDiff(ContTable, Xvar, Yvar, Alpha, M_Prop, M_Upper, M_Lower);`
Parameter	*Description*
`ContTable`	Input contingency table
`Xvar`	X variable name
`Yvar`	Y variable name
`Alpha`	Used to determine the confidence level of $(1\text{-}\alpha/2)$%
`M_Prop`	Difference in proportion (absolute value)
`M_Upper`	Resulting upper limit of the confidence interval
`M_Lower`	Resulting lower limit of the confidence interval

Step 1

Sort the contingency table using both the X and Y variables to guarantee the meaning of the different entries.

```
proc sort data=&ContTable;
 by &Xvar &Yvar;
run;
```

Step 2

Transform the entries of the contingency table into macro variables.

```
data _NULL_;
 set &ContTable;
 call symput ("n_"||left(_N_), COUNT);
run;
```

Step 3

Substitute into Equation 13.6.

```
%let N1star=%eval(&N_1+&N_2);
%let N2star=%eval(&N_3+&N_4);
%let P1=%sysevalf(&N_1/&N1star);
%let P2=%sysevalf(&N_3/&N2star);
%let P1P2=%sysfunc(abs(&p1-&P2));
%let sigma=%sysfunc(sqrt(((&P1*(1-&P1))/&N1star)
                        +((&P2*(1-&P2))/&N2star)));
%let &M_Prop = &P1P2;
%let &M_Upper=%sysevalf(&p1p2
             + &sigma * %sysfunc(probit(1-&alpha/2)));
%let &M_Lower=%sysevalf(&p1p2
             - &sigma * %sysfunc(probit(1-&alpha/2)));
%mend;
```

To demonstrate the use of the macro PropDiff(), we create a contingency table using the following DATA step:

```
DATA contingency;
 INPUT Count Gender $ Response $;
DATALINES;
18540 Female N
587   Female Y
22545 Male   N
987   Male   Y
;
RUN;
```

The following code shows how to call the macro and print the upper and lower limits to the SAS Log.

```
%let ContTable=Contingency;
%let Xvar=Gender;
%let Yvar=Response;
%let Alpha=0.05;
%let Prop=;
%let Upper=;
```

```
%let Lower=;
%PropDiff(&ContTable, &Xvar, &Yvar, &Alpha,
          Prop, Upper, Lower);
%put *********** Prop. Diff.= &Prop;
%put *********** Lower Limit= &Lower;
%put *********** Upper Limit= &Upper;
```

13.4.2 The Odds Ratio

Recall the results of the marketing campaign as shown in Table 13.6.

Table 13.6 Gender distribution of mailing campaign results.

	Response status		
Gender	*Yes*	*No*	*Total*
Female	587	18,540	19,127
Male	987	22,545	23,532
Total	1,574	411,085	42,659

We denote the *probability* of success of the marketing effort in the case of female customers given in the first row π_1. Similarly, we denote the probability of success for male customers, in the second row, π_2. We can determine the probability of *failure* for the first and second rows as $(1 - \pi_1)$ and $(1 - \pi_2)$, respectively. These probabilities are, in principal, unknown. However, we may estimate their values using the ratios p_1 and p_2, as in Equations 13.3 and 13.4.

The *odds* of success in the case of female customers is defined as the ratio of the probability of success to that of failure:

$$\text{Odds}_1 = \frac{\pi_1}{1 - \pi_1}. \tag{13.7}$$

Similarly, the odds of success with male customers is defined as

$$\text{Odds}_2 = \frac{\pi_2}{1 - \pi_2}. \tag{13.8}$$

The *odds ratio* is then defined as the ratio between these two quantities as

$$\theta = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}. \tag{13.9}$$

Because the proportions are the sample *estimates* of the probabilities, we may write the odds ratio in terms of the proportions p_1 and p_2 as

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{13.10}$$

The odds ratio determines the odds of success in row 1 relative to those of row 2. It is always a positive number. Interchanging the rows of the contingency table results in the inverse value of the odds ratio; that is, it changes the value of θ to $1/\theta$. The farther the value of θ from 1.0, in either direction, the more association there is between the variables X and Y.

In the example of Table 13.6 of the campaign responses, the odds ratio is $\theta = 0.723$, which, as in the case of the difference of proportion, shows that there is an association between the response rate and the gender. It should be noted that although computer programs can calculate θ to many significant digits, there is no need to do so because we are using the sample proportions as approximations to the real unknown probabilities. Therefore, interpreting θ for up to, say, three significant digits is usually sufficient.

To calculate the confidence interval for the odds ratio, θ, it is more convenient to use its logarithm, $\log(\theta)$. Therefore, we calculate the confidence interval for the logarithm of the odds ratio, and then use the exponential function to find the actual range. In large samples, the distribution of the logarithm of the odds ratio is normal with a mean of $\log(\theta)$ and a standard deviation, known as the *asymptotic standard error*, denoted ASE, of

$$ASE(\log\theta) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{13.11}$$

Equation 13.11 shows, as expected, that the ASE decreases as the cell counts increase. The $(1-\alpha)\%$ confidence interval of the logarithm of the odds ratio is then given by

$$\log\theta \pm z_{\alpha/2} ASE(\log\theta), \tag{13.12}$$

with $z_{\alpha/2}$ being the standard normal percentile, having right-tail probability of $\alpha/2$.

In the example of the response to the marketing campaign, the bounds of the 95% confidence interval of the logarithm of the odds ratio are −0.428 and −0.219. Using the natural exponential function, the actual bounds on the odds ratio are 0.652 and 0.803. Since this range for θ *does not* contain the unity 1.0, we infer that the true odds for males and females *are* different with a confidence of 95%.

Equation 13.12 fails when one of the cell counts is 0. In this case, the odds ratio is either 0 or ∞. To account for this case, the formulas for θ and ASE are modified by adding $\frac{1}{2}$ to each n_{ij},

$$\tilde{\theta} = \frac{(n_{11}+\frac{1}{2})(n_{22}+\frac{1}{2})}{(n_{12}+\frac{1}{2})(n_{21}+\frac{1}{2})}, \tag{13.13}$$

and

$$\widetilde{ASE}(\log\theta) = \sqrt{\frac{1}{(n_{11}+\frac{1}{2})} + \frac{1}{(n_{12}+\frac{1}{2})} + \frac{1}{(n_{21}+\frac{1}{2})} + \frac{1}{(n_{22}+\frac{1}{2})}}. \qquad (13.14)$$

The modification does not change the values of either θ or ASE when the cell counts are large, as in our example of Table 13.6. The macro `OddsRatio()` calculates the odds ratio and its confidence interval using these modified formulas (see Table 13.7).

Table 13.7 Parameters of macro `OddsRatio()`.

Header	`OddsRatio(ContTable, Xvar, Yvar, Alpha, M_Theta, M_Upper, M_Lower);`
Parameter	*Description*
`ContTable`	Input contingency table
`Xvar`	X variable name
`Yvar`	Y variable name
`Alpha`	Used to determine the confidence level of (1-α/2)%
`M_Theta`	Odds ratio
`M_Upper`	Resulting upper limit of the confidence interval
`M_Lower`	Resulting lower limit of the confidence interval

Step 1

Sort the contingency table by the categories of the X variable and the Y variable.

```
proc sort data=&ContTable;
 by &Xvar &Yvar;
run;
```

Step 2

Convert the count into macro variables that contain the cell counts.

```
data _NULL_;
 set &ContTable;
 call symput ("n_"||left(_N_), COUNT);
run;
```

Step 3

Calculate the odds ratio and the ASE using the modified ratio (just in case any of the cell counts is 0).

```
%let Theta=%sysevalf((&N_1+0.5)*(&N_4+0.5)/
                       ((&N_2+0.5)*(&N_3+0.5)));
%let ASE_log=%sysfunc(sqrt(1/(&N_1+0.5)+ 1/(&N_2+0.5)
                            +1/(&N_3+0.5)+ 1/(&N_4+0.5) ));
```

Step 4

Calculate the confidence interval for the log of the odds ratio, and use the exponential function to obtain the actual limits.

```
%let LogT=%sysfunc(log(&Theta));
%let LogU= %sysevalf(&LogT + &ASE_Log *
                       %sysfunc(probit(1-&alpha/2)));
%let LogL= %sysevalf(&LogT - &ASE_log *
                       %sysfunc(probit(1-&alpha/2)));
%let &M_Theta = &Theta;
%let &M_Upper = %sysfunc(exp(&LogU));
%let &M_Lower = %sysfunc(exp(&LogL));
%mend;
```

13.4.3 The Pearson Statistic

The *Pearson Chi-squared statistic* is another measure that tests the association between the variables X and Y, by comparing the actual cell counts with the *expected* counts, under the assumption of independence. When the resulting categories of the variable Y are independent of those of the variable X, then we expect that the cell counts reflect that, such that the probability of success in a row is independent of the row itself.

Let's demonstrate this idea using a simple example. Consider the data in Table 13.8, which shows the results of a marketing campaign in terms of the credit card type that the customer owns. Table 13.8 shows that the probability of response for the Visa card owners is (8/100) = 8% and for the MasterCard owners is (40/200) = 20%. If the response behavior is independent of the credit card type, both rows should show equal response rate. This in turn would have translated to the cell count of the Yes category being proportional to the sample size for each card group.

Table 13.8 Response versus credit card type.

	Response status		
Credit card	*Yes*	*No*	*Total*
VISA	8	92	100
MasterCard	40	160	200
Total	48	252	300

Since we have a total of 48 `Yes` responders, they should be divided according to the ratio of the total count of each group of card owners, that is 16 for Visa owners and 32 for MasterCard owners. A similar argument could be made for the case of the `No` responders. These cell counts are the *expected* cell counts under the assumption of independence between response result and card type. Table 13.9 shows these counts in parentheses under the actual counts.

Table 13.9 Actual expected counts of response status versus credit card type.

	Response status		
Credit card	*Yes*	*No*	*Total*
Visa	8 (16)	92 (84)	100
MasterCard	40 (32)	160 (168)	200
Total	48	252	300

The expected count of the cell ij, denoted μ_{ij}, is in fact calculated as

$$\mu_{ij} = \frac{n_{i*}n_{*j}}{n}. \tag{13.15}$$

The Pearson Chi-squared statistic is defined as

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}, \tag{13.16}$$

with the summation being over all the cells of the contingency table. It is the sum of the squared difference between the actual and expected cell counts, normalized by the expected counts. For example, using the values in Table 13.9, we calculate X^2 as

$$X^2 = \frac{(8-16)^2}{16} + \frac{(92-84)^2}{84} + \frac{(40-32)^2}{32} + \frac{(160-168)^2}{168} = 7.143.$$

When the cell counts are large, the X^2 follows a χ^2 distribution with degrees of freedom (df) of $(J-1)(I-1)$. Therefore, in the case of contingency tables with two binary variables, $df = 1$. The assumption of large cell counts is usually considered valid for modest values, such as when $\mu_{ij} \geq 5$, which is almost the case in all real business data. We use the χ^2 distribution to find the probability of independence, such that

$$Pr(\text{independence}) = \chi^2(X^2, df). \tag{13.17}$$

The value of $Pr()$ could be calculated using the SAS function `PROBCHI(.,.)`, which taks the values of X^2 and df as arguments. To implement the Pearson Chi-squared statistic and the confidence test, we could either program it using the original

Table 13.10 Parameters of meters of macro `PearChi()`.

Header	`PearChi(DSin, Xvar, Yvar, M_X2, M_pvalue);`
Parameter	*Description*
`DSin`	Input dataset
`Xvar`	X variable name
`Yvar`	Y variable name
`M_X2`	Pearson Chi-squared statistic, X^2
`M_pvalue`	p-value of the χ^2 test of the Pearson statistic

formulas or use `PROC FREQ`, which calculates it along with other measures. The macro `PearChi()` uses `PROC FREQ`, and extracts from its results the probability (p-value) of the χ^2 test of the Pearson statistic (see Table 13.10).

```
%macro PearChi(DSin, Xvar, Yvar, M_X2, M_pvalue);
PROC FREQ data =&DSin NOPRINT;
 TABLES &Xvar * &Yvar/chisq;
 OUTPUT All out=temp_chi chisq;
RUN;

/* Extract the P-value of the Chi square test */
Data _Null_;
set Temp_chi;
call symput("Mpvalue", P_PCHI);
call symput("MX2",_PCHI_);
run;
%let &M_Pvalue=&Mpvalue;
%let &M_X2 =&MX2;
proc datasets library=work nolist;
  delete temp_chi;
quit;
%mend;
```

Let us demonstrate the use of the macro with a simple dataset. The following code generates a dataset with 100 records of two variables, Gender (Female/Male) and Response (Yes/No). The values of the categories of Gender and Response are assigned at random using the SAS uniform distribution random number generator function `RanUni(.)`.

```
data test;
length Response $3. Gender $6. ;
 do i=1 to 100;
  if ranuni(0)>0.8 then Response='Yes';
     else Response ='No';
  if ranuni(0)>0.6 then Gender ='Male';
     else Gender ='Female';
 output;
 end;
 drop i;
run;
```

Finally, we call the macro and display the result of the test in the SAS Log using the following code:

```
%let DSin=test;
%let Xvar=Gender;
%let Yvar=Response;
%let X2=;
%let pvalue=;

%PearChi(&DSin, &XVar, &Yvar, x2, pvalue);
%put The Pearson chi-square Stat.= &X2;
%put Probability of Independence = &pvalue;
```

The small p-value indicates that the two variables *are* associated and the large p-value confirms their independence.

13.4.4 The Likelihood Ratio Statistic

This statistic is similar to the Pearson Chi-squared statistic, except it is derived using the maximum likelihood method. The statistic is defined as

$$G^2 = 2 \sum n_{ij} \log \left(\frac{n_{ij}}{\mu_{ij}} \right), \qquad (13.18)$$

with the summation taken over all the cells of the contingency table. The G^2 statistic also follows the χ^2 distribution, with $(J-1)(I-1)$ degrees of freedom, and is called the *likelihood ratio Chi-squared statistic*. The Pearson statistic and G^2 usually provide similar results for large datasets and result in the same conclusion.

The G^2 statistic and its Chi-squared test are also calculated by `PROC FREQ`. This will be the basis of our implementation of the macro `LikeRatio()`, the macro identical to that used to calculate the Pearson statistic test, with the exception that it extracts the likelihood ratio statistic and its p-value (see Table 13.11).

```
%macro LikeRatio(DSin, Xvar, Yvar, M_G2, M_pvalue);
proc freq data =&DSin noprint;
 tables &Xvar * &Yvar/chisq;
 output All out=temp_chi chisq;
run;
```

Table 13.11 Parameters of macro `LikeRatio()`.

Header	`LikeRatio(DSin, Xvar, Yvar, M_G2, M_pvalue);`
Parameter	*Description*
`DSin`	Input dataset
`Xvar`	X variable name
`Yvar`	Y variable name
`M_G2`	The likelihood ratio statisic, G^2
`M_pvalue`	p-value of the χ^2 test of the G^2 statistic

```
/* Extract the G2 and it sp-value */
Data _Null_;
set Temp_chi;
call symput("Mpvalue", P_LRCHI);
call symput("MG2", _LRCHI_);
run;
%let &M_Pvalue=&Mpvalue;
%let &M_G2 =&MG2;
proc datasets library=work nolist;
  delete temp_chi;
quit;
%mend;
```

Again, a small p-value indicates that the two variables are associated and the large value indicates their independence.

The two macros `PearChi()` and `LikeRatio()` could have been combined into one macro because the option `/CHISQ` of the `TABLES` statement of `PROCFREQ` calculates both statistics and their p-values. We implemented them separately only for clarity.

13.5 Contingency Tables for Multicategory Variables

The extension of the evaluation of the association methods among the variables of the contingency tables to the case of variables with several categories is straightforward. However, because there is no particular category of the Y variable that could be denoted as `Success`, we are restricted to the methods that do not depend on such a definition. Therefore, the difference in proportion and the odds ratio are not applicable.

Equations 13.16 and 13.18, defining the Pearson Chi-squared statistic and the likelihood ratio statistic, are valid for multicategory variables. Furthermore, the implementation of the two macros `PearChi()` and `LikeRatio()` also allows for either or both of the X and Y variables to have more than two categories.

Table 13.12 Parameters of macro `ContnAna()`.

Header	`ContnAna(DSin, VarX, VarY, ResDS);`
Parameter	*Description*
`DSin`	Input dataset
`VarX`	*X* variable name
`VarY`	*Y* variable name
`ResDS`	Results dataset

Therefore, we do not need to provide separate implementations for the case of multicategory nominal variables. However, we present a macro that calculates *all* the measures of association with `PROC FREQ` and extract them with their description to a dataset (see Table 13.12). This macro should replace the macros `PearChi()` and `LikeRatio()`.

In addition to the Pearson statistic and the likelihood ratio, this macro extracts three more statistics that measure association. They are the Mantel-Haenszel, the Phi coefficient, and Cramer's V. These statistics are described in the SAS online documentation of `PROC FREQ`.

The macro calls `PROC FREQ`, stores all the association measures in the dataset `temp_chi`, extracts these statistics, and stores them in the results dataset.

```
%macro ContnAna(DSin, VarX, VarY, ResDS);
/* Calculation of measures of association between
   two categorical variables (VarX, VarY)
   in a dataset (DSin) using PROC FREQ and
   arranging the results in a dataset (ResDS) */

proc freq data =&DSin noprint;
 tables &VarX * &VarY/chisq;
 output All out=temp_chi chisq;
run;

proc sql noprint;
 create table &ResDS
        (SAS_Name char(10), Description char(50), Value num);
 select _PHI_, P_MHCHI, P_LRCHI, P_PCHI, N, _MHCHI_
      , _LRCHI_, DF_MHCHI, DF_LRCHI, DF_PCHI ,_CRAMV_
      ,_CONTGY_ ,_PCHI_
    into :PHI, :P_MHCHI, :P_LRCHI, :P_PCHI, :N, :MHCHI
       , :LRCHI, :DF_MHCHI, :DF_LRCHI, :DF_PCHI, :CRAMV
       , :CONTGY, :PCHI
 from temp_chi;
insert into &ResDS
```

```
values("N", "Number of Subjects in the Stratum",&N)
values("_PCHI_","Chi-Square",&PCHI)
values("DF_PCHI","DF for Chi-Square",&DF_PCHI)
values("P_PCHI","P-value for Chi-Square",&P_PCHI)
values("_MHCHI_","Mantel-Haenszel Chi-Square",&MHCHI)
values("DF_MHCHI","DF for Mantel-Haenszel Chi-Square",
       &DF_MHCHI)
values("P_MHCHI","P-value for Mantel-Haenszel Chi-Square",
       &P_MHCHI)
values("_LRCHI_","Likelihood Ratio Chi-Square",&LRCHI)
values("DF_LRCHI","DF for Likelihood Ratio Chi-Square",
       &DF_LRCHI)
values("P_LRCHI","P-value for Likelihood Ratio Chi-Square",
       &P_LRCHI)
values("_PHI_","Phi Coefficient",&PHI)
values("_CONTGY_","Contingency Coefficient",&CONTGY)
values("_CRAMV_","Cramer's V",&CRAMV)
;
quit;
proc datasets library=work nolist;
  delete temp_chi;
quit;
%mend;
```

13.6 Analysis of Ordinal Variables

When one or both of the row or column variables, X and Y, are ordinal, the methods described in the previous sections are not appropriate. This is because the measures of association, such as X^2 and G^2, are based on the assumption that there is no ordering within the different categories. Adding an ordinal scale to the categories introduces more information in the data that should be used.

In analyzing ordinal as well as continuous variables, two types of models are usually proposed: *parametric* and *nonparametric*. Parametric models assume that the variables follow a particular probability distribution. On the other hand, nonparametric models make no such an assumption and are sometimes called *distribution-free* models.

Before we present the definition of the different measures of association between ordinal variables, we have to discuss the subject of *scores*.

When the categories of either X or Y, or both, are set on an ordinal scale, we call the values of this scale for each category the *scores.* Let us demonstrate this by an example.

Table 13.13 shows the results of cross-tabulating the average monthly rate of credit card usage versus the event of default on the card payment. To express the ordinal nature of the rate of credit card usage, we assign a scale to the different ranges. For example, we assign $1 - 2 \rightarrow 1$, $3 - 5 \rightarrow 2$, and so on. These values are the assigned

Table 13.13 Credit card usage rate versus credit default.

	Credit status			
Average usage	*Good*	*Default*	*Total*	*(%) Default*
1–2	35,784	1,469	37,253	3.94
3–5	45,874	1,457	47,331	3.08
5–9	45,741	2,897	48,638	5.96
9–15	8,547	451	8,998	5.01
>15	6,987	359	7,346	4.89
Total	142,933	6,633	149,566	4.64

scores. Whenever scores are assigned, we have to assign *both* row and column scores. In this example of credit card default, we may assign a score of 1 to the status `good` and 5 to the status `Default`.

For most datasets, the choice of scores has little effect on the final results as long as the data is, more or less, evenly distributed over the different categories. In the preceding example, this is not really the case, because we have fewer observations for the high range of credit card use. Therefore, it is always a good practice to check the quality of the assigned scores by trying different assignment schemes.

For ordinal variables, parametric measures of association are computed using the values of the scores directly. Nonparametric measures work instead with their ranks. In most cases, the scores assigned to ordinal variables mean their ranks. Therefore, given the choice, we prefer to use nonparametric measures.

First, some notation. The row scores are denoted R_i for the score of row i, and column scores are denoted C_j for the score of column j. Furthermore, the average row score is denoted $\bar{R}$ and the average column score is denoted $\bar{C}$. Using this notation, the *Pearson correlation coefficient* is defined as

$$r = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}, \tag{13.19}$$

where the terms ss_{rc}, ss_r, and ss_c are defined as

$$ss_r = \sum_{i=1}^{I} \sum_{j-1}^{J} n_{ij}(R_i - \bar{R})^2, \tag{13.20}$$

$$ss_c = \sum_{i=1}^{I} \sum_{j-1}^{J} n_{ij}(C_i - \bar{C})^2, \tag{13.21}$$

and

$$ss_{rc} = \sum_{i=1}^{I} \sum_{j-1}^{J} n_{ij}(R_i - \bar{R})(C_i - \bar{C}). \tag{13.22}$$

The value of r is between -1 and $+1$. Values close to 0, either positive or negative, indicate lack of correlation between the variables X and Y, and larger values (near -1 or $+1$) are indicators of strong correlation (or anticorrelation in the case of negative r).

The variance of the correlation coefficient r is also calculated by PROC FREQ. However, because this is a parametric measure, the confidence intervals, which can be calculated using the variance and statistical tests, are meaningful only when both the variables X and Y are normally distributed. When this condition cannot be guaranteed in real-life data, we prefer to use nonparametric measures.

PROC FREQ calculates the Pearson correlation coefficient using the option MEASURES in the TABLES and in the OUTPUT statements when we wish to store the results in a dataset. The generated dataset will contain the coefficient in the variable _PCORR_. The following macro wraps PROC FREQ and extracts r (see Table 13.14).

Table 13.14 Parameters of the macro ContPear().

Header	ContPear(DSin, XScore, YScore, M_R);
Parameter	*Description*
DSin	Input dataset
XScore	X variable name (scores)
YScore	Y variable name (scores)
M_R	Output value of r

```
%macro ContPear(DSin, XScore, YScore, M_R);

proc freq data=&DSin noprint;
 tables &XScore*&YScore / measures;
 output measures out=temp_r;
run;

data _NULL_;
 set temp_r;
 call symput("r", _PCORR_);
run;

%let &M_r = &r;

/* clean workspace */
proc datasets nodetails;
 delete temp_r;
quit;

%mend;
```

We now turn our attention to a nonparametric correlation coefficient: the *Spearman rank correlation coefficient*. Basically, it has the same definition as the Pearson coefficient, but it uses the ranks instead of the scores. It can be defined in terms of the ranks of the values of the $X$and Y variables directly as

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}, \tag{13.23}$$

where R_i and S_i are the ranks of X and Y and $\bar{R}$ and $\bar{S}$ are their average ranks, respectively. Note that r_s can also be expressed in terms of the counts of the contingency matrix. The equations in this case are more complex. These expressions can be found in the SAS help on PROC FREQ. The advantage of the Spearman's coefficient is that its significance test is performed simply using the following statistic:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}, \tag{13.24}$$

which follows the Student t-distribution with $n - 2$ degrees of freedom (because two degrees of freedom were lost in calculating $\bar{R}$ and $\bar{S}$).

The biggest advantage of using r_s is that when its significance test shows that there is a correlation, then this result is true regardless of the underlying distribution of the variables.

To get the value of r_s, we can modify the macro ContPear() by making the DATA step read the variable _SCORR_ from the temporary dataset, as shown in the following code of the macro ContSpear() (see Table 13.15).

Table 13.15 Parameters of the macro ContSpear().

Header	ContSpear(DSin, XScore, YScore, M_RS);
Parameter	*Description*
DSin	Input dataset
XScore	X variable name (scores)
YScore	Y variable name (scores)
M_RS	Output value of r_s

```
%macro ContSpear(DSin, XScore, YScore, M_RS);
/* Calculation of Spearman correlation coefficient
    for ordinal variables using the scores given
    in XScore, YScore. The result is stored in M_RS */
```

```
proc freq data=&DSin noprint;
 tables &XScore*&YScore / measures;
output measures out=temp_rs;
run;

data _NULL_;
 set temp_rs;
 call symput("rs", _SCORR_);
run;

%let &M_rs = &rs;

proc datasets nodetails;
 *delete temp_rs;
quit;

%mend;
```

13.7 Implementation Scenarios

This chapter presented several methods of measuring the associations for nominal and ordinal variables. One may summarize possible implementation scenarios as follows.

1. Faced with a large number of variables in a candidate mining view, we could reduce the number of independent variables by considering only those that show a reasonable level of association or correlation with the dependent variable.
2. We can also test the hypothesis that two datasets, such as the training and validation partitions of the mining view, have the same distribution and interrelationships of variables. We perform this analysis by finding the level of association between each of the key variables in the analysis and a dummy variable representing the dataset label. This is particularly useful in checking that the scoring view variables have the same distribution as those used in building the models before committing to the scores produced by these models and using them at face value.
3. During the exploratory data analysis (EDA) procedures, it is always required that we investigate the relationships among the variables. The macros presented in this chapter could prove useful in automating this process.

This Page Intentionally Left Blank