Chapter 26. The FREQ Procedure

EXAMPLES .................................................. 1284
Example 26.1 Creating an Output Data Set with Table Cell Frequencies . . . 1284
Example 26.2 Computing Chi-square Tests for One-Way Frequency Tables . 1287
Example 26.3 Computing Binomial Proportions for One-Way Frequency Ta-
bles .................................................. 1288
Example 26.4 Analyzing a 2x2 Contingency Table .............................. 1288
Example 26.5 Creating an Output Data Set Containing Chi-Square Statistics . 1293
Example 26.6 Computing Cochran-Mantel-Haenszel Statistics for a Strati-
fied Table ........................................... 1295
Example 26.7 Computing the Cochran-Armitage Trend Test ................. 1297
Example 26.8 Computing Friedman’s Chi-Square Statistic ................. 1300
Example 26.9 Testing Marginal Homogeneity with Cochran’s Q ........... 1302

REFERENCES ............................................ 1305
The FREQ procedure produces one-way to $n$-way frequency and crosstabulation (contingency) tables. For two-way tables, PROC FREQ computes tests and measures of association. For $n$-way tables, PROC FREQ does stratified analysis, computing statistics within, as well as across, strata. Frequencies and statistics can also be output to SAS data sets.

For one-way frequency tables, PROC FREQ can compute statistics to test for equal proportions, specified proportions, or the binomial proportion. For contingency tables, PROC FREQ can compute various statistics to examine the relationships between two classification variables adjusting for any stratification variables. PROC FREQ automatically displays the output in a report and can also save the output in a SAS data set.

For some pairs of variables, you may want to examine the existence or the strength of any association between the variables. To determine if an association exists, chi-square tests are computed. To estimate the strength of an association, PROC FREQ computes measures of association that tend to be close to zero when there is no association and close to the maximum (or minimum) value when there is perfect association. The statistics for contingency tables include:

- chi-square tests and measures
- measures of association
- risks (binomial proportions) and risk differences for $2 \times 2$ tables
- odds ratios and relative risks for $2 \times 2$ tables
- Cochran-Armitage test for trend and Jonckheere-Terpstra test
- tests and measures of agreement
- Cochran-Mantel-Haenszel statistics

PROC FREQ computes asymptotic standard errors, confidence intervals, and tests for measures of association and measures of agreement. Exact $p$-values and confidence intervals are available for various test statistics and measures. PROC FREQ also performs stratified analyses that compute statistics within, as well as across, strata for $n$-way tables. The statistics include Cochran-Mantel-Haenszel statistics and measures of agreement.

In choosing measures of association to use in analyzing a two-way table, you should consider the study design (which indicates whether the row and column variables are
dependent or independent), the measurement scale of the variables (nominal, ordinal, or interval), the type of association that each measure is designed to detect, and any assumptions required for valid interpretation of a measure. You should exercise care in selecting measures that are appropriate for your data.

Similar comments apply to the choice and interpretation of the test statistics. For example, the Mantel-Haenszel chi-square statistic requires an ordinal scale for both variables and is designed to detect a linear association. The Pearson chi-square, on the other hand, is appropriate for all variables and can detect any kind of association, but it is less powerful for detecting a linear association because its power is dispersed over a greater number of degrees of freedom (except for $2 \times 2$ tables).

Several SAS procedures produce frequency counts; only PROC FREQ computes chi-square tests for one-way to $n$-way tables and measures of association and agreement for contingency tables. Other procedures to consider for counting are TABULATE, CHART, and UNIVARIATE. When you want to fit models to categorical data, use a procedure such as CATMOD, GENMOD, LOGISTIC, PHREG, or PROBIT. For more information on selecting the appropriate statistical analyses, refer to Agresti (1996) or Stokes, Davis, and Koch (1995).

### Getting Started

#### Frequency Tables and Statistics

The FREQ procedure provides easy access to statistics for testing for association in a crossclassification table.

In this example, high school students applied for courses in a summer enrichment program: these included journalism, art history, statistics, graphics arts, and computer programming. The students accepted were randomly assigned to classes with and without internships in local companies. The following table contains counts of the students who enrolled in the summer program by gender and whether they were assigned an internship slot.

**Table 26.1. Summer Enrichment Data**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Internship</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>boys</td>
<td>yes</td>
<td>35</td>
</tr>
<tr>
<td>boys</td>
<td>no</td>
<td>14</td>
</tr>
<tr>
<td>girls</td>
<td>yes</td>
<td>32</td>
</tr>
<tr>
<td>girls</td>
<td>no</td>
<td>53</td>
</tr>
</tbody>
</table>

The SAS data set `SummerSchool` is created by inputting count data that corresponds to each cell of the table. The following DATA step statements create the SAS data set `SummerSchool`.

SAS OnlineDoc™: Version 7-1
The variable Gender takes the values ‘boys’ or ‘girls’, the variable Internship takes the values ‘yes’ and ‘no’, and the variable School takes the values ‘yes’ and ‘no’. The variable COUNT contains the number of students with the characteristics corresponding to the other variable values. The double at sign (@@) indicates that more than one observation is included on a single data line. In this DATA step, two observations are included on each line.

Researchers are interested in whether there is an association between internship status and summer program enrollment. The Pearson chi-square statistic is an appropriate statistic to assess the association in the corresponding \(2 \times 2\) table. The following PROC FREQ statements specify this analysis.

You specify the table for which you want to compute statistics with the TABLE statement. You specify the statistics you want to compute with options after a slash (/) in the TABLES statement.

```
proc freq data=SummerSchool order=data;
   weight count;
   tables Internship*School / chisq;
run;
```

The ORDER= option controls the order in which variable values are displayed in the rows and columns of the table. By default, the values are arranged according to alphanumeric order. If ORDER=DATA is specified, the data are displayed in the same order as they were encountered in the DATA step. Here, since ‘yes’ appears before ‘no’ in the data, ‘yes’ appears first in any table. Another option for controlling order is to use ORDER=FORMATTED to base the ordering on formatted values.

In the TABLES statement, Internship*School specifies a table comprised of rows of internship status and columns of program attendance. The WEIGHT statement is required when the input data are in count form. The variable specified in the WEIGHT statement identifies the count variable. Finally, the CHISQ option requests that chi-square statistics for assessing association be computed.

Figure 26.1 presents the cross-classification of Internship and School. In each cell, the values printed under the cell count are the table percentage, column percentage, and row percentage, respectively. For example, in the first cell, 63.21 percent of those offered courses with internships accepted them and 36.79 percent did not.
The SAS System

The FREQ Procedure

Table of Internship by School

<table>
<thead>
<tr>
<th>Internship</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>67</td>
<td>30.04</td>
<td>30.04</td>
<td>50.00</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>17.49</td>
<td>63.21</td>
<td>47.53</td>
<td>43.82</td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>67</td>
<td>30.04</td>
<td>30.04</td>
<td>50.00</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>22.42</td>
<td>57.26</td>
<td>52.47</td>
<td>42.74</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>60.09</td>
<td>60.09</td>
<td>100.00</td>
<td>223</td>
</tr>
</tbody>
</table>

Figure 26.1. Cross-Classification Table

The next table displays the statistics produced by the CHISQ option. The Pearson chi-square statistic is labeled ‘Chi-Square’ and has a value of 0.8189 with 1 degree of freedom. The associated p-value is 0.3655, which means that there is no significant evidence of an association between internship status and program acceptance. The other chi-square statistics have similar values and they are asymptotically equivalent. The Fisher’s Exact test takes the value p=0.4122 (two-tailed). The other statistics (Phi Coefficient, Contingency Coefficient, and Cramer’s V) are measures of correlation.

The FREQ Procedure

Statistics for Table of Internship by School

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>0.8189</td>
<td>0.3655</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>0.8202</td>
<td>0.3651</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>0.5899</td>
<td>0.4425</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>0.8153</td>
<td>0.3666</td>
</tr>
<tr>
<td>Fisher’s Exact Test (Left)</td>
<td></td>
<td>0.8513</td>
<td></td>
</tr>
<tr>
<td>(Right)</td>
<td></td>
<td>0.2213</td>
<td></td>
</tr>
<tr>
<td>(2-Tail)</td>
<td></td>
<td>0.4122</td>
<td></td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.0606</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.0605</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.0606</td>
<td></td>
</tr>
</tbody>
</table>

Sample Size = 223

Figure 26.2. Statistics Produced with the CHISQ Option
The analysis, so far, has ignored gender. However, it may be of interest to ask whether program acceptance is associated with internship status after adjusting for gender. You can address this question by doing an analysis of a set of tables, in this case, by analyzing the set consisting of one for boys and one for girls. The Cochran-Mantel-Haenszel statistic is appropriate for this situation: it addresses whether rows and columns are associated after controlling for the stratification variable. In this case, you would be stratifying by gender.

The FREQ statements for this analysis are very similar except that there is a third variable, Gender, in the TABLES statement. When you cross more than two variables, the two rightmost variables construct the rows and columns of the table, respectively, and the leftmost variables determine the stratification.

```
proc freq data=SummerSchool;
    weight count;
    tables Gender*Internship*School / chisq cmh;
run;
```

This execution of PROC FREQ first produces two individual tables, one for boys and one for girls. Chi-square statistics are produced for each individual table. Note that the chi-square statistic for boys is significant at the $\alpha = 0.05$ level of significance. Boys offered a course with an internship are more likely to accept than boys who are not.
The FREQ Procedure

Table 1 of Internship by School
Controlling for Gender=boys

<table>
<thead>
<tr>
<th>Internship</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Percent</td>
</tr>
<tr>
<td>no</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>25.71</td>
</tr>
<tr>
<td></td>
<td>65.85</td>
</tr>
<tr>
<td></td>
<td>48.21</td>
</tr>
<tr>
<td>yes</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>27.62</td>
</tr>
<tr>
<td></td>
<td>45.31</td>
</tr>
<tr>
<td></td>
<td>51.79</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>53.33</td>
</tr>
</tbody>
</table>

Statistics for Table 1 of Internship by School
Controlling for Gender=boys

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>4.2366</td>
<td>0.0396</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>4.2903</td>
<td>0.0383</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>3.4515</td>
<td>0.0632</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>4.1963</td>
<td>0.0405</td>
</tr>
<tr>
<td>Fisher’s Exact Test (Left)</td>
<td></td>
<td>0.9885</td>
<td></td>
</tr>
<tr>
<td>(Right)</td>
<td></td>
<td>0.0311</td>
<td></td>
</tr>
<tr>
<td>(2-Tail)</td>
<td></td>
<td>0.0467</td>
<td></td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.2009</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.1969</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.2009</td>
<td></td>
</tr>
</tbody>
</table>

Sample Size = 105

Figure 26.3. Frequency Table and Statistics for Boys

If you look at the individual table for girls, you see that there is no evidence of association for girls getting internship offers versus those who did not.
Table 2 of Internship by School
Controlling for Gender=girls

<table>
<thead>
<tr>
<th>Internship School</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>23</td>
<td>53</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.49</td>
<td>44.92</td>
<td>64.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30.26</td>
<td>69.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>69.70</td>
<td>62.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>10</td>
<td>32</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.47</td>
<td>27.12</td>
<td>35.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.81</td>
<td>76.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30.30</td>
<td>37.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>85</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.97</td>
<td>72.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Statistics for Table 2 of Internship by School
Controlling for Gender=girls

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>0.5593</td>
<td>0.4546</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>0.5681</td>
<td>0.4510</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>0.2848</td>
<td>0.5936</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>0.5545</td>
<td>0.4565</td>
</tr>
<tr>
<td>Fisher’s Exact Test (Left)</td>
<td></td>
<td>0.8317</td>
<td></td>
</tr>
<tr>
<td>(Right)</td>
<td></td>
<td>0.2994</td>
<td></td>
</tr>
<tr>
<td>(2-Tail)</td>
<td></td>
<td>0.5245</td>
<td></td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.0688</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.0687</td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.0688</td>
<td></td>
</tr>
</tbody>
</table>

Sample Size = 118

Figure 26.4. Frequency Table and Statistics for Girls

These individual table results demonstrate the occasional problems with combining information into one table and not accounting for information in other variables such as Gender. Figure 26.4 contains the CMH results. There are three summary (CMH) statistics: which one you use depends on whether your rows and/or columns have an ordering to them in \( r \times c \) tables. However, in the case of \( 2 \times 2 \) tables, ordering doesn’t matter and all three statistics take the same value. The CMH statistic follows the chi-square distribution under the hypothesis of no association, and here, it takes the value 4.0186 with 1 degree of freedom. The associated \( p \)-value is 0.0450, which indicates a significant association at the \( \alpha = 0.05 \) level.

Thus, when you adjust for the effect of gender in these data, there is an association between internship and program acceptance. But, if you ignore gender, no association is found. Note that the CMH option also produces other statistics, including estimates and confidence limits for relative risk and odds ratios for \( 2 \times 2 \) tables and the Breslow-Day Test. These results are not displayed here.
The FREQ Procedure

Summary Statistics for Internship by School
Controlling for Gender

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
<tr>
<td>3</td>
<td>General Association</td>
<td>1</td>
<td>4.0186</td>
<td>0.0450</td>
</tr>
</tbody>
</table>

Total Sample Size = 223

Figure 26.5.  Test for the Hypothesis of No Association

Agreement Study Example

Medical researchers are interested in evaluating the efficacy of a new treatment for a skin condition. Dermatologists from participating clinics were trained to conduct the study and to evaluate the condition. After the training, two dermatologists examined patients with the skin condition from a pilot study and rated the same patients. The possible evaluations are terrible, poor, marginal, and clear.

Table 26.2 contains the data.

Table 26.2.  Skin Condition Data

<table>
<thead>
<tr>
<th>Dermatologist 1</th>
<th>Terrible</th>
<th>Poor</th>
<th>Marginal</th>
<th>Clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrible</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Poor</td>
<td>5</td>
<td>10</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Marginal</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Clear</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>

The dermatologists’ evaluations of the patients are contained in the variables derm1 and derm2; the variable count is the number of patients given a particular pair of ratings. In order to evaluate the agreement of the diagnoses (a possible contribution to measurement error in the study), the kappa coefficient is computed. You specify the AGREE option in the TABLES statement and use the TEST statement to request a test for the null hypothesis that their agreement is purely by chance. You specify the keyword KAPPA to perform this test for the kappa coefficient. The results are shown in Figure 26.6.

```sas
data SkinCondition;
  input derm1 $ derm2 $ count;
datalines;
terrible terrible 10
terrible poor 4
terrible marginal 1
```

SAS OnlineDoc™: Version 7-1
The FREQ Procedure

Statistics for Table of derm1 by derm2

Simple Kappa Coefficient

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>0.3449</td>
</tr>
<tr>
<td>ASE</td>
<td>0.0724</td>
</tr>
<tr>
<td>95% Lower Conf Bound</td>
<td>0.2030</td>
</tr>
<tr>
<td>95% Upper Conf Bound</td>
<td>0.4868</td>
</tr>
</tbody>
</table>

Test of H0: Kappa = 0

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE under H0</td>
<td>0.0612</td>
</tr>
<tr>
<td>Z</td>
<td>5.6366</td>
</tr>
<tr>
<td>One-sided Pr &gt; Z</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Two-sided Pr &gt;</td>
<td>Z</td>
</tr>
</tbody>
</table>

Sample Size = 88

Figure 26.6. Agreement Study

The kappa coefficient has the value 0.3349, which indicates slight agreement between the dermatologists, and the hypothesis test confirms that you can reject the null hypothesis of no agreement. This conclusion is further supported by the confidence interval of (0.2030, 0.4868), which suggests that the true kappa is greater than zero. The AGREE option also produces Bowker’s test for symmetry and the weighted kappa coefficient, but that output is not shown.
Syntax

The following statements are available in PROC FREQ.

```
PROC FREQ < options > ;
    BY variables ;
    EXACT statistic-options < / computation-option > ;
    OUTPUT < OUT=SAS-data-set > options ;
    TABLES requests < / options > ;
    TEST options ;
    WEIGHT variable ;
```

The PROC FREQ statement is the only required statement for the FREQ procedure. If you specify the following statements, PROC FREQ produces a one-way frequency table for each variable in the most recently created data set.

```
proc freq;
   run;
```

The rest of this section gives detailed syntax information for the BY, EXACT, OUTPUT, TABLES, TEST, and WEIGHT statements in alphabetical order after the description of the PROC FREQ statement. Table 26.3 summarizes the basic functions of each statement.

**Table 26.3. Summary of PROC FREQ Statements**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BY</td>
<td>calculates separate frequency or crosstabulation tables for each BY group.</td>
</tr>
<tr>
<td>EXACT</td>
<td>requests exact tests for specified statistics.</td>
</tr>
<tr>
<td>OUTPUT</td>
<td>creates an output data set that contains specified statistics.</td>
</tr>
<tr>
<td>TABLES</td>
<td>specifies frequency or crosstabulation tables and requests tests and measures of association.</td>
</tr>
<tr>
<td>TEST</td>
<td>requests asymptotic tests for measures of association and agreement.</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>identifies a variable with values that weight each observation.</td>
</tr>
</tbody>
</table>

**PROC FREQ Statement**

```
PROC FREQ < options > ;
```

The PROC FREQ statement invokes the procedure. The following table lists the options available in the PROC FREQ statement. Descriptions follow in alphabetical order.
Table 26.4. PROC FREQ Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA=</td>
<td>specifies the input data set.</td>
</tr>
<tr>
<td>COMPRESS</td>
<td>begins the next one-way table on the current page even if the entire table does not fit on that page.</td>
</tr>
<tr>
<td>FORMCHAR=</td>
<td>specifies the outline and cell divider characters for the cells of the crosstabulation table.</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>suppresses all displayed output.</td>
</tr>
<tr>
<td>ORDER=</td>
<td>specifies the order to list the variable values.</td>
</tr>
<tr>
<td>PAGE</td>
<td>displays one table per page.</td>
</tr>
</tbody>
</table>

You can specify the following options in the PROC FREQ statement.

**COMPRESS**

begins display of the next one-way frequency table on the same page as the preceding one-way table if there is enough space to begin the table. By default, the next one-way table begins on the current page only if the entire table fits on that page. The COMPRESS option is not valid with the PAGE option.

**DATA=SAS-data-set**

names the SAS data set to be analyzed by PROC FREQ. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

**FORMCHAR (1,2,7) =’formchar-string’**

defines the characters to be used for constructing the outlines and dividers for the cells of contingency tables. The FORMCHAR= option can specify 20 different SAS formatting characters used to display output; however, PROC FREQ uses only the first, second, and seventh formatting characters. Therefore, the proper specification for PROC FREQ is FORMCHAR(1,2,7)= ’formchar-string’. The formchar-string should be three characters long. The characters are used to denote (1) vertical separator, (2) horizontal separator, and (7) vertical-horizontal intersection. You can use any character in formchar-string, including hexadecimal characters. If you use hexadecimal characters, you must put an x after the closing quote. For information on which hexadecimal codes to use for which characters, consult the documentation for your hardware.

Specifying all blanks for formchar-string produces tables with no outlines or dividers:

```
formchar (1,2,7)= ’ ’
```

If you do not specify the FORMCHAR= option, PROC FREQ uses the default

```
formchar (1,2,7)= ’|+-’
```

Refer to the CALENDAR, PLOT, and TABULATE procedures in the *SAS Procedures Guide* for more information on form characters.
Table 26.5. Formatting Characters Used by PROC FREQ

<table>
<thead>
<tr>
<th>Position</th>
<th>Default</th>
<th>Used to Draw</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>vertical separators</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>horizontal separators</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>intersections of vertical and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>horizontal separators</td>
</tr>
</tbody>
</table>

NOPRINT

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 14, “Using the Output Delivery System.”

Note: A NOPRINT option is also available in the TABLES statement. It suppresses display of the crosstabulation tables but allows display of the requested statistics.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the order in which the values of the frequency and crosstabulation table variables are to be reported. The following table shows how PROC FREQ interprets values of the ORDER= option.

DATA orders values according to their order in the input data set.

FORMATTED orders values by their formatted values. This order is operating environment dependent. By default, the order is ascending.

FREQ orders values by descending frequency count.

INTERNAL orders values by their unformatted values, which yields the same order that the SORT procedure does. This order is operating environment dependent.

By default, ORDER=INTERNAL. The ORDER= option does not apply to missing values, which are always ordered first.

PAGE displays only one table per page. Otherwise, PROC FREQ displays multiple tables per page as space permits. The PAGE option is not valid with the COMPRESS option.

BY Statement

BY variables;

You can specify a BY statement with PROC FREQ to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the FREQ procedure. The NOTSORTED option does not mean
that the data are unsorted but rather that the data are arranged in groups (ac-
cording to values of the BY variables) and that these groups are not necessarily
in alphabetical or increasing numeric order.

- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in SAS Language
Reference: Concepts. For more information on the DATASETS procedure, refer to
the discussion in the SAS Procedures Guide.

**EXACT Statement**

```sas
EXACT statistic-options < / computation-option > ;
```

The EXACT statement requests exact tests or confidence bounds for the specified
statistics. The `statistic-options` specify the statistics for which to provide exact tests or
confidence bounds. The `computation-option` specifies an option for the computation
of exact statistics. Computation of exact statistics can require a large amount of time
and memory.

**Statistic-Options**

The `statistic-options` specify the statistics for which exact tests or confidence bounds
are computed. PROC FREQ can compute exact $p$-values for the following hypo-
thesis tests: chi-square goodness-of-fit test for one-way tables; Pearson chi-
square, likelihood-ratio chi-square, Mantel-Haenszel chi-square, Fisher’s exact test,
Jonckheere-Terpstra test, Cochran-Armitage test for trend, and McNemar’s test for
two-way tables. PROC FREQ can also compute exact $p$-values for tests of the fol-
lowing statistics: Pearson correlation coefficient, Spearman correlation coefficient,
simple kappa coefficient, and weighted kappa coefficient. PROC FREQ can com-
pute exact $p$-values for the binomial proportion test, as well as an exact confidence
bound for the binomial proportion. Additionally, PROC FREQ can compute exact
confidence bounds for odds ratios for $2 \times 2$ tables.

Table 26.6 lists the available `statistic-options` and the exact statistics computed. The
option names are identical to the corresponding options in the TABLES statement
and the OUTPUT statement. You can request exact computations for groups of statis-
tics by using options that are identical to the following TABLES statement options:
CHISQ, MEASURES, and AGREE. For example, when you specify the CHISQ op-
tion in the EXACT statement, PROC FREQ computes exact $p$-values for the Pearson
chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square tests. You
request exact $p$-values for an individual test by specifying one of the `statistic-options`
shown in Table 26.6.
Table 26.6. EXACT Statement Statistic-Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Exact Statistics Computed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREE</td>
<td>McNemar’s test for $2 \times 2$ tables, simple kappa coefficient, and weighted kappa coefficient</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>binomial proportion test for one-way tables</td>
</tr>
<tr>
<td>CHISQ</td>
<td>chi-square goodness-of-fit test for one-way tables; Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square tests for two-way tables</td>
</tr>
<tr>
<td>FISHER</td>
<td>Fisher’s exact test</td>
</tr>
<tr>
<td>JT</td>
<td>Jonckheere-Terpstra test</td>
</tr>
<tr>
<td>KAPPA</td>
<td>test for the simple kappa coefficient</td>
</tr>
<tr>
<td>LRCHI</td>
<td>likelihood-ratio chi-square test</td>
</tr>
<tr>
<td>MCNEM</td>
<td>McNemar’s test</td>
</tr>
<tr>
<td>MEASURES</td>
<td>tests for the Pearson correlation and the Spearman correlation, and the odds ratio confidence bounds for $2 \times 2$ tables</td>
</tr>
<tr>
<td>MHCHI</td>
<td>Mantel-Haenszel chi-square test</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio confidence bounds for $2 \times 2$ tables</td>
</tr>
<tr>
<td>PCHI</td>
<td>Pearson chi-square test</td>
</tr>
<tr>
<td>PCORR</td>
<td>test for the Pearson correlation coefficient</td>
</tr>
<tr>
<td>SCORR</td>
<td>test for the Spearman correlation coefficient</td>
</tr>
<tr>
<td>TREND</td>
<td>Cochran-Armitage test for trend</td>
</tr>
<tr>
<td>WTKAP</td>
<td>test for the weighted kappa coefficient</td>
</tr>
</tbody>
</table>

**Computation-Option**

The *computation-option* specifies an option for computation of exact statistics. You can specify the following *computation-option* in the EXACT statement.

**MAXTIME=value**

specifies the maximum time (in seconds) that PROC FREQ can use to compute an exact $p$-value. If the procedure does not complete the computation within the specified time, it terminates the computation. The value of the MAXTIME= option must be a positive number.

See the “Computational Resources section” on page 1273 for more information.

**Using TABLES Statement Options with the EXACT Statement**

If you use only one TABLES statement, you do not need to specify options in the TABLES statement that are identical to options appearing in the EXACT statement. PROC FREQ automatically invokes the corresponding TABLES statement option when you specify the option in the EXACT statement. However, when you use multiple TABLES statements and want exact computations, you must specify options in the TABLES statement to compute the desired statistics. PROC FREQ then performs exact computations for all statistics that are also specified in the EXACT statement.
OUTPUT Statement

```
OUTPUT < OUT= SAS-data-set > options ;
```

The OUTPUT statement creates a SAS data set containing statistics computed by PROC FREQ. The variables contain statistics for each two-way table or stratum, as well as summary statistics across all strata.

Only one OUTPUT statement is allowed for each execution of PROC FREQ. You must specify a TABLES statement with the OUTPUT statement. If you use multiple TABLES statements, the contents of the OUTPUT data set correspond to the last TABLES statement. If you use multiple table requests in a TABLES statement, the contents of the OUTPUT data set correspond to the last table request.

For more information, see the section “Output Data Sets” on page 1275.

Note that you can use the Output Delivery System (ODS) to create a SAS data set from any piece of PROC FREQ output. For more information, see Table 26.11 on page 1282 and Chapter 14, “Using the Output Delivery System.”

You can specify the following options in an OUTPUT statement.

**OUT=SAS-data-set**

names the output data set. If you omit the OUT= option, the data set is named DATA\(n\), where \(n\) is the smallest integer that makes the name unique.

**options**

specifies the statistics that you want in the output data set. Available statistics are those produced by PROC FREQ for each one-way or two-way table, as well as the summary statistics across all strata. When you request a statistic, the OUTPUT data set contains that estimate or test statistic plus any associated standard error, confidence bounds, \(p\)-values, and degrees of freedom. You can output statistics by using group options identical to those specified in the TABLES statement: AGREE, ALL, CHISQ, CMH, and MEASURES. Alternatively, you can request an individual statistic by specifying one of the options shown in the following table.
### Table 26.7. OUTPUT Statement Options and Required TABLES Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Output Data Set Statistics</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREE</td>
<td>McNemar’s test for $2 \times 2$ tables, simple kappa coefficient, and weighted kappa coefficient; for square tables with more than two response categories, Bowker’s test of symmetry; for multiple strata, overall simple and weighted kappa statistics, and tests for equal kappas among strata; for multiple strata with two response categories, Cochran’s Q test</td>
<td>AGREE</td>
</tr>
<tr>
<td>AJCHI</td>
<td>continuity adjusted chi-square for $2 \times 2$ tables</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>ALL</td>
<td>all statistics under CHISQ, MEASURES, and CMH, and the number of nonmissing subjects</td>
<td>ALL</td>
</tr>
<tr>
<td>BDCHI</td>
<td>Breslow-Day test</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>BIN</td>
<td>chi-square goodness-of-fit test for one-way tables; for two-way tables, Pearson chi-square, likelihood ratio chi-square, continuity-adjusted chi-square for $2 \times 2$ tables, Mantel-Haenszel chi-square, Fisher’s exact test for $2 \times 2$ tables, phi coefficient, contingency coefficient, and Cramer’s V</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>CHISQ</td>
<td>Cochran-Mantel-Haenszel correlation, row mean scores (ANOVA), and general association statistics; for $2 \times 2$ tables, logit and Mantel-Haenszel adjusted odds ratios, relative risks, and Breslow-Day test</td>
<td>ALL or CMH</td>
</tr>
<tr>
<td>CMH</td>
<td>same as CMH, but excludes general association and row mean scores (ANOVA) statistics</td>
<td>ALL or CMH or CMH1</td>
</tr>
<tr>
<td>CMH1</td>
<td>same as CMH, but excludes the general association statistic</td>
<td>ALL or CMH or CMH1</td>
</tr>
<tr>
<td>CMH2</td>
<td>Cochran-Mantel-Haenszel correlation statistic</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>CMHGA</td>
<td>Cochran-Mantel-Haenszel general association statistic</td>
<td>ALL or CMH</td>
</tr>
<tr>
<td>CMHRMS</td>
<td>Cochran-Mantel-Haenszel row mean scores (ANOVA) statistic</td>
<td>ALL or CMH or CMH2</td>
</tr>
</tbody>
</table>
### Table 26.7. (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Output Data Set Statistics</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCHQ</td>
<td>Cochran’s $Q$</td>
<td>AGREE</td>
</tr>
<tr>
<td>CONTGY</td>
<td>contingency coefficient</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>CRAMV</td>
<td>Cramer’s $V$</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>EKQAP</td>
<td>test for equal simple kappas</td>
<td>AGREE</td>
</tr>
<tr>
<td>EKWKP</td>
<td>test for equal weighted kappas</td>
<td>AGREE</td>
</tr>
<tr>
<td>FISHER</td>
<td>Fisher’s exact test</td>
<td>ALL or CHISQ *</td>
</tr>
<tr>
<td>GAMMA</td>
<td>gamma</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>JT</td>
<td>Jonckheere-Terpstra test</td>
<td>JT</td>
</tr>
<tr>
<td>KAPPA</td>
<td>simple kappa coefficient</td>
<td>AGREE</td>
</tr>
<tr>
<td>KEN TD</td>
<td>Kendall’s tau-b</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>LAMCR</td>
<td>lambda asymmetric $(C</td>
<td>R)$</td>
</tr>
<tr>
<td>LAMDAS</td>
<td>lambda symmetric</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>LAMRC</td>
<td>lambda asymmetric $(R</td>
<td>C)$</td>
</tr>
<tr>
<td>LGOR</td>
<td>adjusted logit odds ratio</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>LGRRC1</td>
<td>adjusted column 1 logit relative risk</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>LGRRC2</td>
<td>adjusted column 2 logit relative risk</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>LRCHI</td>
<td>likelihood ratio chi-square</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>MCNEM</td>
<td>McNemar’s test</td>
<td>AGREE</td>
</tr>
<tr>
<td>MEASURES</td>
<td>gamma, Kendall’s tau-b, Stuart’s tau-c, Somers’ $D(C</td>
<td>R)$, Somers’ $D(R</td>
</tr>
<tr>
<td>MHCHI</td>
<td>Mantel-Haenszel chi-square</td>
<td>ALL or CHISQ</td>
</tr>
<tr>
<td>MHOR</td>
<td>adjusted Mantel-Haenszel odds ratio</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>MHRRC1</td>
<td>adjusted column 1 Mantel-Haenszel relative risk</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>MHRRC2</td>
<td>adjusted column 2 Mantel-Haenszel relative risk</td>
<td>ALL or CMH or CMH1 or CMH2</td>
</tr>
<tr>
<td>N</td>
<td>number of nonmissing subjects for the stratum</td>
<td></td>
</tr>
<tr>
<td>NMISS</td>
<td>number of missing subjects for the stratum</td>
<td></td>
</tr>
</tbody>
</table>

*ALL and CHISQ compute Fisher’s exact test for 2 $\times$ 2 tables. Use the FISHER option to compute Fisher’s exact test for general $r$$\times$$c$ tables.
Using the TABLES Statement with the OUTPUT Statement

In order to specify that the OUTPUT data set contain a particular statistic, you must have PROC FREQ compute the statistic by using the corresponding option in the TABLES statement or the EXACT statement. For example, you cannot specify the option PCHI (Pearson chi-square) in the OUTPUT statement without also specifying a TABLES statement option or an EXACT statement option to compute the Pearson chi-square. The TABLES statement option ALL or CHISQ computes the Pearson chi-square. Additionally, if you have only one TABLES statement, the EXACT statement
option CHISQ or PCHI computes the Pearson chi-square.

## TABLES Statement

**TABLES** requests < / options > ;

The TABLES statement requests one-way to \( n \)-way frequency and crosstabulation tables and the statistics for those tables.

If you omit the TABLES statement, PROC FREQ generates one-way frequency tables for all data set variables that are not listed in the other statements.

The following argument is required in the TABLES statement.

### requests

specifies the frequency and crosstabulation tables to produce. A request is composed of one variable name or several variable names separated by asterisks. To request a one-way frequency table, use a single variable. To request a two-way crosstabulation table, use an asterisk between two variables. To request a multiway table (an \( n \)-way table, where \( n \geq 2 \)), separate the desired variables with asterisks. The unique values of these variables form the rows, columns, and strata of the table.

For two-way to multiway tables, the values of the last variable form the crosstabulation table columns, while the values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables forms one stratum. PROC FREQ produces a separate crosstabulation table for each stratum. For example, a specification of A*B*C*D in a TABLES statement produces \( k \) tables, where \( k \) is the number of different combinations of values for A and B. Each table lists the values for C down the side and the values for D across the top.

You can use multiple TABLES statements in the PROC FREQ step. PROC FREQ builds all the table requests in one pass of the data, so that there is essentially no loss of efficiency. You can also specify any number of table requests in a single TABLES statement. To specify multiple table requests quickly, use a grouping syntax by placing parentheses around several variables and joining other variables or variable combinations. For example, the following statements illustrate grouping syntax.

### Table 26.8. Grouping Syntax

<table>
<thead>
<tr>
<th>Request</th>
<th>Equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>tables A*(B C);</td>
<td>tables A<em>B A</em>C;</td>
</tr>
<tr>
<td>tables (A B)*(C D);</td>
<td>tables A<em>C B</em>C A<em>D B</em>D;</td>
</tr>
<tr>
<td>tables (A B C)*D;</td>
<td>tables A<em>D B</em>D C*D;</td>
</tr>
<tr>
<td>tables A–C;</td>
<td>tables A B C;</td>
</tr>
<tr>
<td>tables (A–C)*D;</td>
<td>tables A<em>D B</em>D C*D;</td>
</tr>
</tbody>
</table>

### Without Options

If you request a one-way frequency table for a variable without specifying options, PROC FREQ produces frequencies, cumulative frequencies, percentages of the total frequency, and cumulative percentages for each value of the variable. If you request a
two-way or an $n$-way crosstabulation table without specifying options, PROC FREQ produces crosstabulation tables that include cell frequencies, cell percentages of the total frequency, cell percentages of row frequencies, and cell percentages of column frequencies. The procedure excludes observations with missing values from the table but displays the total frequency of missing observations below each table.

### Options

The following table lists the options available with the TABLES statement. Descriptions follow in alphabetical order.

**Table 26.9. TABLES Statement Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control Statistical Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>AGREE</td>
<td>requests tests and measures of classification agreement</td>
</tr>
<tr>
<td>ALL</td>
<td>requests tests and measures of association produced by CHISQ, MEASURES, and CMH</td>
</tr>
<tr>
<td>ALPHA=</td>
<td>sets the confidence level for confidence bounds</td>
</tr>
<tr>
<td>BINOMIAL</td>
<td>requests binomial proportion, confidence bounds and test for one-way tables</td>
</tr>
<tr>
<td>CHISQ</td>
<td>requests chi-square tests and measures of association based on chi-square</td>
</tr>
<tr>
<td>CL</td>
<td>requests confidence limits for the MEASURES statistics</td>
</tr>
<tr>
<td>CMH</td>
<td>requests all Cochran-Mantel-Haenszel statistics</td>
</tr>
<tr>
<td>CMH1</td>
<td>requests the CMH correlation statistic, and adjusted relative risks and odds ratios</td>
</tr>
<tr>
<td>CMH2</td>
<td>requests CMH correlation and row mean scores (ANOVA) statistics, and adjusted relative risks and odds ratios</td>
</tr>
<tr>
<td>CONVERGE=</td>
<td>specifies convergence criterion to compute polychoric correlation</td>
</tr>
<tr>
<td>FISHER</td>
<td>requests Fisher’s exact test for tables larger than $2 \times 2$</td>
</tr>
<tr>
<td>JT</td>
<td>requests Jonckheere-Terpstra test</td>
</tr>
<tr>
<td>MAXITER=</td>
<td>specifies maximum number of iterations to compute polychoric correlation</td>
</tr>
<tr>
<td>MEASURES</td>
<td>requests measures of association and their asymptotic standard errors</td>
</tr>
<tr>
<td>MISSING</td>
<td>treats missing values as nonmissing</td>
</tr>
<tr>
<td>PLCORR</td>
<td>requests polychoric correlation</td>
</tr>
<tr>
<td>RELRISK</td>
<td>requests relative risk measures for $2 \times 2$ tables</td>
</tr>
<tr>
<td>RISKDIFF</td>
<td>requests risks and risk differences for $2 \times 2$ tables</td>
</tr>
<tr>
<td>SCORES=</td>
<td>specifies the type of row and column scores</td>
</tr>
<tr>
<td>TESTF=</td>
<td>specifies expected frequencies for a one-way table chi-square test</td>
</tr>
<tr>
<td>TESTP=</td>
<td>specifies expected proportions for a one-way table chi-square test</td>
</tr>
<tr>
<td>TREND</td>
<td>requests Cochran-Armitage test for trend</td>
</tr>
<tr>
<td><strong>Control Additional Table Information</strong></td>
<td></td>
</tr>
<tr>
<td>CELLCHI2</td>
<td>displays each cell’s contribution to the total Pearson chi-square statistic</td>
</tr>
<tr>
<td>CUMCOL</td>
<td>displays the cumulative column percentage in each cell</td>
</tr>
<tr>
<td>DEVIATION</td>
<td>displays the deviation of the cell frequency from the expected value for each cell</td>
</tr>
</tbody>
</table>
Table 26.9. (continued)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPECTED</td>
<td>displays the expected cell frequency for each cell</td>
</tr>
<tr>
<td>MISSPRINT</td>
<td>displays missing value frequencies</td>
</tr>
<tr>
<td>SPARSE</td>
<td>lists all possible combinations of variable levels even when a combination</td>
</tr>
<tr>
<td></td>
<td>does not occur</td>
</tr>
<tr>
<td>TOTPCT</td>
<td>displays percentage of total frequency on ( n )-way tables when ( n &gt; 2 )</td>
</tr>
</tbody>
</table>

**Control Displayed Output**

| NOCOL        | suppresses display of the column percentage for each cell                  |
| NOCUM        | suppresses display of cumulative frequencies and cumulative percentages in |
|              | one-way frequency tables and in list format                               |
| NOFREQ       | suppresses display of the frequency count for each cell                    |
| NOPERCENT    | suppresses display of the percentage, row percentage, and column percentage in crosstabulation tables, or percentages and cumulative percentages in one-way frequency tables and in list format |
| NOPRINT      | suppresses display of tables but displays statistics                      |
| NOROW        | suppresses display of the row percentage for each cell                    |
| LIST         | displays two-way to \( n \)-way tables in list format                     |
| PRINTKWT     | displays kappa coefficient weights                                         |
| SCOROUT      | displays the row and the column scores                                    |

**Create an Output Data Set**

| OUT=         | specifies an output data set to contain variable values and frequency counts |
| OUTEXPECT    | includes the expected frequency of each cell in the output data set         |
| OUTFREQ      | includes the percentage of column frequency, row frequency, and            |
|              | two-way table frequency in the output data set                            |

You can specify the following options in a TABLES statement.

**AGREE < (WT=FC) >**

requests tests and measures of classification agreement for square tables. The AGREE option provides McNemar’s test for \( 2 \times 2 \) tables and Bowker’s test of symmetry for tables with more than two response categories. The AGREE option also produces the simple kappa coefficient, the weighted kappa coefficient, the asymptotic standard errors for the simple and weighted kappas, and the corresponding confidence bounds. When there are multiple strata, the AGREE option provides overall simple and weighted kappas as well as tests for equal kappas among strata. When there are multiple strata and two response categories, PROC FREQ computes Cochran’s \( Q \) test. For more information, see the section “Tests and Measures of Agreement” on page 1259.

The (WT=FC) specification requests that PROC FREQ use Fleiss-Cohen weights to compute the weighted kappa coefficient. By default, PROC FREQ uses Cicchetti-Allison weights. See the section “Weighted Kappa Coefficient” on page 1261 for more information. You can specify the option PRINTKWT to display the kappa coefficient weights.
requests all of the tests and measures that are computed by the CHISQ, MEASURES, and CMH options. The number of CMH statistics computed can be controlled by the CMH1 and CMH2 options.

sets the confidence level for confidence bounds. The value of the Alpha= option must be between 0.0001 and 0.9999, and the default is 0.05. A confidence level of \( \alpha \) results in \( 100(1-\alpha)\% \) confidence bounds. The default of Alpha=0.05 results in 95% confidence bounds. If \( \alpha \) is between 0 and 1 but outside the range of 0.0001 to 0.9999, PROC FREQ uses the closest range endpoint. For example, if you specify Alpha=0.000001, PROC FREQ uses 0.0001 to determine confidence bounds.

requests the binomial proportion for one-way tables. This is the proportion of observations for the first variable level that appears in the output. The BINOMIAL option also provides the asymptotic standard error, asymptotic and exact confidence intervals, and the asymptotic test for the binomial proportion. To specify the null hypothesis proportion value for the test, use the p= specification. If you omit p=value, PROC FREQ uses 0.5 as the default for the test. See the section “Binomial Proportion” on page 1250 for more information.

 displays each cell’s contribution to the total Pearson chi-square statistic, which is computed as

\[
\frac{(\text{frequency} - \text{expected})^2}{\text{expected}}
\]

The CELLCHI2 option is valid for contingency tables but has no effect on tables that are produced with the LIST option.

requests chi-square tests of homogeneity or independence and measures of association based on chi-square. The tests include the Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square. The measures include the phi coefficient, the contingency coefficient, and Cramer’s \( V \). For 2 \( \times \) 2 tables, the CHISQ option includes Fisher’s exact test and the continuity-adjusted chi-square. For one-way tables, the CHISQ option requests a chi-square goodness-of-fit test for equal proportions. If you specify the null hypothesis proportions with the TESTP= option, then PROC FREQ computes a chi-square goodness-of-fit test for the specified proportions. If you specify null hypothesis frequencies with the TESTF= option, PROC FREQ computes a chi-square goodness-of-fit test for the specified frequencies.

See the section “Chi-Square Tests and Statistics” on page 1236 for more information.

requests confidence bounds for the MEASURES statistics. If you omit the MEASURES option, the CL option invokes MEASURES. The FREQ procedure determines the confidence coefficient using the Alpha= option, which by default equals 0.05 and produces 95% confidence bounds. For more information, see the section
“Confidence Bounds” on page 1241.

CMH
requests Cochran-Mantel-Haenszel statistics, which test for association between the row and column variables after adjusting for the remaining variables in a multiway table. In addition, for $2 \times 2$ tables, PROC FREQ computes the adjusted Mantel-Haenszel and logit estimates of the odds ratios and relative risks and the corresponding confidence bounds. For the stratified $2 \times 2$ case, PROC FREQ computes the Breslow-Day test for homogeneity of odds ratios. The CMH1 and CMH2 options control the number of CMH statistics that PROC FREQ computes. For more information, see the section “Cochran-Mantel-Haenszel Statistics” on page 1265.

CMH1
requests the Cochran-Mantel-Haenszel correlation statistic and, for $2 \times 2$ tables, the adjusted Mantel-Haenszel and logit estimates of the odds ratios and relative risks and the corresponding confidence bounds. For the stratified $2 \times 2$ case, PROC FREQ computes the Breslow-Day test for homogeneity of odds ratios. Except for $2 \times 2$ tables, the CMH1 option requires less memory than the CMH option, which can require an enormous amount for large tables.

CMH2
requests the Cochran-Mantel-Haenszel correlation statistic, row mean scores (ANOV A) statistic, and, for $2 \times 2$ tables, the adjusted Mantel-Haenszel and logit estimates of the odds ratios and relative risks and the corresponding confidence bounds. For the stratified $2 \times 2$ case, PROC FREQ computes the Breslow-Day test for homogeneity of odds ratios. Except for $2 \times 2$ tables with two columns, the CMH2 option requires less memory than the CMH option, which can require an enormous amount for large tables.

CONVERGE=value
specifies the convergence criterion for computing the polychoric correlation when the PLCORR option is specified. The value of the CONVERGE= option must be a positive number; by default, CONVERGE=0.0001. Iterative computation of the polychoric correlation stops when the convergence measure falls below the value of the CONVERGE= option or when the number of iterations specified by the MAXITER= option is exceeded, whichever happens first. See the section “Polychoric Correlation” on page 1248 for more information.

CUMCOL
displays the cumulative column percentages in the cells of the crosstabulation table.

DEVIATION
displays the deviation of the cell frequency from the expected frequency for each cell of the crosstabulation table. The DEVIATION option is valid for contingency tables but has no effect on tables produced with the LIST option.

FISHER | EXACT
requests Fisher’s exact test for tables that are larger than $2 \times 2$. This test is also known as the Freeman-Halton test. For more information, see the section “Fisher’s Exact Test” on page 1239 and the “EXACT Statement” section on page 1215.
If you omit the CHISQ option in the TABLES statement, the FISHER option invokes CHISQ. You can also request Fisher’s exact test by specifying the FISHER option in the EXACT statement.

**Caution:** For tables with many rows or columns or with large total frequency, PROC FREQ may require a large amount of time or memory to compute exact $p$-values (see the section “Computational Resources” on page 1273).

**EXPECTED**

displays the expected cell frequencies under the hypothesis of independence (or homogeneity). The EXPECTED option is valid for crosstabulation tables but has no effect on tables produced with the LIST option.

**JT**

performs the Jonckheere-Terpstra test. For more information, see the section “Jonckheere-Terpstra Test” on page 1257.

**LIST**

displays two-way to $n$-way tables in a list format rather than as crosstabulation tables. PROC FREQ ignores the LIST option when you request statistical tests or measures of association.

**MAXITER=number**

specifies the maximum number of iterations for computing the polychoric correlation when the PLCORR option is specified. The value of the MAXITER= option must be a positive integer; by default, MAXITER=20. Iterative computation of the polychoric correlation stops when the number of iterations specified by the MAXITER= option is exceeded or when the convergence measures fall below the value of the CONVERGE= option, whichever happens first. For more information see the section “Polychoric Correlation” on page 1248.

**MEASURES**

requests several measures of association and their asymptotic standard errors (ASE). The measures include gamma, Kendall’s tau-$b$, Stuart’s tau-$c$, Somers’ $D(C|R)$, Somers’ $D(R|C)$, the Pearson and Spearman correlation coefficients, lambda (symmetric and asymmetric), uncertainty coefficients (symmetric and asymmetric), and, for $2 \times 2$ tables, the odds ratio, column 1 relative risk, column 2 relative risk, and the corresponding confidence bounds. For more information, see the section “Measures of Association” on page 1241.

**MISSING**

treats missing values as nonmissing and includes them in calculations of percentages and other statistics. For more information, see the section “Missing Values” on page 1233.

**MISSPRINT**

displays missing value frequencies for all tables, even though PROC FREQ does not use the frequencies in the calculation of statistics. For more information, see the section “Missing Values” on page 1233.

**NOCOL**

suppresses the display of column percentages in cells of the crosstabulation table.
NOCUM
suppresses the display of cumulative frequencies and cumulative percentages for one-
way frequency tables and for frequencies in list format.

NOFREQ
suppresses the display of cell frequencies for a crosstabulation table. This also sup-
presses frequencies for row totals.

NOPERCENT
suppresses the display of cell percentages, row total percentages, and column total
percentages for a crosstabulation table. For one-way frequency tables and frequencies
in list format, the NOPERCENT option suppresses the display of percentages and
cumulative percentages.

NOPRINT
suppresses the display of frequency and crosstabulation tables but displays all re-
quested tests and statistics. Use the NOPRINT option in the PROC FREQ statement
to suppress the display of all tables.

NOROW
suppresses the display of row percentages in cells of the crosstabulation table.

OUT=SAS-data-set
names the output data set that contains variable values and frequency counts. The
variable COUNT contains the frequencies and the variable PERCENT contains the
percentages. If more than one table request appears in the TABLES statement, the
contents of the data set correspond to the last table request in the TABLES statement.
For more information, see the section “Output Data Sets” on page 1275 and see the
following descriptions for the options OUTEXPECT and OUTPCT.

OUTEXPECT
includes the expected frequency in the output data set when you specify the OUT=
option in the TABLES statement. The variable EXPECTED contains the expected
frequency for each table cell. For more information, see the section “Output Data
Sets” on page 1275.

OUTPCT
includes the following additional variables in the output data set when you specify
the OUT= option in the TABLES statement:

- PCT_COL the percentage of column frequency
- PCT_ROW the percentage of row frequency
- PCT_TABL the percentage of stratum frequency, for \( n \)-way tables where \( n > 2 \)

For more information, see the section “Output Data Sets” on page 1275.

PLCORR
requests the polychoric correlation coefficient. For \( 2 \times 2 \) tables, this statistic is more
commonly known as the tetrachoric correlation coefficient, and it is labeled as such
in the displayed output. If you omit the MEASURES option, the PLCORR option in-
vokes MEASURES. For more information, see the section “Polychoric Correlation”
on page 1248 and the descriptions for the CONVERGE= and MAXITER= options in this list.

**PRINTKWT**

Displays the weights PROC FREQ uses to compute the weighted kappa coefficient. You must also specify the AGREE option, which requests the weighted kappa coefficient. You can specify (WT=FC) with the AGREE option to request Fleiss-Cohen weights. By default, PROC FREQ uses Cicchetti-Allison weights. See the section “Weighted Kappa Coefficient” on page 1261 for more information.

**RELRISK**

Requests relative risk measures and their confidence bounds for 2 × 2 tables. These measures include the odds ratio and the column 1 and 2 relative risks. For more information, see the section “Odds Ratio and Relative Risks for 2×2 Tables” on page 1253. You can also obtain the RELRISK measures by specifying the MEASURES option, which produces other measures of association in addition to the relative risks.

**RISDIFF**

Requests column 1 and 2 risks (or binomial proportions), risk differences, and their confidence bounds for 2 × 2 tables. See the section “Risks and Risk Differences” on page 1252 for more information.

**SCORES=type**

Specifies the type of row and column scores that PROC FREQ uses with the Mantel-Haenszel chi-square, Pearson correlation, Cochran-Armitage test for trend, weighted kappa coefficient, and Cochran-Mantel-Haenszel statistics, where type is one of the following (the default is SCORE=TABLE):

- MODRIDIT
- RANK
- RIDIT
- TABLE

By default, the row or column scores are the integers 1,2,... for character variables and the actual variable values for numeric variables. Using other types of scores yields nonparametric analyses. For more information, see the section “Scores” on page 1235.

**SCOROUT**

Displays the row and the column scores. You specify the score type with the SCORES= option. PROC FREQ uses the scores when it calculates the Mantel-Haenszel chi-square, Pearson correlation, Cochran-Armitage test for trend, weighted kappa coefficient, or Cochran-Mantel-Haenszel statistics. The SCOROUT option displays the row and column scores only when statistics are computed for two-way tables. To store the scores in an output data set, use the Output Delivery System. For more information, see the section “Scores” on page 1235.

**SPARSE**

Lists all possible combinations of the variable values for an n-way table when n > 1,
even if a combination does not occur in the data. The SPARSE option has no effect unless you also specify the LIST or OUT= option. When you use the SPARSE and LIST options, PROC FREQ lists any combination of values with a frequency count of zero. When you use the SPARSE and OUT= options, PROC FREQ includes empty crosstabulation table cells in the output data set. For more information, see the section “Missing Values” on page 1233.

**TESTF=**(values)
specifies the null hypothesis frequencies for a one-way chi-square test for specified frequencies. You can separate values with blanks or commas. The sum of the frequency values must equal the total frequency for the one-way table. The number of TESTF= values must equal the number of variable levels in the one-way table. List these values in the order in which the corresponding variable levels appear in the output. If you omit the CHISQ option, the TESTF= option invokes CHISQ. For more information, see the section “Chi-Square Test for One-Way Tables” on page 1237.

**TESTP=**(values)
specifies the null hypothesis proportions for a one-way chi-square test for specified proportions. You can separate values with blanks or commas. Specify values in probability form as numbers between 0 and 1, where the proportions sum to 1. Or specify values in percentage form as numbers between 0 and 100, where the percentages sum to 100. The number of TESTP= values must equal the number of variable levels in the one-way table. List these values in the order in which the corresponding variable levels appear in the output. If you omit the CHISQ option, the TESTP= option invokes CHISQ. For more information, see the section “Chi-Square Test for One-Way Tables” on page 1237.

**TOTPCT**
displays the percentage of total frequency on crosstabulation tables, for n-way tables where n > 2. This percentage is also available with the LIST option or as the PERCENT variable in the OUT= output data set.

**TREND**
performs the Cochran-Armitage test for trend. The table must be 2 × C or R × 2. For more information, see the section “Cochran-Armitage Test for Trend” on page 1255.

---

**TEST Statement**

```
TEST options ;
```

The TEST statement requests asymptotic tests for the specified measures of association and measures of agreement. You must use a TABLES statement with the TEST statement.

**options**
specifies the statistics for which to provide asymptotic tests. The available statistics are those measures of association and agreement listed in Table 26.10. The option names are identical to those in the TABLES statement and the OUTPUT statement.
You can request all available tests for groups of statistics by using group options MEASURES or AGREE. Or you can request tests individually by using one of the options shown in Table 26.10.

For each measure of association or agreement that you specify, the TEST statement provides an asymptotic test that the measure equals zero. When you request an asymptotic test, PROC FREQ gives the asymptotic standard error under the null hypothesis, the test statistic, and the $p$-values. Additionally, PROC FREQ reports the confidence bounds for that measure. The ALPHA= option in the TABLES statement determines the confidence level, which by default equals 0.05 and provides 95% confidence bounds. For more information, see the sections “Asymptotic Tests” on page 1241 and “Confidence Bounds” on page 1241, and see the “Statistical Computations” sections (beginning on page 1234) describing the individual measures.

In addition to these asymptotic tests, exact tests for selected measures of association and agreement are available with the EXACT statement. See the section “EXACT Statement” on page 1215 for more information.

Table 26.10. TEST Statement Options and Required TABLES Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Asymptotic Tests Computed</th>
<th>Required TABLES Statement Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGREE</td>
<td>simple kappa coefficient and weighted kappa coefficient</td>
<td>AGREE</td>
</tr>
<tr>
<td>GAMMA</td>
<td>gamma</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>KAPPA</td>
<td>simple kappa coefficient</td>
<td>AGREE</td>
</tr>
<tr>
<td>KENTB</td>
<td>Kendall’s tau-$b$</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>LAMCR</td>
<td>lambda asymmetric (C</td>
<td>R)</td>
</tr>
<tr>
<td>LAMDAS</td>
<td>lambda symmetric</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>LAMRC</td>
<td>lambda asymmetric (R</td>
<td>C)</td>
</tr>
<tr>
<td>MEASURES</td>
<td>gamma, Kendall’s tau-$b$, Stuart’s tau-$c$, Somers’ $D(C</td>
<td>R)$, Somers’ $D(R</td>
</tr>
<tr>
<td>PCORR</td>
<td>Pearson correlation coefficient</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>SCORR</td>
<td>Spearman correlation coefficient</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>SMDCR</td>
<td>Somers’ $D(C</td>
<td>R)$</td>
</tr>
<tr>
<td>SMDRC</td>
<td>Somers’ $D(R</td>
<td>C)$</td>
</tr>
<tr>
<td>STUTC</td>
<td>Stuart’s tau-$c$</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>U</td>
<td>symmetric uncertainty coefficient</td>
<td>ALL or MEASURES</td>
</tr>
<tr>
<td>UCR</td>
<td>uncertainty coefficient (C</td>
<td>R)</td>
</tr>
<tr>
<td>URC</td>
<td>uncertainty coefficient (R</td>
<td>C)</td>
</tr>
<tr>
<td>WTKAP</td>
<td>weighted kappa coefficient</td>
<td>AGREE</td>
</tr>
</tbody>
</table>
WEIGHT Statement

**WEIGHT** variable;

The WEIGHT statement specifies a numeric variable with a value that represents the frequency of the observation. The WEIGHT statement is most commonly used to input cell count data. See the “Inputting Frequency Counts” section on page 1231 for more information. If you use the WEIGHT statement, PROC FREQ assumes that an observation represents \( n \) observations, where \( n \) is the value of variable. The value of the weight variable need not be an integer, but when a value is missing or zero, PROC FREQ ignores the corresponding observation. If a WEIGHT statement does not appear, each observation has a default weight of 1. The sum of the weight variable values represents the total number of observations.

If any value of the weight variable is negative, PROC FREQ displays the frequencies (as measured by the weighted values) but does not compute percentages and other statistics. If you create an output data set using the OUT= option in the TABLES statement, PROC FREQ creates the PERCENT variable and assigns a missing value for each observation. PROC FREQ also assigns missing values to the variables that the OUTEXPECT and OUTPCT options create. You cannot create an output data set using the OUTPUT statement since statistics are not computed when there are negative weights.

Details

Inputting Frequency Counts

PROC FREQ can use either raw data or cell count data to produce frequency and crosstabulation tables. Raw data, also known as case-record data, report the data as one record for each subject or sample member. Cell count data report the data as a table, listing all possible combinations of data values along with the frequency counts. This way of presenting data often appears in published results.

The following DATA step statements store raw data in a SAS data set:

```plaintext
data Raw;
  input Subject $ R C @@;
datalines;
  01 1 1 02 1 1 03 1 1 04 1 1 05 1 1
  06 1 2 07 1 2 08 1 2 09 2 1 10 2 1
  11 2 1 12 2 1 13 2 2 14 2 2 14 2 2
;```

SAS OnlineDoc™: Version 7-1
You can store the same data as cell counts using the following DATA step statements:

```sas
data CellCounts;
  input R C Count @@;
datalines;
  1 1 5 1 2 3
  2 1 4 2 2 3
;```

The variable R contains the values for the rows, and the variable C contains the values for the columns. The Count variable contains the cell count for each row and column combination.

Both the Raw data set and the CellCounts data set produce identical frequency counts, two-way tables, and statistics. With the CellCounts data set, you must use a WEIGHT statement to specify that the Count variable contains cell counts. For example, to create a two-way crosstabulation table, submit the following statements:

```sas
proc freq data=CellCounts;
  weight Count;
  tables R*C;
run;
```

### Grouping with Formats

PROC FREQ groups a variable’s values according to its formatted values. If you assign a format to a variable with a FORMAT statement, PROC FREQ formats the variable values before dividing observations into the levels of a frequency or crosstabulation table.

For example, suppose that a variable X has the values 1.1, 1.4, 1.7, 2.1, and 2.3. Each of these values appears as a level in the frequency table. If you decide to round each value to a single digit, include the following statement in the PROC FREQ step:

```sas
format X 1.;
```

Now the table lists the frequency count for formatted level 1 as two and formatted level 2 as three.

PROC FREQ treats formatted character variables in the same way. The formatted values are used to group the observations into the levels of a frequency table or crosstabulation table. PROC FREQ uses the entire value of a character format to classify an observation.

You can also use the FORMAT statement to assign formats that were created with the FORMAT procedure to the variables. User-written formats determine the number of levels for a variable and provide labels for a table. If you use the same data with different formats, then you can produce frequency counts and statistics for different classifications of the variable values.
When you use PROC FORMAT to create a user-written format that combines missing and nonmissing values into one category, PROC FREQ treats the entire category of formatted values as missing. For example, a questionnaire codes 1 as yes, 2 as no, and 8 as a no answer. The following PROC FORMAT step creates a user-written format:

```sas
proc format;
  value Questfmt 1 ='Yes'
                 2 ='No'
                 8,.='Missing';
run;
```

When you use a FORMAT statement to assign `Questfmt` to a variable, the variable’s frequency table no longer includes a frequency count for the response of 8. You must use the MISSING or MISSPRINT option in the TABLES statement to list the frequency for no answer. The frequency count for this level includes observations with either a value of 8 or a missing value (.).

The frequency or crosstabulation table lists the values of both character and numeric variables in ascending order based on internal (unformatted) variable values unless you change the order with the ORDER= option. To list the values in ascending order by formatted values, use ORDER=FORMATTED in the PROC FREQ statement.

For more information on the FORMAT statement, refer to SAS Language Reference: Concepts.

### Missing Values

By default, PROC FREQ excludes missing values before it constructs the frequency and crosstabulation tables. PROC FREQ also excludes missing values before computing statistics. However, the total frequency of observations with missing values is displayed below each table. The following options change the way in which PROC FREQ handles missing values:

- **MISSPRINT** includes missing value frequencies in frequency or crosstabulation tables.
- **MISSING** includes missing values in percentage and statistical calculations.

The OUT= option in the TABLES statement includes an observation in the output data set that contains the frequency of missing values. The NMISS option in the OUTPUT statement creates a variable in the output data set that contains the number of missing values.

Figure 26.7 shows three ways in which PROC FREQ handles missing values. The first table uses the default method; the second table uses the MISSPRINT option; and the third table uses the MISSING option.
### Figure 26.7. Missing Values in Frequency Tables

When a combination of variable values for a crosstabulation is missing, PROC FREQ assigns zero to the frequency count for the table cell. By default, PROC FREQ omits missing combinations in list format and in the output data set that is created in a TABLES statement. To include the missing combinations, use the SPARSE option with the LIST or OUT= option in the TABLES statement.

PROC FREQ treats missing BY variable values like any other BY variable value. The missing values form a separate BY group. When the value of a WEIGHT variable is missing, PROC FREQ excludes the observation from the analysis.

## Statistical Computations

### Definitions and Notation

In this chapter, a two-way table represents the crosstabulation of variables $X$ and $Y$. Let the rows of the table be labeled by the values $X_i$, $i = 1, 2, \ldots, R$, and the columns by $Y_j$, $j = 1, 2, \ldots, C$. Let $n_{ij}$ denote the cell frequency in the $i$th row and the $j$th column and define the following:
\[ n_i = \sum_j n_{ij} \quad \text{(row totals)} \]
\[ n_j = \sum_i n_{ij} \quad \text{(column totals)} \]
\[ n = \sum_i \sum_j n_{ij} \quad \text{(overall total)} \]
\[ p_{ij} = \frac{n_{ij}}{n} \quad \text{(cell percentages)} \]
\[ p_i = \frac{n_i}{n} \quad \text{(row percentages)} \]
\[ p_j = \frac{n_j}{n} \quad \text{(column percentages)} \]
\[ R_i = \text{score for row } i \]
\[ C_j = \text{score for column } j \]
\[ \bar{R} = \sum_i n_i R_i / n \quad \text{(average row score)} \]
\[ \bar{C} = \sum_j n_j C_j / n \quad \text{(average column score)} \]
\[ A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl} \]
\[ D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl} \]
\[ P = \sum_i \sum_j n_{ij} A_{ij} \quad \text{(twice the number of concordances)} \]
\[ Q = \sum_i \sum_j n_{ij} D_{ij} \quad \text{(twice the number of discordances)} \]

**Scores**

PROC FREQ uses scores for the variable values when computing the Mantel-Haenszel chi-square, Pearson correlation, Cochran-Armitage test for trend, weighted kappa coefficient, and Cochran-Mantel-Haenszel statistics. The SCORES= option in the TABLES statement specifies the score type that PROC FREQ uses. The available score types are TABLE, RANK, RIDIT, and MODRIDIT scores. The default score type is TABLE.

For numeric variables, table scores are the values of the row and column levels. If the row or column variables are formatted, then the table score is the internal numeric value corresponding to that level. If two or more numeric values are classified into the same formatted level, then the internal numeric value for that level is the smallest of these values. For character variables, table scores are defined as the row numbers and column numbers (that is, 1 for the first row, 2 for the second row, and so on).
Rank scores, which you can use to obtain nonparametric analyses, are defined by

Row scores: \[ R_{1i} = \sum_{k < i} n_k + \frac{(n_i + 1)}{2} \quad i = 1, 2, \ldots, R \]

Column scores: \[ C_{1j} = \sum_{l < j} n_l + \frac{(n_j + 1)}{2} \quad j = 1, 2, \ldots, C \]

Note that rank scores yield midranks for tied values.

Ridit scores (Bross 1958; Mack and Skillings 1980) also yield nonparametric analyses, but they are standardized by the sample size. Ridit scores are derived from rank scores as

\[ R_{2i} = \frac{R_{1i}}{n} \]
\[ C_{2j} = \frac{C_{1j}}{n} \]

Modified ridit (MODRIDIT) scores (van Elteren 1960; Lehmann 1975), which also yield nonparametric analyses, represent the expected values of the order statistics for the uniform distribution on (0,1). Modified ridit scores are derived from rank scores as

\[ R_{3i} = \frac{R_{1i}}{(n + 1)} \]
\[ C_{3j} = \frac{C_{1j}}{(n + 1)} \]

**Chi-Square Tests and Statistics**

When you specify the CHISQ option in the TABLES statement, PROC FREQ performs the following chi-square tests for each two-way table: Pearson chi-square, continuity-adjusted chi-square for \( 2 \times 2 \) tables, likelihood-ratio chi-square, Mantel-Haenszel chi-square, and Fisher’s exact test for \( 2 \times 2 \) tables. Also, PROC FREQ computes the following statistics derived from the Pearson chi-square: the phi coefficient, the contingency coefficient, and Cramer’s \( V \). PROC FREQ computes Fisher’s exact test for general \( R \times C \) tables when you specify the FISHER (or EXACT) option in the TABLES statement, or, equivalently, when you specify the FISHER option in the EXACT statement.

For one-way frequency tables, PROC FREQ performs a chi-square goodness-of-fit test when you specify the CHISQ option. The other chi-square tests and statistics described in this section are defined only for two-way tables and so are not computed for one-way frequency tables.

All the two-way test statistics described in this section test the null hypothesis of no association between the row variable and the column variable. When the sample
size \( n \) is large, these test statistics are distributed approximately as chi-square when the null hypothesis is true. When the sample size is not large, exact tests may be useful. PROC FREQ computes exact tests for the following chi-square statistics when you specify the corresponding option in the EXACT statement: Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square. See the section “Exact Statistics” beginning on page 1271 for more information.

Note that the Mantel-Haenszel chi-square statistic is appropriate only when both variables lie on an ordinal scale. The other chi-square tests and statistics in this section are appropriate for either nominal or ordinal variables. The following sections give the formulas that PROC FREQ uses to compute the chi-square tests and statistics. For further information on the formulas and on the applicability of each statistic, refer to Agresti (1996), Stokes, Davis, and Koch (1995), and the other references cited for each statistic.

### Chi-Square Test for One-Way Tables

For one-way frequency tables, the CHISQ option in the TABLES statement computes a chi-square goodness-of-fit test. Let \( C \) denote the number of classes, or levels, in the one-way table. Let \( f_i \) denote the frequency of class \( i \) (or the number of observations in class \( i \)) for \( i = 1, 2, \ldots, C \). Then PROC FREQ computes the chi-square statistic as

\[
Q_P = \sum_{i=1}^{C} \frac{(f_i - e_i)^2}{e_i}
\]

where \( e_i \) is the expected frequency for class \( i \) under the null hypothesis.

In the test for equal proportions, which is the default for the CHISQ option, the null hypothesis specifies equal proportions of the total sample size for each class. Under this null hypothesis, the expected frequency for each class equals the total sample size divided by the number of classes,

\[
e_i = \frac{n}{C} \quad \text{for} \ i = 1, 2, \ldots, C
\]

In the test for specified frequencies, which PROC FREQ computes when you input null hypothesis frequencies using the TESTF= option, the expected frequencies are those TESTF= values. In the test for specified proportions, which PROC FREQ computes when you input null hypothesis proportions using the TESTP= option, the expected frequencies are determined from the TESTP= proportions \( p_i \), as

\[
e_i = p_i \cdot n \quad \text{for} \ i = 1, 2, \ldots, C
\]

Under the null hypothesis (of equal proportions, specified frequencies, or specified proportions), this test statistic has an asymptotic chi-square distribution, with \( C - 1 \) degrees of freedom. In addition to the asymptotic test, PROC FREQ computes the exact one-way chi-square test when you specify the CHISQ option in the EXACT statement.
Chi-Square Test for Two-Way Tables

The Pearson chi-square statistic for two-way tables involves the differences between the observed and expected frequencies, where the expected frequencies are computed under the null hypothesis of independence. The chi-square statistic is computed as

\[ Q_P = \sum_{i} \sum_{j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \]

where

\[ e_{ij} = \frac{n_i \cdot n_j}{n} \]

When the row and column variables are independent, \( Q_P \) has an asymptotic chi-square distribution with \((R - 1)(C - 1)\) degrees of freedom. For large values of \( Q_P \), this test rejects the null hypothesis in favor of the alternative hypothesis of general association. In addition to the asymptotic test, PROC FREQ computes the exact chi-square test when you specify the PCHI or CHISQ option in the EXACT statement.

For a \( 2 \times 2 \) table, the Pearson chi-square is also appropriate for testing the equality of two binomial proportions or, for \( R \times 2 \) and \( 2 \times C \) tables, the homogeneity of proportions. Refer to Fienberg (1980).

Likelihood-Ratio Chi-Square Test

The likelihood-ratio chi-square statistic involves the ratios between the observed and expected frequencies. The statistic is computed as

\[ G^2 = 2 \sum_{i} \sum_{j} n_{ij} \ln \left( \frac{n_{ij}}{e_{ij}} \right) \]

When the row and column variables are independent, \( G^2 \) has an asymptotic chi-square distribution with \((R - 1)(C - 1)\) degrees of freedom. In addition to the asymptotic test, PROC FREQ computes the exact test when you specify the LRCHI or CHISQ option in the EXACT statement.

Continuity-Adjusted Chi-Square Test

The continuity-adjusted chi-square statistic for \( 2 \times 2 \) tables is similar to the Pearson chi-square, except that it is adjusted for the continuity of the chi-square distribution. The continuity-adjusted chi-square is most useful for small sample sizes. The use of the continuity adjustment is controversial; this chi-square test is more conservative, and more like Fisher’s exact test, when your sample size is small. As the sample size increases, the statistic becomes more and more like the Pearson chi-square. The statistic is computed as

\[ Q_C = \sum_{i} \sum_{j} \left[ \frac{\max(0, |n_{ij} - e_{ij}| - 0.5)}{e_{ij}} \right]^2 \]

Under the null hypothesis of independence, \( Q_C \) has an asymptotic chi-square distribution with \((R - 1)(C - 1)\) degrees of freedom.
Mantel-Haenszel Chi-Square Test
The Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between the row variable and the column variable. Both variables must lie on an ordinal scale. The statistic is computed as

\[ Q_{MH} = (n - 1)r^2 \]

where \( r^2 \) is the Pearson correlation between the row variable and the column variable. For a description of the Pearson correlation, see the “Pearson Correlation Coefficient” section on page 1245. The Pearson correlation and, thus, the Mantel-Haenszel chi-square statistic use the scores that you specify in the SCORES= option in the TABLES statement.

Under the null hypothesis of no association, \( Q_{MH} \) has an asymptotic chi-square distribution with 1 degree of freedom. In addition to the asymptotic test, PROC FREQ computes the exact test when you specify the MHCHI or CHISQ option in the EXACT statement.

Refer to Mantel and Haenszel (1959) and Landis, Heyman, and Koch (1978).

Fisher’s Exact Test
2 × 2 Tables
For 2 × 2 tables, Fisher’s exact test is the probability of observing a table that gives at least as much evidence of association as the one actually observed, given that the null hypothesis is true. The row and column margins are assumed to be fixed. The hypergeometric probability, \( p \), of every possible table is computed, and the \( p \)-value is defined as

\[ PROB = \sum_A p \]

For a two-sided alternative hypothesis, \( A \) is the set of tables with \( p \) less than or equal to the probability of the observed table. A small two-sided \( p \)-value supports the alternative hypothesis of association between the row and column variables.

One-sided tests are defined in terms of the frequency of the cell in the first row and first column (the (1,1) cell). For a left-sided alternative hypothesis, \( A \) is the set of tables where the frequency in the (1,1) cell is less than or equal to that of the observed table. A small left-sided \( p \)-value supports the alternative hypothesis that the probability of an observation being in the first cell is less than expected under the null hypothesis of independent row and column variables.

Similarly, for a right-sided alternative hypothesis, \( A \) is the set of tables where the frequency in the (1,1) cell is greater than or equal to that of the observed table. A small right-sided \( p \)-value supports the alternative that the probability of an observation being in the first cell is greater than expected under the null hypothesis.

Because the (1,1) cell frequency completely determines the 2 × 2 table when the marginal row and column sums are fixed, these one-sided alternatives can be equivalently stated in terms of other cell probabilities or ratios of cell probabilities. The
left-sided alternative is equivalent to an odds ratio greater than 1, where the odds ratio equals \( \frac{n_{11} n_{22}}{n_{12} n_{21}} \). Additionally, the left-sided alternative is equivalent to the column 1 risk for row 1 being less than the column 1 risk for row 2, \( p_{1|1} < p_{1|2} \). Similarly, the right-sided alternative is equivalent to the column 1 risk for row 1 being greater than the column 1 risk for row 2, \( p_{1|1} > p_{1|2} \). Refer to Agresti (1996).

**R \times C Tables**

Fisher’s exact test was extended to general \( R \times C \) tables by Freeman and Halton (1951), and this test is also known as the Freeman-Halton test. For \( R \times C \) tables, the two-sided \( p \)-value is defined the same as it is for \( 2 \times 2 \) tables. \( A \) is the set of all tables with \( p \) less than or equal to the probability of the observed table. A small \( p \)-value supports the alternative hypothesis of association between the row and column variables. For \( R \times C \) tables, Fisher’s exact test is inherently two-sided. The alternative hypothesis is defined only in terms of general, and not linear, association. Therefore, PROC FREQ does not compute right-sided or left-sided \( p \)-values for general \( R \times C \) tables.

For \( R \times C \) tables, PROC FREQ computes Fisher’s exact test using the network algorithm of Mehta and Patel (1983), which provides a faster and more efficient solution than direct enumeration. See the section “Exact Statistics” beginning on page 1271 for more details.

**Phi Coefficient**

The phi coefficient is a measure of association derived from the Pearson chi-square statistic. It has the range \( -1 \leq \phi \leq 1 \) for \( 2 \times 2 \) tables. Otherwise, the range is \( 0 \leq \phi \leq \min(\sqrt{R-1}, \sqrt{C-1}) \) (Liebetrau 1983). The phi coefficient is computed as

\[
\phi = \frac{n_{11} n_{22} - n_{12} n_{21}}{\sqrt{n_{1.} n_{2.} n_{.1} n_{.2}}} \quad \text{for} \ 2 \times 2 \text{ tables}
\]

\[
\phi = \sqrt{Q_P / n} \quad \text{otherwise}
\]


**Contingency Coefficient**

The contingency coefficient is a measure of association derived from the Pearson chi-square. It has the range \( 0 \leq P \leq \sqrt{(m-1)/m} \), where \( m = \min(R, C) \) (Liebetrau 1983). The contingency coefficient is computed as

\[
P = \sqrt{\frac{Q_P}{Q_P + n}}
\]

Refer to Kendall and Stuart (1979, pp. 587–588).

**Cramer’s V**

Cramer’s V is a measure of association derived from the Pearson chi-square. It is designed so that the attainable upper bound is always 1. It has the range \( -1 \leq V \leq 1 \)
for $2 \times 2$ tables; otherwise, the range is $0 \leq V \leq 1$. Cramer’s $V$ is computed as

$$V = \phi \quad \text{for } 2 \times 2 \text{ tables}$$

$$V = \sqrt{\frac{Q_p/n}{\min(R-1,C-1)}} \quad \text{otherwise}$$

Refer to Kendall and Stuart (1979, p. 588).

### Measures of Association

When you specify the MEASURES option in the TABLES statement, PROC FREQ computes several statistics that describe the association between the two variables of the contingency table. The following are measures of ordinal association that consider whether the variable $Y$ tends to increase as $X$ increases: gamma, Kendall’s tau-$b$, Stuart’s tau-$c$, and Somers’ $D$. These measures are appropriate for ordinal variables, and they classify pairs of observations as concordant or discordant. A pair is concordant if the observation with the larger value of $X$ also has the larger value of $Y$. A pair is discordant if the observation with the larger value of $X$ has the smaller value of $Y$. Refer to Agresti (1996) and the other references cited in the discussion of each measure of association.

The Pearson correlation coefficient and the Spearman rank correlation coefficient are also appropriate for ordinal variables. The Pearson correlation describes the strength of the linear association between the row and column variables, and it is computed using the row and column scores specified by the SCORES= option in the TABLES statement. The Spearman correlation is computed with rank scores. The polychoric correlation (requested by the PLCORR option) also requires ordinal variables and assumes that the variables have an underlying bivariate normal distribution. The following measures of association do not require ordinal variables, but they are appropriate for nominal variables: lambda asymmetric, lambda symmetric, and uncertainty coefficients.

PROC FREQ computes estimates of the measures according to the formulas given in the discussion of each measure of association. For each measure, PROC FREQ computes an asymptotic standard error (ASE), which is the square root of the asymptotic variance denoted by $\text{var}$ in the following sections.

### Confidence Bounds

If you specify the CL option in the TABLES statement, PROC FREQ computes asymptotic confidence bounds for all MEASURES statistics. The confidence coefficient is determined according to the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence bounds. The confidence bounds are computed as

$$\text{est} \pm z_{\alpha/2} \cdot \text{ASE}$$

where $\text{est}$ is the estimate of the measure, $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, and ASE is the asymptotic standard error of the estimate.
Asymptotic Tests

For each measure that you specify in the TEST statement, PROC FREQ computes an asymptotic test of the null hypothesis that the measure equals zero. Asymptotic tests are available for the following measures of association: gamma, Kendall’s tau-\(b\), Stuart’s tau-\(c\), Somers’ \(D(R|C)\), Somers’ \(D(C|R)\), the Pearson correlation coefficient, and the Spearman rank correlation coefficient. To compute an asymptotic test, PROC FREQ uses a standardized test statistic \(z\), which has an asymptotic standard normal distribution under the null hypothesis. The standardized test statistic is computed as

\[
z = \frac{est}{\sqrt{var_0(est)}}
\]

where \(est\) is the estimate of the measure and \(var_0(est)\) is the variance of the estimate under the null hypothesis. Formulas for \(var_0(est)\) are given in the discussion of each measure of association.

Note that the ratio of \(est\) to \(\sqrt{var_0(est)}\) is the same for the following measures: gamma, Kendall’s tau-\(b\), Stuart’s tau-\(c\), Somers’ \(D(R|C)\), and Somers’ \(D(C|R)\). Therefore, the tests for these measures are identical. For example, the \(p\)-values for the test of \(H_0: \text{gamma} = 0\) equal the \(p\)-values for the test of \(H_0: \text{tau-}b = 0\).

PROC FREQ computes one-sided and two-sided \(p\)-values for each of these tests. When the test statistic \(z\) is greater than its null hypothesis expected value of zero, PROC FREQ computes the right-sided \(p\)-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided \(p\)-value supports the alternative hypothesis that the true value of the measure is greater than zero. When the test statistic is less than or equal to zero, PROC FREQ computes the left-sided \(p\)-value, which is the probability of a smaller value of the statistic occurring under the null hypothesis. A small left-sided \(p\)-value supports the alternative hypothesis that the true value of the measure is less than zero. The one-sided \(p\)-value \(P_1\) can be expressed as

\[
P_1 = \text{Prob}(Z > z) \quad \text{if } z > 0
\]

\[
P_1 = \text{Prob}(Z < z) \quad \text{if } z \leq 0
\]

where \(Z\) has a standard normal distribution. The two-sided \(p\)-value \(P_2\) is computed as

\[
P_2 = \text{Prob}(|Z| > |z|)
\]

Exact Tests

Exact tests are available for two measures of association, the Pearson correlation coefficient and the Spearman rank correlation coefficient. If you specify the PCORR option in the EXACT statement, PROC FREQ computes the exact test of the hypothesis that the Pearson correlation equals zero. If you specify the SCORR option in the EXACT statement, PROC FREQ computes the exact test of the hypothesis that the
Spearman correlation equals zero. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

**Gamma**

The estimator of gamma is based only on the number of concordant and discordant pairs of observations. It ignores tied pairs (that is, pairs of observations that have equal values of \(X\) or equal values of \(Y\)). Gamma is appropriate only when both variables lie on an ordinal scale. It has the range \(-1 \leq \Gamma \leq 1\). If the two variables are independent, then the estimator of gamma tends to be close to zero. Gamma is estimated by

\[
G = \frac{P - Q}{P + Q}
\]

with asymptotic variance

\[
\text{var} = \frac{16}{(P + Q)^4} \sum_i \sum_j n_{ij} (QA_{ij} - PD_{ij})^2
\]

The variance of the estimator under the null hypothesis that gamma equals zero is computed as

\[
\text{var}_0(G) = \frac{4}{(P + Q)^2} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P - Q)^2/n \right)
\]

For \(2 \times 2\) tables, gamma is equivalent to Yule’s \(Q\). Refer to Goodman and Kruskal (1979), Agresti (1990), and Brown and Benedetti (1977).

**Kendall’s Tau-b**

Kendall’s tau-b is similar to gamma except that tau-b uses a correction for ties. Tau-b is appropriate only when both variables lie on an ordinal scale. Tau-b has the range \(-1 \leq \tau_b \leq 1\). It is estimated by

\[
t_b = \frac{P - Q}{\sqrt{w_rw_c}}
\]

with

\[
\text{var} = \frac{1}{w^4} \left( \sum_i \sum_j n_{ij} (2w_d_{ij} + t_b w_{ij})^2 - n^2 t_b^2 (w_r + w_c)^2 \right)
\]

where

\[
w = \sqrt{w_rw_c}
\]

\[
w_r = n^2 - \sum_i n_i^2
\]
\[ w_c = n^2 - \sum_j n_j^2 \]
\[ d_{ij} = A_{ij} - D_{ij} \]
\[ v_{ij} = n_i w_c + n_j w_r \]

The variance of the estimator under the null hypothesis that tau-b equals zero is computed as

\[ \text{var}_0(t_b) = \frac{4}{w_r w_c} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P - Q)^2 / n \right) \]

Refer to Kendall (1955) and Brown and Benedetti (1977).

**Stuart’s Tau-c**

Stuart’s tau-c makes an adjustment for table size in addition to a correction for ties. Tau-c is appropriate only when both variables lie on an ordinal scale. Tau-c has the range \(-1 \leq \tau_c \leq 1\). It is estimated by

\[ t_c = \frac{m(P - Q)}{n^2(m - 1)} \]

with

\[ \text{var} = \frac{4m^2}{(m - 1)^2 n^4} \left( \sum_i \sum_j n_{ij} d_{ij}^2 - (P - Q)^2 / n \right) \]

where

\[ m = \min(R, C) \]
\[ d_{ij} = A_{ij} - D_{ij} \]

The variance of the estimator under the null hypothesis that tau-c equals zero is

\[ \text{var}_0(t_c) = \text{var} \]

Refer to Brown and Benedetti (1977).

**Somers’ D (C | R) and D (R | C)**

Somers’ D(C|R) and Somers’ D(R|C) are asymmetric modifications of tau-b. C|R denotes that the row variable X is regarded as an independent variable, while the
column variable \( Y \) is regarded as dependent. Similarly, \( R|C \) denotes that the column variable \( Y \) is regarded as an independent variable, while the row variable \( X \) is regarded as dependent. Somers’ \( D \) differs from \( \tau -b \) in that it uses a correction only for pairs that are tied on the independent variable. Somers’ \( D \) is appropriate only when both variables lie on an ordinal scale. It has the range \(-1 \leq D \leq 1\). Formulas for Somers’ \( D(R|C) \) are obtained by interchanging the indices.

\[
D(C|R) = \frac{P - Q}{w_r}
\]

with

\[
\text{var} = \frac{4}{w_r^4} \sum_i \sum_j n_{ij} (w_rd_{ij} - (P - Q)(n - n_i))^2
\]

where

\[
w_r = n^2 - \sum_i n_i^2.
\]

\[
d_{ij} = A_{ij} - D_{ij}
\]

The variance of the estimator under the null hypothesis that \( D(C|R) \) equals zero is computed as

\[
\text{var}_0(D(C|R)) = \frac{4}{w_r^2} \left( \sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P - Q)^2/n \right)
\]

Refer to Somers (1962), Goodman and Kruskal (1979), and Liebetrau (1983).

**Pearson Correlation Coefficient**

PROC FREQ computes the Pearson correlation coefficient using the scores specified in the \texttt{SCORES=} option. The Pearson correlation is appropriate only when both variables lie on an ordinal scale. It has the range \(-1 \leq \rho \leq 1\). The Pearson correlation coefficient is computed as

\[
r = \frac{u}{w} = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}
\]

with

\[
\text{var} = \frac{1}{w^4} \sum_i \sum_j n_{ij} \left( w(R_i - \bar{R})(C_j - \bar{C}) - \frac{b_{ij}v}{2w} \right)^2
\]
The row scores \( R_i \) and the column scores \( C_j \) are determined by the SCORES= option in the TABLES statement, and

\[
ss_r = \sum_i \sum_j n_{ij} (R_i - \bar{R})^2
\]

\[
ss_c = \sum_i \sum_j n_{ij} (C_j - \bar{C})^2
\]

\[
ss_{rc} = \sum_i \sum_j n_{ij} (R_i - \bar{R})(C_j - \bar{C})
\]

\[
b_{ij} = (R_i - \bar{R})^2 ss_c + (C_j - \bar{C})^2 ss_r
\]

\[
v = ss_{rc}
\]

\[
w = \sqrt{ss_r ss_c}
\]

Refer to Snedecor and Cochran (1989) and Brown and Benedetti (1977).

To compute an asymptotic test for the Pearson correlation, PROC FREQ uses a standardized test statistic \( r^* \), which has an asymptotic standard normal distribution under the null hypothesis that the correlation equals zero. The standardized test statistic is computed as

\[
r^* = \frac{r}{\sqrt{\text{var}_0(r)}}
\]

where \( \text{var}_0(r) \) is the variance of the correlation under the null hypothesis.

\[
\text{var}_0(r) = \frac{\sum_i \sum_j n_{ij} (R_i - \bar{R})^2 (C_j - \bar{C})^2 - ss_{rc}^2/n}{ss_r ss_c}
\]

The asymptotic variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. Refer to Brown and Benedetti (1977).

PROC FREQ also computes the exact test for the hypothesis that the Pearson correlation equals zero when you specify the PCORR option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

**Spearman Rank Correlation Coefficient**

The Spearman correlation coefficient is computed using rank scores \( R1_i \) and \( C1_j \), defined in the section “Scores” beginning on page 1235. It is appropriate only when both variables lie on an ordinal scale. It has the range \(-1 \leq \rho_s \leq 1\). The Spearman correlation coefficient is computed as

\[
r_s = \frac{v}{w}
\]
with

\[ \text{var} = \frac{1}{n^2w^4} \sum_i \sum_j n_{ij}(z_{ij} - \bar{z})^2 \]

where

\[
\begin{align*}
  v &= \sum_i \sum_j n_{ij}R(i)C(j) \\
  w &= \frac{1}{12}\sqrt{FG} \\
  F &= n^3 - \sum_i n_i^3 \\
  G &= n^3 - \sum_j n_j^3 \\
  R(i) &= R1_i - n/2 \\
  C(j) &= C1_j - n/2 \\
  z &= \frac{1}{n} \sum_i \sum_j n_{ij}z_{ij} \\
  z_{ij} &= wv_{ij} - vw_{ij} \\
  v_{ij} &= n \left( R(i)C(j) + \frac{1}{2} \sum_l n_{il}C(l) + \frac{1}{2} \sum_k n_{kj}R(k) + \sum_l \sum_{k>i} n_{kl}C(l) + \sum_k \sum_{l>j} n_{kl}R(k) \right) \\
  w_{ij} &= \frac{-n}{96w}(F_{n_j}^2 + G_{n_i}^2)
\end{align*}
\]

Refer to Snedecor and Cochran (1989) and Brown and Benedetti (1977).

To compute an asymptotic test for the Spearman correlation, PROC FREQ uses a standardized test statistic \( r_s^* \), which has an asymptotic standard normal distribution under the null hypothesis that the correlation equals zero. The standardized test statistic is computed as

\[
r_s^* = \frac{r_s}{\sqrt{\text{var}_0(r_s)}}
\]
where \( \text{var}_0(r_s) \) is the variance of the correlation under the null hypothesis.

\[
\text{var}_0(r_s) = \frac{1}{n^2w^2} \sum_i \sum_j n_{ij}(v_{ij} - \bar{v})^2
\]

where

\[
\bar{v} = \frac{1}{n^2} \sum_i \sum_j n_{ij}v_{ij}/n
\]

The asymptotic variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. Refer to Brown and Benedetti (1977).

PROC FREQ also computes the exact test for the hypothesis that the Spearman rank correlation equals zero when you specify the SCORR option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

**Polychoric Correlation**

When you specify the PLCORR option in the TABLES statement, PROC FREQ computes the polychoric correlation. This measure of association is based on the assumption that the ordered, categorical variables of the frequency table have an underlying bivariate normal distribution. For \( 2 \times 2 \) tables, the polychoric correlation is also known as the tetrachoric correlation. Refer to Drasgow (1986) for an overview of polychoric correlation. The polychoric correlation coefficient is the maximum likelihood estimate of the product-moment correlation between the normal variables, estimating thresholds from the observed table frequencies. The range of the polychoric correlation is from -1 to 1. Olsson (1979) gives the likelihood equations and an asymptotic covariance matrix for the estimates.

To estimate the polychoric correlation, PROC FREQ iteratively solves the likelihood equations by a Newton-Raphson algorithm using the Pearson correlation coefficient as the initial approximation. Iteration stops when the convergence measure falls below the convergence criterion or when the maximum number of iterations is reached, whichever occurs first. The CONVERGE= option sets the convergence criterion, and the default value is 0.0001. The MAXITER= option sets the maximum number of iterations, and the default value is 20.

**Lambda Asymmetric**

Asymmetric lambda, \( \lambda(C|R) \), is interpreted as the probable improvement in predicting the column variable \( Y \) given knowledge of the row variable \( X \). Asymmetric lambda has the range \( 0 \leq \lambda(C|R) \leq 1 \). It is computed as

\[
\lambda(C|R) = \frac{\sum_i r_i - r}{n - r}
\]
Measures of Association

[Equation]

\[ \text{var} = \frac{n - \sum_i r_i}{(n - r)^3} \left( \sum_i r_i + r - 2 \sum_i (r_i | l_i = l) \right) \]

where

\[ r_i = \max_j (n_{ij}) \]
\[ r = \max_j (n_{.,j}) \]

Also, let \( l_i \) be the unique value of \( j \) such that \( r_i = n_{ij} \), and let \( l \) be the unique value of \( j \) such that \( r = n_{.,j} \).

Because of the uniqueness assumptions, ties in the frequencies or in the marginal totals must be broken in an arbitrary but consistent manner. In case of ties, \( l \) is defined here as the smallest value of \( j \) such that \( r = n_{.,j} \). For a given \( i \), if there is at least one value \( j \) such that \( n_{ij} = r_i = c_j \), then \( l_i \) is defined here to be the smallest such value of \( j \). Otherwise, if \( n_{il} = r_i \), then \( l_i \) is defined to be equal to \( l \). If neither condition is true, then \( l_i \) is taken to be the smallest value of \( j \) such that \( n_{ij} = r_i \). The formulas for lambda asymmetric \((R|C)\) can be obtained by interchanging the indices.

Refer to Goodman and Kruskal (1979).

**Lambda Symmetric**

The nondirectional lambda is the average of the two asymmetric lambdas, \( \lambda(C|R) \) and \( \lambda(R|C) \). Lambda symmetric has the range \( 0 \leq \lambda \leq 1 \). Lambda symmetric is defined as

\[ \lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c} = \frac{w - v}{w} \]

with

\[ \text{var} = \frac{1}{w^4} \left( wvy - 2w^2 \left[ n - \sum_i \sum_j (n_{ij} | j = l_i, i = k_j) \right] - 2v^2(n - n_{kl}) \right) \]

where

\[ c_j = \max_i (n_{ij}) \]
\[ c = \max_i (n_{.,i}) \]
\[ w = 2n - r - c \]
\[ v = 2n - \sum_i r_i - \sum_j c_j \]
\[ x = \sum_i (r_i | l_i = l) + \sum_j (c_j | k_j = k) + r_k + c_l \]
\[ y = 8n - w - v - 2x \]

Refer to Goodman and Kruskal (1979).

**Uncertainty Coefficients (C|R) and (R|C)**

The uncertainty coefficient, \( U(C|R) \), is the proportion of uncertainty (entropy) in the column variable \( Y \) that is explained by the row variable \( X \). It has the range \( 0 \leq U(C|R) \leq 1 \). The formulas for \( U(R|C) \) can be obtained by interchanging the indices.

\[ U(C|R) = \frac{H(X) + H(Y) - H(XY)}{H(Y)} = \frac{v}{w} \]

with

\[ \text{var} = \frac{1}{n^2 w^4} \sum_i \sum_j n_{ij} \left( H(Y) \ln \left( \frac{n_{ij}}{n_i} \right) + (H(X) - H(XY)) \ln \left( \frac{n_{ij}}{n} \right) \right)^2 \]

where

\[ v = H(X) + H(Y) - H(XY) \]
\[ w = H(Y) \]
\[ H(X) = -\sum_i \left( \frac{n_i}{n} \right) \ln \left( \frac{n_i}{n} \right) \]
\[ H(Y) = -\sum_j \left( \frac{n_{.j}}{n} \right) \ln \left( \frac{n_{.j}}{n} \right) \]
\[ H(XY) = -\sum_i \sum_j \left( \frac{n_{ij}}{n} \right) \ln \left( \frac{n_{ij}}{n} \right) \]

Refer to Theil (1972, pp. 115–120) and Goodman and Kruskal (1979).

**Uncertainty Coefficient (U)**

The uncertainty coefficient, \( U \), is the symmetric version of the two asymmetric coefficients. It has the range \( 0 \leq U \leq 1 \). It is defined as

\[ U = \frac{2(H(X) + H(Y) - H(XY))}{H(X) + H(Y)} \]

with

\[ \text{var} = 4 \sum_i \sum_j n_{ij} \left( H(XY) \ln \left( \frac{n_{i.} n_{.j}}{n^2} \right) - (H(X) + H(Y)) \ln \left( \frac{n_{i.} n_{.j}}{n^2 (H(X) + H(Y))^4} \right) \right)^2 \]
Refer to Goodman and Kruskal (1979).

**Binomial Proportion**

When you specify the BINOMIAL option in the TABLES statement, PROC FREQ computes a binomial proportion for one-way tables. This is the proportion of observations in the first variable level, or class, that appears in the output.

\[ \hat{p} = \frac{n_1}{n} \]

where \( n_1 \) is the frequency for the first level and \( n \) is the total frequency for the one-way table. The standard error for the binomial proportion is computed as

\[ se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

Using the normal approximation to the binomial distribution, PROC FREQ constructs asymptotic confidence bounds for \( p \) according to

\[ \hat{p} \pm z_{\alpha/2} \cdot se(\hat{p}) \]

where \( z_{\alpha/2} \) is the 100\((1 - \alpha/2)\) percentile of the standard normal distribution. The confidence level \( \alpha \) is determined by the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence bounds. Additionally, PROC FREQ computes exact confidence bounds for the binomial proportion using the \( F \) distribution method given in Collett (1991) and also described by Leemis and Trivedi (1996).

PROC FREQ computes an asymptotic test of the hypothesis that the binomial proportion equals \( p_0 \), where the value of \( p_0 \) is specified by the P= option in the TABLES statement. If you do not specify a value for the P= option, PROC FREQ uses \( p_0 = 0.5 \) by default. The asymptotic test statistic is

\[ z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \]

PROC FREQ computes one-sided and two-sided \( p \)-values for this test. When the test statistic \( z \) is greater than zero, its expected value under the null hypothesis, PROC FREQ computes the right-sided \( p \)-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided \( p \)-value supports the alternative hypothesis that the true value of the proportion is greater than \( p_0 \). When the test statistic is less than or equal to zero, PROC FREQ computes the left-sided \( p \)-value, which is the probability of a smaller value of the statistic occurring under the null hypothesis. A small left-sided \( p \)-value supports the alternative hypothesis that the true value of the proportion is less than \( p_0 \). The one-sided \( p \)-value \( P_1 \) can be expressed as

\[ P_1 = \text{Prob} \left( Z > z \right) \quad \text{if} \quad z > 0 \]

\[ P_1 = \text{Prob} \left( Z < z \right) \quad \text{if} \quad z \leq 0 \]
where $Z$ has a standard normal distribution. The two-sided $p$-value $P_2$ is computed as

$$P_2 = \text{Prob} ( |Z| > |z| )$$

When you specify the BINOMIAL option in the EXACT statement, PROC FREQ also computes an exact test of the null hypothesis $H_0: p = p_0$. To compute this exact test, PROC FREQ uses the binomial probability function

$$\text{Prob} (X = x \mid p_0) = \binom{n}{x} p_0^x (1 - p_0)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

where the variable $X$ has a binomial distribution with parameters $n$ and $p_0$. To compute $\text{Prob}(X \leq n_1)$, PROC FREQ sums these binomial probabilities over $x$ from zero to $n_1$. To compute $\text{Prob}(X \geq n_1)$, PROC FREQ sums these binomial probabilities over $x$ from $n_1$ to $n$. Then the exact one-sided $p$-value is

$$P_1 = \min (\text{Prob}(X \leq n_1 \mid p_0), \text{Prob}(X \geq n_1 \mid p_0))$$

and the exact two-sided $p$-value is

$$P_2 = 2 \cdot P_1$$

**Risks and Risk Differences**

The RISKDIFF option in the TABLES statement provides estimates of risks (or binomial proportions) and risk differences for $2 \times 2$ tables. This analysis may be appropriate when comparing the proportion of some characteristic for two groups, where row 1 and row 2 correspond to the two groups, and the columns correspond to two possible characteristics or outcomes. For example, the row variable might be a treatment or dose, and the column variable might be the response. Refer to Collett (1991), Fleiss (1981), and Stokes, Davis, and Koch (1995).

Let the frequencies of the $2 \times 2$ table be represented as follows.

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_1$.</td>
</tr>
<tr>
<td>Row 2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_2$.</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{-1}$</td>
<td>$n_{-2}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

The column 1 risk for row 1 is the proportion of row 1 observations classified in column 1,

$$P_{111} = \frac{n_{11}}{n_1}$$

This estimates the conditional probability of the column 1 response, given the first level of the row variable.
The column 1 risk for row 2 is the proportion of row 2 observations classified in column 1,

\[ p_{12} = \frac{n_{21}}{n_2} \]

and the overall column 1 risk is the proportion of all observations classified in column 1,

\[ p_1 = \frac{n_1}{n} \]

The column 1 risk difference compares the risks for the two rows, and it is computed as the column 1 risk for row 1 minus the column 1 risk for row 2,

\[ (p_{\text{diff}})_1 = p_{11} - p_{12} \]

The risks and risk difference are defined similarly for column 2.

The standard error of the column 1 risk estimate for row \( i \) is computed as

\[ se(p_{1i}) = \sqrt{\frac{p_{1i}(1 - p_{1i})}{n_i}}. \]

The standard error of the overall column 1 risk estimate is computed as

\[ se(p_1) = \sqrt{\frac{p_1(1 - p_1)}{n}} \]

If the two rows represent independent binomial samples, the standard error for the column 1 risk difference is computed as

\[ se((p_{\text{diff}})_1) = \sqrt{\frac{\text{var}(p_{11}) + \text{var}(p_{12})}{n}} \]

The standard errors are computed in a similar manner for the column 2 risks and risk difference.

Using the normal approximation to the binomial distribution, PROC FREQ constructs asymptotic confidence bounds for the risks and risk differences according to

\[ est \pm z_{\alpha/2} \cdot se(est) \]

where \( est \) is the estimate, \( z_{\alpha/2} \) is the 100(1 - \( \alpha/2 \)) percentile of the standard normal distribution, and \( se(est) \) is the standard error of the estimate. The confidence level \( \alpha \) is determined from the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence bounds.

PROC FREQ computes exact confidence bounds for the column 1, column 2, and overall risks using the \( F \) distribution method given in Collett (1991) and also described by Leemis and Trivedi (1996). PROC FREQ does not provide exact confidence bounds for the risk differences. Refer to Agresti (1992) for a discussion of...
issues involved in constructing exact confidence bounds for differences of proportions.

**Odds Ratio and Relative Risks for 2×2 Tables**

**Odds Ratio (Case-Control Studies)**

The odds ratio is a useful measure of association for a variety of study designs. For a retrospective design called a *case-control study*, the odds ratio can be used to estimate the relative risk when the probability of positive response is small (Agresti 1990). In a case-control study, two independent samples are identified based on a binary (yes-no) response variable, and the conditional distribution of a binary explanatory variable is examined, within fixed levels of the response variable. Refer to Stokes, Davis, and Koch (1995) and Agresti (1996).

The odds of a positive response (column 1) in row 1 is \( n_{11}/n_{12} \). Similarly, the odds of a positive response in row 2 is \( n_{21}/n_{22} \). The odds ratio is formed as the ratio of the row 1 odds to the row 2 odds. The odds ratio for 2×2 tables is defined as

\[
\text{OR} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}
\]

The odds ratio can be any nonnegative number. When the row and column variables are independent, the true value of the odds ratio equals 1. An odds ratio greater than 1 indicates that the odds of a positive response are higher in row 1 than in row 2. Values less than 1 indicate the odds of positive response are higher in row 2. The strength of association increases with the deviation from 1.

The transformation \( G = (\text{OR} - 1)/(\text{OR} + 1) \) transforms the odds ratio to the range \((-1, 1)\) with \( G = 0 \) when \( \text{OR} = 1 \); \( G = -1 \) when \( \text{OR} = 0 \); and \( G \) approaches 1 as \( \text{OR} \) approaches infinity. \( G \) is the gamma statistic, which PROC FREQ computes when you specify the MEASURES option.

The asymptotic 100(1 − \( \alpha \))% confidence bounds for the odds ratio are

\[
( \text{OR} \cdot \exp(-z\sqrt{v}), \text{OR} \cdot \exp(z\sqrt{v}) )
\]

where

\[
v = \text{var}(\ln \text{OR}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}
\]

and \( z \) is the 100(1 − \( \alpha/2 \)) percentile of the standard normal distribution. If any of the four cell frequencies are zero, the estimates are not computed.

When you specify option OR in the EXACT statement, PROC FREQ computes exact confidence bounds for the odds ratio using an iterative algorithm based on that presented by Thomas (1971). Because this is a discrete problem, the confidence coefficient for these exact confidence bounds is not exactly 1 − \( \alpha \) but is at least 1 − \( \alpha \). Thus, these confidence bounds are conservative. Refer to Agresti (1992).
Relative Risks (Cohort Studies)

These measures of relative risk are useful in cohort (prospective) study designs, where two samples are identified based on the presence or absence of an explanatory factor. The two samples are observed in future time for the binary (yes-no) response variable under study. Relative risk measures are also useful in cross-sectional studies, where two variable are observed simultaneously. Refer to Stokes, Davis, and Koch (1995) and Agresti (1996).

The column 1 relative risk is the ratio of the column 1 risks for row 1 to row 2. The column 1 risk for row 1 is the proportion of the row 1 observations classified in column 1,

\[ p_{1|1} = \frac{n_{11}}{n_1}. \]

Similarly, the column 1 risk for row 2 is

\[ p_{1|2} = \frac{n_{21}}{n_2}. \]

The column 1 relative risk is then computed as

\[ RR_1 = \frac{p_{1|1}}{p_{1|2}} \]

A relative risk greater than 1 indicates that the probability of positive response is greater in row 1 than in row 2. Similarly, a relative risk less than 1 indicates that the probability of positive response is less in row 1 than in row 2. The strength of association increases with the deviation from 1.

The asymptotic 100(1 - \( \alpha \))% confidence bounds for the column 1 relative risk are

\( \left( RR_1 \cdot \exp(-z \sqrt{v}), \ RR_1 \cdot \exp(z \sqrt{v}) \right) \)

where

\[ v = \text{var}(\ln RR_1) = \frac{1 - p_{1|1}}{n_{11}} + \frac{1 - p_{1|2}}{n_{21}} \]

and \( z \) is the 100(1 - \( \alpha \)/2) percentile of the standard normal distribution. If either \( n_{11} \) or \( n_{21} \) is zero, the estimates are not computed.

PROC FREQ computes the column 2 relative risks in a similar manner.

**Cochran-Armitage Test for Trend**

The TREND option in the TABLES statement requests the Cochran-Armitage test for trend, which tests for trend in binomial proportions across levels of a single factor or covariate. This test is appropriate for a contingency table where one variable has two levels and the other variable is ordinal. The two-level variable represents the response, and the other variable represents an explanatory variable with ordered levels. When the contingency table has two columns and \( R \) rows, PROC FREQ tests
for trend across the \( R \) levels of the row variable, and the binomial proportion is computed as the proportion of observations in the first column. When the table has two rows and \( C \) columns, PROC FREQ tests for trend across the \( C \) levels of the column variable, and the binomial proportion is computed as the proportion of observations in the first row.

The trend test is based upon the regression coefficient for the weighted linear regression of the binomial proportions on the scores of the levels of the explanatory variable. Refer to Margolin (1988) and Agresti (1990). If the contingency table has two columns and \( R \) rows, the trend test statistic is computed as

\[
T = \frac{\sum_{i=1}^{R} n_{i1}(R_i - \bar{R})}{\sqrt{p_{1}(1 - p_{1})s^2}}
\]

where

\[
s^2 = \sum_{i=1}^{R} n_i(R_i - \bar{R})^2
\]

The row scores \( R_i \) are determined by the value of the SCORES= option in the TABLES statement. By default, PROC FREQ uses table scores. For character variables, the table scores for the row variable are the row numbers (for example, 1 for the first row, 2 for the second row, and so on). For numeric variables, the table score for each row is the numeric value of the row level. When you perform the trend test, the explanatory variable may be numeric (for example, dose of a test substance), and these variable values may be appropriate scores. If the explanatory variable has ordinal levels that are not numeric, you can assign meaningful scores to the variable levels. Sometimes equidistant scores, such as the table scores for a character variable, may be appropriate. For more information on choosing scores for the trend test, refer to Margolin (1988).

The null hypothesis for the Cochran-Armitage test is no trend, which means that the binomial proportion \( p_{i1} = n_{i1}/n_i \) is the same for all levels of the explanatory variable. Under this null hypothesis, the trend test statistic is asymptotically distributed as a standard normal random variable. In addition to this asymptotic test, PROC FREQ can compute the exact trend test, which you request by specifying the TREND option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

PROC FREQ computes one-sided and two-sided \( p \)-values for the trend test. When the test statistic is greater than its null hypothesis expected value of zero, PROC FREQ computes the right-sided \( p \)-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided \( p \)-value supports the alternative hypothesis of increasing trend in binomial proportions from row 1 to row \( R \). When the test statistic is less than or equal to zero, PROC FREQ outputs the left-sided \( p \)-value. A small left-sided \( p \)-value supports the alternative of decreasing trend. The one-sided \( p \)-value \( P_1 \) can be expressed as

\[
P_1 = \text{Prob} (\text{Trend Statistic} > T) \quad \text{if} \quad T > 0
\]
The Jonckheere-Terpstra test statistic is computed by first forming \( R(R-1)/2 \) Mann-Whitney counts \( M_{i,i'} \), where \( i < i' \), for pairs of rows in the contingency table,

\[
M_{i,i'} = \begin{cases} 
\text{number of times } X_{i,j} < X_{i',j'}, \\
\quad j = 1, \ldots, n_i; \quad j' = 1, \ldots, n_{i'} 
\end{cases} + \frac{1}{2} \begin{cases} 
\text{number of times } X_{i,j} = X_{i',j'}, \\
\quad j = 1, \ldots, n_i; \quad j' = 1, \ldots, n_{i'} 
\end{cases}
\]

where \( X_{i,j} \) is response \( j \) in row \( i \). Then the Jonckheere-Terpstra test statistic is computed as

\[
J = \sum_{1 \leq i < i' \leq R} \sum M_{i,i'}
\]

This test rejects the null hypothesis of no difference among classes for large values of \( J \). Asymptotic \( p \)-values for the Jonckheere-Terpstra test are obtained by using
the normal approximation for the distribution of the standardized test statistic. The standardized test statistic is computed as

\[ J^* = \frac{J - E_0(J)}{\sqrt{\text{var}_0(J)}} \]

where \( E_0(J) \) and \( \text{var}_0(J) \) are the expected value and variance of the test statistic under the null hypothesis.

\[ E_0(J) = \left( n^2 - \sum_i n_i^2 \right) / 4 \]

\[ \text{var}_0(J) = A / 72 + B / [36n(n - 1)(n - 2)] + C / [8n(n - 1)] \]

where

\[ A = n(n - 1)(2n + 5) - \sum_i n_i(n_i - 1)(2n_i + 5) - \sum_j n_j(n_j - 1)(2n_j + 5) \]

\[ B = \left[ \sum_i n_i(n_i - 1)(n_i - 2) \right] \left[ \sum_j n_j(n_j - 1)(n_j - 2) \right] \]

\[ C = \left[ \sum_i n_i(n_i - 1) \right] \left[ \sum_j n_j(n_j - 1) \right] \]

In addition to this asymptotic test, PROC FREQ can compute the exact Jonckheere-Terpstra test, which you request by specifying the JT option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

PROC FREQ computes one-sided and two-sided \( p \)-values for the Jonckheere-Terpstra test. When the standardized test statistic is greater than its null hypothesis expected value of zero, PROC FREQ computes the right-sided \( p \)-value, which is the probability of a larger value of the statistic occurring under the null hypothesis. A small right-sided \( p \)-value supports the alternative hypothesis of increasing order from row 1 to row \( R \). When the standardized test statistic is less than or equal to zero, PROC FREQ computes the left-sided \( p \)-value. A small left-sided \( p \)-value supports the alternative of decreasing order from row 1 to row \( R \). The one-sided \( p \)-value \( P_1 \) can be expressed as

\[ P_1 = \text{Prob} ( \text{Std JT Statistic} > J^* ) \quad \text{if } J^* > 0 \]

\[ P_1 = \text{Prob} ( \text{Std JT Statistic} < J^* ) \quad \text{if } J^* \leq 0 \]
The two-sided $p$-value $P_2$ is computed as

$$P_2 = \text{Prob}\left( |\text{Std JT Statistic}| > |J^*| \right)$$

**Tests and Measures of Agreement**

When you specify the AGREE option in the TABLES statement, PROC FREQ computes tests and measures of agreement for square tables (that is, for tables where the number of rows equals the number of columns). For two-way tables, these tests and measures include McNemar’s test for $2 \times 2$ tables, Bowker’s test of symmetry, the simple kappa coefficient, and the weighted kappa coefficient. For multiple strata ($n$-way tables, where $n > 2$), PROC FREQ computes the overall simple kappa coefficient and the overall weighted kappa coefficient, as well as tests for equal kappas (simple and weighted) among strata. Cochran’s $Q$ is computed for multi-way tables when each variable has two levels, that is, for $2 \times 2 \times \cdots \times 2$ tables.

PROC FREQ computes the kappa coefficients (simple and weighted), their asymptotic standard errors, and their confidence bounds when you specify the AGREE option in the TABLES statement. If you also specify the KAPPA option in the TEST statement, then PROC FREQ computes the asymptotic test of the hypothesis that simple kappa equals zero. Similarly, if you specify the WTKAP option in the TEST statement, PROC FREQ computes the asymptotic test for weighted kappa.

In addition to the asymptotic tests described in this section, PROC FREQ computes the exact $p$-value for McNemar’s test when you specify the option MCNEM in the EXACT statement. For the kappa statistics, PROC FREQ computes the exact test of the hypothesis that kappa (or weighted kappa) equals zero when you specify the option KAPPA (or WTKAP) in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

The discussion of each test and measures of agreement provides the formulas that PROC FREQ uses to compute the AGREE statistics. For information on the use and interpretation of these statistics, refer to Agresti (1990), Agresti (1996), Fleiss (1981), and the other references cited for each statistic.

**McNemar’s Test**

PROC FREQ computes McNemar’s test for $2 \times 2$ tables when you specify the AGREE option. McNemar’s test is appropriate when you are analyzing data from matched pairs of subjects with a dichotomous (yes-no) response. It tests the null hypothesis of marginal homogeneity, or $p_{11} = p_{12}$. McNemar’s test is computed as

$$Q_M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Under the null hypothesis, $Q_M$ has an asymptotic chi-square distribution with one degree of freedom. Refer to McNemar (1947), as well as the references cited in the preceding section. In addition to the asymptotic test, PROC FREQ also computes the exact $p$-value for McNemar’s test when you specify the MCNEM option in the EXACT statement.
Bowker’s Test of Symmetry

For Bowker’s test of symmetry, the null hypothesis is that the probabilities in the square table satisfy symmetry or that \( p_{ij} = p_{ji} \) for all pairs of table cells. When there are more than two categories, Bowker’s test of symmetry is calculated as

\[
Q_B = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}
\]

For large samples, \( Q_B \) has an asymptotic chi-square distribution with \( R(R - 1)/2 \) degrees of freedom under the null hypothesis of symmetry of the expected counts. Refer to Bowker (1948). For two categories, this test of symmetry is identical to McNemar’s test.

Simple Kappa Coefficient

The simple kappa coefficient, introduced by Cohen (1960), is a measure of interrater agreement:

\[
\hat{\kappa} = \frac{P_o - P_e}{1 - P_e}
\]

where \( P_o = \sum_i p_{ii} \) and \( P_e = \sum_i p_i p_i \). If the two response variables are viewed as two independent ratings of the \( n \) subjects, the kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds chance agreement, kappa is positive, with its magnitude reflecting the strength of agreement. Although this is unusual in practice, kappa is negative when the observed agreement is less than chance agreement. The minimum value of kappa is between \(-1\) and 0, depending on the marginal proportions.

The asymptotic variance of the simple kappa coefficient can be estimated by the following, according to Fleiss, Cohen, and Everitt (1969):

\[
\text{var} = \frac{A + B - C}{(1 - P_e)^2 n}
\]

where

\[
A = \sum_i p_{ii} \left[ 1 - (p_i + p_i)(1 - \hat{\kappa}) \right]^2
\]

\[
B = (1 - \hat{\kappa})^2 \sum_{i \neq j} \sum p_{ij}(p_i + p_j)^2
\]

and

\[
C = \left[ \hat{\kappa} - P_e(1 - \hat{\kappa}) \right]^2
\]
PROC FREQ computes confidence bounds for the simple kappa coefficient according to

\[ \hat{k} \pm z_{\alpha/2} \cdot \sqrt{\text{var}(\hat{k})} \]

where \( z_{\alpha/2} \) is the 100(1 - \( \alpha \)/2) percentile of the standard normal distribution. The value of \( \alpha \) is determined by the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence bounds.

To compute an asymptotic test for the kappa coefficient, PROC FREQ uses a standardized test statistic \( \hat{k}^* \), which has an asymptotic standard normal distribution under the null hypothesis that kappa equals zero. The standardized test statistic is computed as

\[ \hat{k}^* = \frac{\hat{k}}{\sqrt{\text{var}_0(\hat{k})}} \]

where \( \text{var}_0(\hat{k}) \) is the variance of the kappa coefficient under the null hypothesis.

\[ \text{var}_0(\hat{k}) = \frac{P_e + P_e^2 - \sum_i p_i p_i (p_i + p_i)}{(1 - P_e)^2 n} \]

Refer to Fleiss (1981).

In addition to the asymptotic test for kappa, PROC FREQ computes the exact test when you specify the KAPPA or AGREE option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

**Weighted Kappa Coefficient**

The weighted kappa coefficient is a generalization of the simple kappa coefficient, using weights to quantify the relative difference between categories. For \( 2 \times 2 \) tables, the weighted kappa coefficient equals the simple kappa coefficient. PROC FREQ displays the weighted kappa coefficient only for tables larger than \( 2 \times 2 \). PROC FREQ computes the weights from the column scores, using either the Cicchetti-Allison weight type or the Fleiss-Cohen weight type, both of which are described in the following section. The weights \( w_{ij} \) are constructed so that 0 ≤ \( w_{ij} \) < 1 for all \( i \neq j \), \( w_{ii} = 1 \) for all \( i \), and \( w_{ij} = w_{ji} \). The weighted kappa coefficient is defined as

\[ \hat{k}_w = \frac{P_{o[w]} - P_{e[w]}}{1 - P_{e[w]}} \]

where

\[ P_{o[w]} = \sum_i \sum_j w_{ij} p_{ij} \]
and

\[ P_{e[w]} = \sum_i \sum_j w_{ij} p_i p_j \]

The asymptotic variance of the weighted kappa coefficient can be estimated by the following, according to Fleiss, Cohen, and Everitt (1969):

\[ \text{var} = \frac{\sum_i \sum_j p_{ij} \left[ w_{ij} - (\overline{w}_i + \overline{w}_j)(1 - \hat{\kappa}_w) \right]^2}{(1 - P_{e[w]})^2 n} - \left[ \hat{\kappa}_w - P_{e[w]}(1 - \hat{\kappa}_w) \right]^2 \]

where

\[ \overline{w}_i = \sum_j p_{.j} w_{ij} \]

and

\[ \overline{w}_j = \sum_i p_{i.} w_{ij} \]

PROC FREQ computes confidence bounds for the weighted kappa coefficient according to

\[ \hat{\kappa}_w \pm z_{\alpha/2} \cdot \sqrt{\text{var}} \]

where \( z_{\alpha/2} \) is the 100(1 - \( \alpha/2 \)) percentile of the standard normal distribution. The value of \( \alpha \) is determined by the value of the ALPHA= option, which, by default, equals 0.05 and produces 95% confidence bounds.

To compute an asymptotic test for the weighted kappa coefficient, PROC FREQ uses a standardized test statistic \( \hat{\kappa}_w^* \), which has an asymptotic standard normal distribution under the null hypothesis that weighted kappa equals zero. The standardized test statistic is computed as

\[ \hat{\kappa}_w^* = \frac{\hat{\kappa}_w}{\sqrt{\text{var}_0(\hat{\kappa}_w)}} \]

where \( \text{var}_0(\hat{\kappa}_w) \) is the variance of the weighted kappa coefficient under the null hypothesis.

\[ \text{var}_0(\hat{\kappa}_w) = \frac{\sum_i \sum_j p_{ij} \left[ w_{ij} - (\overline{w}_i + \overline{w}_j) \right]^2 - P_{e[w]}^2}{(1 - P_{e[w]})^2 n} \]

Refer to Fleiss (1981).
In addition to the asymptotic test for weighted kappa, PROC FREQ computes the exact test when you specify the WTKAP or AGREE option in the EXACT statement. See the section “Exact Statistics” beginning on page 1271 for information on exact tests.

**Weights**

PROC FREQ computes kappa coefficient weights using the column scores and one of two available weight types. The column scores are determined by the SCORES= option in the TABLES statement. The two available weight types are Cicchetti-Allison and Fleiss-Cohen, and PROC FREQ uses the Cicchetti-Allison type by default. If you specify (WT=FC) with the AGREE option, then PROC FREQ uses the Fleiss-Cohen weight type to construct kappa weights.

PROC FREQ computes Cicchetti-Allison kappa coefficient weights using a form similar to that given by Cicchetti and Allison (1971).

\[ w_{ij} = 1 - \frac{|C_i - C_j|}{C_C - C_1} \]

where \( C_i \) is the score for column \( i \), and \( C \) is the number of categories or columns. You can specify the score type using the SCORES= option in the TABLES statement; if you do not specify the SCORES= option, PROC FREQ uses table scores. For numeric variables, table scores are the values of the numeric row and column headings. You can assign numeric values to the categories in a way that reflects their level of similarity. For example, suppose you have four categories and order them according to similarity. If you assign them values of 0, 2, 4, and 10, the following weights are used for computing the weighted kappa coefficient: \( w_{12} = 0.8, w_{13} = 0.6, w_{14} = 0, w_{23} = 0.8, w_{24} = 0.2, \) and \( w_{34} = 0.4 \). Note that when there are only two categories (that is, \( C = 2 \)), the weighted kappa coefficient is identical to the simple kappa coefficient.

If you specify (WT=FC) with the AGREE option in the TABLES statement, PROC FREQ computes Fleiss-Cohen kappa coefficient weights using a form similar to that given by Fleiss and Cohen (1973).

\[ w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_C - C_1)^2} \]

For the preceding example, the weights used for computing the weighted kappa coefficient are: \( w_{12} = 0.96, w_{13} = 0.84, w_{14} = 0, w_{23} = 0.96, w_{24} = 0.36, \) and \( w_{34} = 0.64 \).

**Overall Kappa Coefficient**

When there are multiple strata, PROC FREQ combines the stratum-level estimates of kappa into an overall estimate of the supposed common value of kappa. Assume there are \( q \) strata, indexed by \( h = 1, 2, \ldots, q \), and let \( \text{var}(\kappa_h) \) denote the squared standard error of \( \kappa_h \). Then the estimate of the overall kappa, according to Fleiss (1981), is computed as

\[ \hat{\kappa}_{overall} = \frac{\sum_{h=1}^{q} \frac{\hat{\kappa}_h}{\text{var}(\kappa_h)}}{\sum_{h=1}^{q} \frac{1}{\text{var}(\kappa_h)}} \]
An estimate of the overall weighted kappa is computed in a similar manner.

**Tests for Equal Kappa Coefficients**

The following chi-square statistic, with \( q - 1 \) degrees of freedom, is used to test whether the values of kappa are equal among the \( q \) strata:

\[
Q_K = \sum_{h=1}^{q} \left( \frac{\hat{\kappa}_h - \hat{\kappa}_\text{overall}}{\text{var}(\hat{\kappa}_h)} \right)^2
\]

A similar test is performed for weighted kappa coefficients.

**Cochran’s Q Test**

Cochran’s \( Q \) is computed for multi-way tables when each variable has two levels, that is, for \( 2 \times 2 \cdot \cdot \cdot 2 \) tables. Cochran’s \( Q \) statistic is used to test the homogeneity of the one-dimensional margins. Let \( m \) denote the number of variables and \( N \) denote the total number of subjects. Then Cochran’s \( Q \) statistic is computed as

\[
Q_C = (m - 1) \frac{m \sum_{j=1}^{m} T_j^2 - T^2}{mT - \sum_{k=1}^{N} S_k^2}
\]

where \( T_j \) is the number of positive responses for variable \( j \), \( T \) is the total number of positive responses over all variables, and \( S_k \) is the number of positive responses for subject \( k \). Under the null hypothesis, Cochran’s \( Q \) is an approximate chi-square statistic with \( m - 1 \) degrees of freedom. Refer to Cochran (1950). When there are only two binary response variables \( (m = 2) \), Cochran’s \( Q \) simplifies to McNemar’s test. When there are more than two response categories, you can test for marginal homogeneity using the repeated measures capabilities of the CATMOD procedure.

**Tables with Zero Rows and Columns**

For multiway tables, PROC FREQ does not compute CHISQ or MEASURES statistics for a stratum with a zero row or a zero column, because most of these statistics are undefined in this case. For a two-way table where there is no stratification, the analysis includes only those levels that occur with nonzero weight. However, PROC FREQ does compute AGREE statistics for stratified tables with a zero row or a zero column. The analysis includes all row and column variable levels that occur in any stratum. It does not include levels that do not occur in any stratum, even if such observations are in the data set with zero weight, because PROC FREQ does not process observations with zero weights (as described in the section “WEIGHT Statement” on page 1231).

To include a variable level with no observations in the analysis, you can assign an extremely small weight (such as 1E-8) to an observation with that variable level. Then the analysis includes this variable level, but the statistic value remains unchanged because the weight is so small. For example, suppose you need to compute a kappa coefficient for data from two raters. One rater uses all possible ratings (say, 1, 2, 3, 4, and 5), but another rater uses only four of the available ratings (1, 2, 3, and 4). You can create an observation where the second rater uses the rating level 5 and assign it
Cochran-Mantel-Haenszel Statistics

For \( n \)-way crosstabulation tables, consider the following example:

```plaintext
proc freq;
  tables A*B*C*D / cmh;
run;
```

The CMH option in the TABLES statement gives a stratified statistical analysis of the relationship between \( C \) and \( D \), after controlling for \( A \) and \( B \). The stratified analysis provides a way to adjust for the possible confounding effects of \( A \) and \( B \) without being forced to estimate parameters for them. The analysis produces Cochran-Mantel-Haenszel statistics, and for \( 2 \times 2 \) tables, it includes estimation of the common odds ratio, common relative risks, and the Breslow-Day test for homogeneity of the odds ratios.

Let the number of strata be denoted by \( q \), indexing the strata by \( h = 1, 2, \ldots, q \). Each stratum contains a contingency table with \( X \) representing the row variable and \( Y \) representing the column variable. For table \( h \), denote the cell frequency in row \( i \) and column \( j \) by \( n_{hij} \), with corresponding row and column marginal totals denoted by \( n_{hi} \) and \( n_{hj} \), and the overall stratum total by \( n_h \).

Because the formulas for the Cochran-Mantel-Haenszel statistics are more easily defined in terms of matrices, the following notation is used. Vectors are presumed to be column vectors unless they are transposed \( (') \).

\[
\begin{align*}
  \mathbf{n}'_{hi} &= (n_{h1i}, n_{h2i}, \ldots, n_{hiC}) & (1 \times C) \\
  \mathbf{n}'_h &= (\mathbf{n}'_{h1}, \mathbf{n}'_{h2}, \ldots, \mathbf{n}'_{hR}) & (1 \times RC) \\
  \mathbf{p}_{hi} &= \frac{n_{hi}}{n_h} & (1 \times 1) \\
  \mathbf{p}_{hj} &= \frac{n_{hj}}{n_h} & (1 \times 1) \\
  \mathbf{P}'_{h*} &= (p_{h1*}, p_{h2*}, \ldots, p_{hR*}) & (1 \times R) \\
  \mathbf{P}'_{h.*} &= (p_{h.*1}, p_{h.*2}, \ldots, p_{h.C*}) & (1 \times C)
\end{align*}
\]

Assume that the strata are independent and that the marginal totals of each stratum are fixed. The null hypothesis, \( H_0 \), is that there is no association between \( X \) and \( Y \) in any of the strata. The corresponding model is the multiple hypergeometric; this implies that, under \( H_0 \), the expected value and covariance matrix of the frequencies are, respectively,

\[
\begin{align*}
  \mathbf{m}_h &= \mathbf{E}[\mathbf{n}_h | H_0] = n_h (\mathbf{P}_{h.*} \otimes \mathbf{P}_{h.*}) \\
  \text{var}[\mathbf{n}_h | H_0] &= c \left( (\mathbf{D}_{\mathbf{P}_{h.*}} - \mathbf{P}_{h.*} \mathbf{P}'_{h.*}) \otimes (\mathbf{D}_{\mathbf{P}_{h.*}} - \mathbf{P}_{h.*} \mathbf{P}'_{h.*}) \right)
\end{align*}
\]
where

\[ c = \frac{n_h^2}{n_h - 1} \]

and where \( \otimes \) denotes Kronecker product multiplication and \( D_a \) is a diagonal matrix with elements of \( a \) on the main diagonal.

The generalized CMH statistic (Landis, Heyman, and Koch 1978) is defined as

\[ Q_{CMH} = G'V_G^{-1}G \]

where

\[
\begin{align*}
G &= \sum_h B_h (n_h - m_h) \\
V_G &= \sum_h B_h (\text{Var}(n_h | H_0)) B'_h
\end{align*}
\]

and where

\[ B_h = C_h \otimes R_h \]

is a matrix of fixed constants based on column scores \( C_h \) and row scores \( R_h \). When the null hypothesis is true, the CMH statistic has an asymptotic chi-square distribution with degrees of freedom equal to the rank of \( B_h \). If \( V_G \) is found to be singular, PROC FREQ prints a message and sets the value of the CMH statistic to missing.

PROC FREQ computes three CMH statistics using this formula for the generalized CMH statistic, with different row and column score definitions for each statistic. The CMH statistics that PROC FREQ computes are the correlation statistic, the ANOVA (row mean scores) statistic, and the general association statistic. These statistics test the null hypothesis of no association against different alternative hypotheses. The following sections describe the computation of these CMH statistics.

**Caution:** The CMH statistics have low power for detecting an association in which the patterns of association for some of the strata are in the opposite direction of the patterns displayed by other strata. Thus, a nonsignificant CMH statistic suggests either that there is no association or that no pattern of association has enough strength or consistency to dominate any other pattern.

**Correlation Statistic**

The correlation statistic, popularized by Mantel and Haenszel (1959) and Mantel (1963), has one degree of freedom and is known as the Mantel-Haenszel statistic.

The alternative hypothesis for the correlation statistic is that there is a linear association between \( X \) and \( Y \) in at least one stratum. If either \( X \) or \( Y \) does not lie on an ordinal (or interval) scale, then this statistic is not meaningful.
To compute the correlation statistic, PROC FREQ uses the formula for the generalized CMH statistic with the row and column scores determined by the SCORES= option in the TABLES statement. See the section “Scores” on page 1235 for more information on the available score types. The matrix of row scores \( \mathbf{R}_h \) has dimension \( 1 \times R \), and the matrix of column scores \( \mathbf{C}_h \) has dimension \( 1 \times C \).

When there is only one stratum, this CMH statistic reduces to \((n - 1)r^2\), where \( r \) is the Pearson correlation coefficient between \( X \) and \( Y \). When nonparametric (RANK or RIDIT) scores are specified, then the statistic reduces to \((n - 1)r_s^2\), where \( r_s \) is the Spearman rank correlation coefficient between \( X \) and \( Y \). When there is more than one stratum, then this CMH statistic becomes a stratum-adjusted correlation statistic.

### ANOVA (Row Mean Scores) Statistic

The ANOVA statistic can be used only when the column variable \( Y \) lies on an ordinal (or interval) scale so that the mean score of \( Y \) is meaningful. For the ANOVA statistic, the mean score is computed for each row of the table, and the alternative hypothesis is that, for at least one stratum, the mean scores of the \( R \) rows are unequal. In other words, the statistic is sensitive to location differences among the \( R \) distributions of \( Y \).

The matrix of column scores \( \mathbf{C}_h \) has dimension \( 1 \times C \), the column scores are determined by the SCORES= option. The matrix of row scores \( \mathbf{R}_h \) has dimension \( (R - 1) \times R \) and is created internally by PROC FREQ as

\[
\mathbf{R}_h = [\mathbf{I}_{R-1}, -\mathbf{J}_{R-1}]
\]

where \( \mathbf{I}_{R-1} \) is an identity matrix of rank \( R - 1 \), and \( \mathbf{J}_{R-1} \) is an \((R - 1) \times 1\) vector of ones. This matrix has the effect of forming \( R - 1 \) independent contrasts of the \( R \) mean scores.

When there is only one stratum, this CMH statistic is essentially an analysis of variance (ANOVA) statistic in the sense that it is a function of the variance ratio \( F \) statistic that would be obtained from a one-way ANOVA on the dependent variable \( Y \). If nonparametric scores are specified in this case, then the ANOVA statistic is a Kruskal-Wallis test.

If there is more than one stratum, then this CMH statistic corresponds to a stratum-adjusted ANOVA or Kruskal-Wallis test. In the special case where there is one subject per row and one subject per column in the contingency table of each stratum, this CMH statistic is identical to Friedman’s chi-square. See Example 26.8 on page 1300 for an illustration.

### General Association Statistic

The alternative hypothesis for the general association statistic is that, for at least one stratum, there is some kind of association between \( X \) and \( Y \). This statistic is always interpretable because it does not require an ordinal scale for either \( X \) or \( Y \).

For the general association statistic, the matrix \( \mathbf{R}_h \) is the same as the one used for the ANOVA statistic. The matrix \( \mathbf{C}_h \) is defined similarly as

\[
\mathbf{C}_h = [\mathbf{I}_{C-1}, -\mathbf{J}_{C-1}]
\]
PROC FREQ generates both score matrices internally. When there is only one stratum, then the general association CMH statistic reduces to \( Q_P(n - 1)/n \), where \( Q_P \) is the Pearson chi-square statistic. When there is more than one stratum, then the CMH statistic becomes a stratum-adjusted Pearson chi-square statistic. Note that a similar adjustment can be made by summing the Pearson chi-squares across the strata. However, the latter statistic requires a large sample size in each stratum to support the resulting chi-square distribution with \( q(R - 1)(C - 1) \) degrees of freedom. The CMH statistic requires only a large overall sample size since it has only \( (R - 1)(C - 1) \) degrees of freedom.

Refer to Cochran (1954); Mantel and Haenszel (1959); Mantel (1963); Birch (1965); Landis, Heyman, and Koch (1978).

**Adjusted Odds Ratio and Relative Risk Estimates**

The CMH option provides adjusted odds ratio and relative risk estimates for stratified 2×2 tables. For each of these measures, PROC FREQ computes the Mantel-Haenszel estimate and the logit estimate. These estimates apply to \( n \)-way table requests in the TABLES statement, when the row and column variables both have only two levels. For example,

```sas
proc freq;
   tables A*B*C*D / cmh;
run;
```

In this example, if the row and columns variables C and D both have two levels, PROC FREQ provides odds ratio and relative risk estimates, adjusting for the confounding variables A and B.

The choice of an appropriate measure depends on the study design. For case-control (retrospective) studies, the odds ratio is appropriate. For cohort (prospective) or cross-sectional studies, the relative risk is appropriate. See the section “Odds Ratio and Relative Risks for 2×2 Tables” beginning on page 1253 for more information on these measures.

Throughout this section, \( z \) denotes the \( 100(1 - \alpha/2) \) percentile of the standard normal distribution.

**Odds Ratio, Case-Control Studies**

**Mantel-Haenszel Estimator**

The Mantel-Haenszel estimate of the common odds ratio is computed as

\[
OR_{MH} = \frac{\sum_h n_{h11} n_{h22}/n_h}{\sum_h n_{h12} n_{h21}/n_h}
\]

It is always computed unless the denominator is zero. Refer to Mantel and Haenszel (1959) and Agresti (1990).

Using the estimated variance for \( \log(OR_{MH}) \) given by Robins, Breslow, and Greenland (1986), PROC FREQ computes the corresponding \( 100(1 - \alpha)\% \) confidence
bounds for the odds ratio as

\[
\left( \text{OR}_{\text{MH}} \cdot \exp(-z\hat{\sigma}), \ OR_{\text{MH}} \cdot \exp(z\hat{\sigma}) \right)
\]

where

\[
\hat{\sigma}^2 = \text{var}[\ln(\text{OR}_{\text{MH}})] = \\
= \frac{\sum_h (n_{h11} + n_{h22})(n_{h11} n_{h22})/w_h^2}{2 \left( \sum_h n_{h11} n_{h22}/n_h \right)^2} + \frac{\sum_h [(n_{h11} + n_{h22})(n_{h12} n_{h21}) + (n_{h12} + n_{h21})(n_{h11} n_{h22})]/n_h^2}{2 \left( \sum_h n_{h11} n_{h22}/n_h \right) \left( \sum_h n_{h12} n_{h21}/n_h \right)} + \frac{\sum_h (n_{h12} + n_{h21})(n_{h12} n_{h21})/n_h^2}{2 \left( \sum_h n_{h12} n_{h21}/n_h \right)^2}
\]

Note that the Mantel-Haenszel odds ratio estimator is less sensitive to small \(n_h\) than the logit estimator.

**Logit Estimator**

The adjusted logit estimate of the odds ratio (Woolf 1955) is computed as

\[
\text{OR}_L = \exp \left( \frac{\sum_h w_h \ln(\text{OR}_h)}{\sum_h w_h} \right)
\]

and the corresponding \(100(1 - \alpha)\)% confidence bounds are

\[
\left( \text{OR}_L \cdot \exp \left( \frac{-z}{\sqrt{\sum_h w_h}} \right), \ \text{OR}_L \cdot \exp \left( \frac{z}{\sqrt{\sum_h w_h}} \right) \right)
\]

where \(\text{OR}_h\) is the odds ratio for stratum \(h\), and

\[
w_h = \frac{1}{\text{var}(\ln \text{OR}_h)}
\]

If any cell frequency in a stratum \(h\) is zero, then PROC FREQ adds 0.5 to each cell of the stratum before computing \(\text{OR}_h\) and \(w_h\) (Haldane 1955), and prints a warning.

**Relative Risks, Cohort Studies**

**Mantel-Haenszel Estimator**

The Mantel-Haenszel estimate of the common relative risk for column 1 is computed as

\[
\text{RR}_{\text{MH}} = \frac{\sum_h n_{h11} n_{h22}/n_h}{\sum_h n_{h21} n_{h11}/n_h}
\]
It is always computed unless the denominator is zero. Refer to Mantel and Haenszel (1959) and Agresti (1990).

Using the estimated variance for $\log(\text{RR}_{\text{MH}})$ given by Greenland and Robins (1985), PROC FREQ computes the corresponding $100(1 - \alpha)\%$ confidence bounds for the relative risk as

$$\left( \text{RR}_{\text{MH}} \cdot \exp(-z\hat{\sigma}), \text{RR}_{\text{MH}} \cdot \exp(z\hat{\sigma}) \right)$$

where

$$\hat{\sigma}^2 = \text{var}[\ln(\text{RR}_{\text{MH}})] = \frac{\sum_h(n_{h1} \cdot n_{h2} \cdot n_{h1} - n_{h11} \cdot n_{h21} \cdot n_h)/n_h^2}{(\sum_h n_{h11} \cdot n_{h2} / n_h)(\sum_h n_{h21} \cdot n_{h1} / n_h)}$$

**Logit Estimator**

The adjusted logit estimate of the common relative risk for column 1 is computed as

$$\text{RR}_L = \exp\left( \frac{\sum_h w_h \ln \text{RR}_h}{\sum w_h} \right)$$

and the corresponding $100(1 - \alpha)\%$ confidence bounds are

$$\left( \text{RR}_L \exp\left( \frac{-z}{\sqrt{\sum_h w_h}} \right), \text{RR}_L \exp\left( \frac{z}{\sqrt{\sum_h w_h}} \right) \right)$$

where $\text{RR}_h$ is the column 1 relative risk estimate for stratum $h$, and

$$w_h = \frac{1}{\text{var}(\ln \text{RR}_h)}$$

If $n_{h11}$ or $n_{h21}$ is zero, then PROC FREQ adds 0.5 to each cell of the stratum before computing $\text{RR}_h$ and $w_h$, and prints a warning. Refer to Kleinbaum, Kupper, and Morgenstern (1982, Sections 17.4 and 17.5).

**Breslow-Day Test for Homogeneity of the Odds Ratios**

When you specify the CMH option, PROC FREQ computes the Breslow-Day test for stratified analysis of $2 \times 2$ tables. It tests the null hypothesis that the odds ratios for the $q$ strata are all equal. When the null hypothesis is true, the statistic has an asymptotic chi-square distribution with $q - 1$ degrees of freedom.

The Breslow-Day statistic is computed as

$$Q_{\text{BD}} = \frac{\sum_h (n_{h11} - E(n_{h11} | \text{OR}_{\text{MH}}))^2}{\text{var}(n_{h11} | \text{OR}_{\text{MH}})}$$
where $E$ and $\text{var}$ denote expected value and variance, respectively. The summation does not include any table with a zero row or column. If $\text{OR}_{\text{MH}}$ equals zero or if it is undefined, then PROC FREQ does not compute the statistic and prints a warning message.

**Caution:** Unlike the Cochran-Mantel-Haenszel statistics, the Breslow-Day test requires a large sample size within each stratum, and this limits its usefulness. In addition, the validity of the CMH tests does not depend on any assumption of homogeneity of the odds ratios; therefore, the Breslow-Day test should never be used as such an indicator of validity.

Refer to Breslow and Day (1994).

**Exact Statistics**

Exact statistics can be useful in situations where the asymptotic assumptions are not met, and so the asymptotic $p$-values are not close approximations for the true $p$-values. Standard asymptotic methods involve the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. When the sample size is not large, asymptotic results may not be valid, with the asymptotic $p$-values differing perhaps substantially from the exact $p$-values. Asymptotic results may also be unreliable when the distribution of the data is sparse, skewed, or heavily tied. Refer to Agresti (1996) and Bishop, Fienberg, and Holland (1975). Exact computations are based on the statistical theory of exact conditional inference for contingency tables, reviewed by Agresti (1992).

PROC FREQ provides exact $p$-values for the following tests for two-way tables: Pearson chi-square, likelihood-ratio chi-square, Mantel-Haenszel chi-square, Fisher’s exact test, Jonckheere-Terpstra test, Cochran-Armitage test for trend, and McNemar’s test. PROC FREQ also computes exact $p$-values for tests of hypotheses that the following statistics equal zero: Pearson correlation coefficient, Spearman correlation coefficient, simple kappa coefficient, and weighted kappa coefficient. Additionally, PROC FREQ computes exact confidence bounds for the odds ratio for $2 \times 2$ tables. For one-way frequency tables, PROC FREQ provides the exact chi-square goodness-of-fit test (for equal proportions or for proportions or frequencies that you specify). Also for one-way tables, PROC FREQ provides exact confidence bounds for the binomial proportion and an exact test for the binomial proportion value.

The following sections summarize the exact computational algorithms, define the exact $p$-values that PROC FREQ computes, and discuss the computational resource requirements.

**Computational Algorithms**

PROC FREQ computes exact $p$-values for general $R \times C$ tables using the network algorithm developed by Mehta and Patel (1983). This algorithm provides a substantial advantage over direct enumeration, which can be very time-consuming and feasible only for small problems. Refer to Agresti (1992) for a review of algorithms for computation of exact $p$-values, and refer to Mehta, Patel, and Tsiatis (1984) and Mehta, Patel, and Senchaudhuri (1991) for information on the performance of the network algorithm.

The reference set for a given contingency table is the set of all contingency tables
with the observed marginal row and column sums. Corresponding to this reference set, the network algorithm forms a directed acyclic network consisting of nodes in a number of stages. A path through the network corresponds to a distinct table in the reference set. The distances between nodes are defined so that the total distance of a path through the network is the corresponding value of the test statistic. At each node, the algorithm computes the shortest and longest path distances for all the paths that pass through that node. For statistics that can be expressed as a linear combination of cell frequencies multiplied by increasing row and column scores, PROC FREQ computes shortest and longest path distances using the algorithm given in Agresti, Mehta, and Patel (1990). For statistics of other forms, PROC FREQ computes an upper bound for the longest path and a lower bound for the shortest path, following the approach of Valz and Thompson (1994).

The longest and shortest path distances or bounds for a node are compared to the value of the test statistic to determine whether all paths through the node contribute to the p-value, none of the paths through the node contribute to the p-value, or neither of these situations occur. If all paths through the node contribute, the p-value is incremented accordingly, and these paths are eliminated from further analysis. If no paths contribute, these paths are eliminated from the analysis. Otherwise, the algorithm continues, still processing this node and the associated paths. The algorithm finishes when all nodes have been accounted for, incrementing the p-value accordingly, or eliminated.

In applying the network algorithm, PROC FREQ uses full precision to represent all statistics, row and column scores, and other quantities involved in the computations. Although it is possible to use rounding to improve the speed and memory requirements of the algorithm, PROC FREQ does not do this since it can result in reduced accuracy of the p-values.

PROC FREQ computes exact confidence bounds for the odds ratio according to an iterative algorithm based on that presented by Thomas (1971). Refer also to Gart (1971). Because this is a discrete problem, the confidence coefficient is not exactly $1 - \alpha$, but it is at least $1 - \alpha$. Thus, these confidence bounds are conservative.

For one-way tables, PROC FREQ computes the exact chi-square goodness-of-fit test by the method of Radlow and Alf (1975). PROC FREQ generates all possible one-way tables with the observed total sample size and number of categories. For each possible table, PROC FREQ compares its chi-square value with the value for the observed table. If the table’s chi-square value is greater than or equal to the observed chi-square, PROC FREQ increments the exact p-value by the probability of that table, which is calculated under the null hypothesis using the multinomial frequency distribution. By default, the null hypothesis states that all categories have equal proportions. If you specify null hypothesis proportions or frequencies using the TESTP= or TESTF= option in the TABLES statement, then PROC FREQ calculates the exact chi-square test based on that null hypothesis.

For binomial proportions in one-way tables, PROC FREQ computes exact confidence bounds using the $F$ distribution method given in Collett (1991) and also described by Leemis and Trivedi (1996). PROC FREQ computes the exact test for a binomial proportion ($H_0 : p = p_0$) by summing binomial probabilities over all alternatives.
See the section “Binomial Proportion” on page 1250 for details. By default, PROC FREQ uses \( p_0 = 0.5 \) as the null hypothesis proportion. Alternatively, you can specify the null hypothesis proportion with the P= option in the TABLES statement.

**Definition of p-Values**

For several tests in PROC FREQ, the test statistic is nonnegative, and large values of the test statistic indicate a departure from the null hypothesis. Such tests include Pearson’s chi-square, the likelihood-ratio chi-square, the Mantel-Haenszel chi-square, Fisher’s exact test for tables larger than \( 2 \times 2 \) tables, McNemar’s test, and the one-way chi-square goodness-of-fit test. The exact \( p \)-value for these nondirectional tests is the sum of probabilities for those tables having a test statistic greater than or equal to the value of the observed test statistic.

There are other tests where it may be appropriate to test against either a one-sided or a two-sided alternative hypothesis. For example, when you test the null hypothesis that the true parameter value equals 0 (\( T = 0 \)), the alternative of interest may be one-sided (\( T \leq 0 \), or \( T \geq 0 \)) or two-sided (\( T \neq 0 \)). Such tests include the Pearson correlation coefficient, Spearman correlation coefficient, Jonckheere-Terpstra test, Cochran-Armitage test for trend, simple kappa coefficient, and weighted kappa coefficient. For these tests, PROC FREQ outputs the right-sided \( p \)-value when the observed value of the test statistic is greater than its expected value. The right-sided \( p \)-value is the sum of probabilities for those tables having a test statistic greater than or equal to the observed test statistic. Otherwise, when the test statistic is less than or equal to its expected value, PROC FREQ outputs the left-sided \( p \)-value. The left-sided \( p \)-value is the sum of probabilities for those tables having a test statistic less than or equal to the one observed. The one-sided \( p \)-value \( P_1 \) can be expressed as

\[
P_1 = \text{Prob} \left( \text{Test Statistic} \geq t \right) \quad \text{if} \quad t > E_0(T)
\]

\[
P_1 = \text{Prob} \left( \text{Test Statistic} \leq t \right) \quad \text{if} \quad t \leq E_0(T)
\]

where \( t \) is the observed value of the test statistic and \( E_0(T) \) is the expected value of the test statistic under the null hypothesis. PROC FREQ computes the two-sided \( p \)-value as the sum of the one-sided \( p \)-value and the corresponding area in the opposite tail of the distribution of the statistic, equidistant from the expected value. The two-sided \( p \)-value \( P_2 \) can be expressed as

\[
P_2 = \text{Prob} \left( |\text{Test Statistic} - E_0(T)| \geq |t - E_0(T)| \right)
\]

**Computational Resources**

PROC FREQ uses relatively fast and efficient algorithms for exact computations. These recently developed algorithms, together with improvements in computer power, make it feasible now to perform exact computations for data sets where previously only asymptotic methods could be applied. Nevertheless, there are still large problems that may require a prohibitive amount of time and memory for exact computations, depending on the speed and memory available on your computer. For large problems, consider whether exact methods are really needed or whether asymptotic
methods might give results quite close to the exact results, while requiring much less computer time and memory.

A formula does not exist that can predict in advance how much time and memory are needed to compute an exact $p$-value for a certain problem. The time and memory required depend on several factors, including which test is being performed, the total sample size, the number of rows and columns, and the specific arrangement of the observations into table cells. Generally, larger problems (in terms of total sample size, number of rows, and number of columns) tend to require more time and memory. Additionally, for a fixed total sample size, time and memory requirements tend to increase as the number of rows and columns increases, since this corresponds to an increase in the number of tables in the reference set. Also for a fixed sample size, time and memory requirements increase as the marginal row and column totals become more homogeneous. Refer to Agresti, Mehta, and Patel (1990) and Gail and Mantel (1977).

At any time while PROC FREQ is computing exact $p$-values, you can terminate the computations by pressing the system interrupt key sequence (refer to the SAS Companion for your system) and choosing to stop computations. After you terminate exact computations, PROC FREQ completes all other remaining tasks. The procedure produces the requested output and reports missing values for any exact $p$-values that were not computed by the time of termination.

You can also use the MAXTIME= option in the EXACT statement to limit the amount of time PROC FREQ uses for exact computations. You specify a MAXTIME= value that is the maximum amount of time (in seconds) that PROC FREQ can use to compute an exact $p$-value. If PROC FREQ does not finish computing an exact $p$-value within that time, it terminates the computation and completes all other remaining tasks.

### Computational Resources

For each variable in a table request, PROC FREQ stores all of the levels in memory. If all variables are numeric and not formatted, this requires about 84 bytes for each variable level. When there are character variables or formatted numeric variables, the memory that is required depends on the formatted variable lengths, with longer formatted lengths requiring more memory. The number of levels for each variable is limited only by the largest integer that your operating environment can store.

For any single crosstabulation table requested, PROC FREQ builds the entire table in memory, regardless of whether the table has zero cell counts. Thus, if the numeric variables A, B, and C each have 10 levels, PROC FREQ requires 2520 bytes to store the variable levels for the table request A*B*C, as follows:

$$3 \text{ variables} \times 10 \text{ levels/variable} \times 84 \text{ bytes/level}$$

In addition, PROC FREQ requires 8000 bytes to store the table cell frequencies

$$1000 \text{ cells} \times 8 \text{ bytes/cell}$$
even though there may be only 10 observations.

When the variables have many levels or when there are many multiway tables, your computer may not have enough memory to construct the tables. If PROC FREQ runs out of memory while constructing tables, it stops collecting levels for the variable with the most levels and returns the memory that is used by that variable. The procedure then builds the tables that do not contain the disabled variables.

If there is not enough memory for your table request and if increasing the available memory is impractical, you can reduce the number of multiway tables or variable levels. If you are not using the CMH or AGREE option in the TABLES statement to compute statistics across strata, reduce the number of multiway tables by using PROC SORT to sort the data set by one or more of the variables or by using the DATA step to create an index for the variables. Then remove the sorted or indexed variables from the TABLES statement and include a BY statement that uses these variables. You can also reduce memory requirements by using a FORMAT statement in the PROC FREQ step to reduce the number of levels. Additionally, reducing the formatted variable lengths reduces the amount of memory that is needed to store the variable levels. For more information on using formats, see the “Grouping with Formats” section on page 1232.

---

**Output Data Sets**

PROC FREQ produces two types of output data sets that you can use with other statistical and reporting procedures. These data sets are produced as follows:

- specifying a TABLES statement with an OUT= option creates an output data set that contains frequency or crosstabulation table counts and percentages.
- specifying an OUTPUT statement creates an output data set that contains statistics.

PROC FREQ does not display the output data sets. Use PROC PRINT, PROC REPORT, or any other SAS reporting tool to display an output data set.

**Contents of the TABLES Statement Output Data Set**

The OUT= option in the TABLES statement creates an output data set that contains one observation for each combination of the variable values (or table cell) in the last table request. By default, each observation contains the frequency and percentage for the table cell. When the input data set contains missing values, the output data set also contains an observation with the frequency of missing values. The output data set includes the following variables:

- BY variables
- table request variables, such as A, B, C, and D in the table request A*B*C*D
- COUNT, a variable containing the cell frequency
- PERCENT, a variable containing the cell percentage
If you specify the OUTEXPECT and OUTPCT options in the TABLES statement, the output data set also contains expected frequencies and row, column, and table percentages, respectively. The additional variables are

- **EXPECTED**, a variable containing the expected frequency
- **PCT_TABL**, a variable containing the percentage of two-way table frequency, for \( n \)-way tables where \( n > 2 \)
- **PCT_ROW**, a variable containing the percentage of row frequency
- **PCT_COL**, a variable containing the percentage of column frequency

When you submit the following statements

```sas
proc freq;
   tables A A*B / out=D;
run;
```

the output data set D contains frequencies and percentages for the last table request, \( A \times B \). If \( A \) has two levels (1 and 2), \( B \) has three levels (1, 2, and 3), and no table cell count is zero or missing, the output data set D includes six observations, one for each combination of \( A \) and \( B \). The first observation corresponds to \( A=1 \) and \( B=1 \); the second observation corresponds to \( A=1 \) and \( B=2 \); and so on. The data set includes the variables COUNT and PERCENT. The value of COUNT is the number of observations with the given combination of \( A \) and \( B \) values. The value of PERCENT is the percent of the total number of observations having that \( A \) and \( B \) combination.

When PROC FREQ combines different variable values into the same formatted level, the output data set contains the smallest internal value for the formatted level. For example, suppose a variable \( X \) has the values 1.1, 1.4, 1.7, 2.1, and 2.3. When you submit the statement

```sas
format X 1.;
```

in a PROC FREQ step, the formatted levels listed in the frequency table for \( X \) are 1 and 2. If you create an output data set with the frequency counts, the internal values of \( X \) are 1.1 and 1.7. To report the internal values of \( X \) when you display the output data set, use a format of 3.1 with \( X \).

**Contents of the OUTPUT Statement Output Data Set**

The OUTPUT statement creates a SAS data set containing the statistics that PROC FREQ computes for the last table request. You specify which statistics to store in the output data set. There is an observation with the specified statistics for each stratum or two-way table. If PROC FREQ computes summary statistics for a stratified table, the output data set also contains a summary observation with those statistics.

The OUTPUT data set can include the following variables:

- **BY** variables
- variables that identify the stratum, such as \( A \) and \( B \) in the table request \( A*B*C*D \)
The output data set also includes variables with the p-values and degrees of freedom, asymptotic standard error (ASE), or confidence bounds when PROC FREQ computes these values for a specified statistic.

The variable names for the specified statistics in the output data set are the names of the options enclosed in underscores. PROC FREQ forms variable names for the corresponding p-values, degrees of freedom, or confidence bounds by combining the name of the option with the appropriate prefix from the following list:

- DF_  – degrees of freedom
- E_   – asymptotic standard error (ASE)
- L_   – lower confidence bound
- U_   – upper confidence bound
- E0_  – ASE under the null hypothesis
- Z_   – standardized value
- P_   – p-value
- P2_  – two-sided p-value
- PL_  – left-sided p-value
- PR_  – right-sided p-value
- XP_  – exact p-value
- XP2_ – exact two-sided p-value
- XPL_ – exact left-sided p-value
- XPR_ – exact right-sided p-value
- XL_  – exact lower confidence bound
- XR_  – exact upper confidence bound

For example, variable names created for the Pearson chi-square, its degrees of freedom, its p-values are _PCHI_, DF_PCHI, and P_PCHI, respectively.

If the length of the prefix plus the statistic option exceeds eight characters, PROC FREQ truncates the option so that the name of the new variable is eight characters long.

---

**Displayed Output**

PROC FREQ displays one-way frequency tables for all one-way table requests in the TABLES statements, unless you specify the NOPRINT option in the PROC statement or the NOPRINT option in the TABLES statement. For a one-way table showing the frequency distribution of a single variable, PROC FREQ displays the following information:

- the name of the variable and its values
- Frequency counts, giving the number of observations that have each value
- specified Test Frequency counts, if you specify the CHISQ and TESTF= options to request a chi-square goodness-of-fit test for specified frequencies
- Percent, giving the percentage of the total number of observations with that value. (The NOPERCENT option suppresses this information.)
- specified Test Percents, if you specify the CHISQ and TESTP= options to request a chi-square goodness-of-fit test for specified percents. (The NOPERCENT option suppresses this information.)

- Cumulative Frequency counts, giving the sum of the frequency counts of that value and all other values listed above it in the table. The last cumulative frequency is the total number of nonmissing observations. (The NOCUM option suppresses this information.)

- Cumulative Percent values, giving the percentage of the total number of observations with that value and all others previously listed in the table. (The NOCUM or the NOPERCENT option suppresses this information.)

- Frequency Missing, or the number of observations with missing values

For one-way tables, two statistical options are available in the TABLES statement. The CHISQ option provides a chi-square goodness-of-fit test, and the BINOMIAL option provides binomial proportion statistics. PROC FREQ displays the following information, unless you specify the NOPRINT option in the PROC statement:

- If you specify the CHISQ option for a one-way table, PROC FREQ provides a chi-square goodness-of-fit test, displaying the Chi-Square statistic, the degrees of freedom (DF), and the probability value (Pr > ChiSq). If you specify the CHISQ option in the EXACT statement, PROC FREQ also displays the exact probability value for this test.

- If you specify the BINOMIAL option for a one-way table, PROC FREQ displays the estimate of the binomial Proportion, which is the proportion of observations in the first class listed in the one-way table. PROC FREQ also displays the asymptotic standard error (ASE) and the asymptotic and exact confidence bounds for this estimate. For the binomial proportion test, PROC FREQ displays the asymptotic standard error under the null hypothesis (ASE Under H0), the standardized test statistic (Z), and the one-sided and two-sided probability values. If you specify the BINOMIAL option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test.

PROC FREQ displays all multiway table requests in the TABLES statements, unless you specify the NOPRINT option in the PROC statement or the NOPRINT option in the TABLES statement. PROC FREQ displays multiway tables either as crosstabulation tables (the default) or as lists (when you specify the LIST option).

For two-way to multiway crosstabulation tables, the values of the last variable in the table request form the table columns. The values of the next-to-last variable form the rows. Each level (or combination of levels) of the other variables form one stratum. PROC FREQ produces a separate two-way crosstabulation table for each stratum. Each cell of a crosstabulation table may contain the following information:

- Frequency, giving the number of observations that have the indicated values of the two variables. (The NOFREQ option suppresses this information.)
- the Expected cell frequency under the hypothesis of independence, if you specify the EXPECTED option
- the Deviation of the cell frequency from the expected value, if you specify the DEVIATION option
- Cell Chi-Square, which is the cell’s contribution to the total chi-square statistic, if you specify the CELLCHI2 option
- Tot Pct, or the cell’s percentage of the total frequency, for n-way tables when \( n > 2 \), if you specify the TOTPCT option
- Percent, the cell’s percentage of the total frequency. (The NOPERCENT option suppresses this information.)
- Row Pct, or the row percentage, the cell’s percentage of the total frequency count for that cell’s row. (The NOROW option suppresses this information.)
- Col Pct, or column percentage, the cell’s percentage of the total frequency count for that cell’s column. (The NOCOL option suppresses this information.)
- Cumulative Col%, or cumulative column percent, if you specify the CUMCOL option
- Frequency Missing, or the number of observations with missing values

If you specify the LIST option in the TABLES statement, PROC FREQ displays multiway tables in a list format rather than as crosstabulation tables. PROC FREQ ignores the LIST option when you also request statistical options. For a multiway table in list format, PROC FREQ displays the following information:

- the variable names and values
- Frequency counts, giving the number of observations with the indicated combination of variable values
- Percent, the cell’s percentage of the total number of observations. (The NOPERCENT option suppresses this information.)
- Cumulative Frequency counts, giving the sum of the frequency counts of that cell and all other cells listed above it in the table. The last cumulative frequency is the total number of nonmissing observations. (The NOCUM option suppresses this information.)
- Cumulative Percent values, giving the percentage of the total number of observations for that cell and all others previously listed in the table. (The NOCUM or the NOPERCENT option suppresses this information.)
- Frequency Missing, or the number of observations with missing values

PROC FREQ computes statistical tests and measures for crosstabulation tables, depending on which statements and options you specify. You can suppress the display of all these results by specifying the NOPRINT option in the PROC statement. With any of the following information, PROC FREQ also displays the Sample Size and the Frequency Missing.
• If you specify the SCOROUT option, PROC FREQ displays the Row Scores and Column Scores that it uses for statistical computations. The Row Scores table displays the row variable values and the Score corresponding to each value. The Column Scores table displays the column variable values and the corresponding Scores. PROC FREQ also identifies the score type used to compute the row and column scores. You can specify the score type with the SCORES= option in the TABLES statement.

• If you specify the CHISQ option, PROC FREQ displays the following statistics for each two-way table: Pearson Chi-Square, Likelihood-Ratio Chi-Square, Continuity-Adjusted Chi-Square, Mantel-Haenszel Chi-Square, Fisher’s Exact Test (for 2 × 2 tables), the Phi Coefficient, the Contingency Coefficient, and Cramer’s V. For each test statistic, PROC FREQ also displays the degrees of freedom (DF) and the probability value (Prob). If you specify the PCHI option, the LRCHI option, or the MHCHI option in the EXACT statement, PROC FREQ also displays the exact probability value for the Pearson Chi-Square, the Likelihood-Ratio Chi-Square, or the Mantel-Haenszel Chi-Square, respectively. If you specify the CHISQ option in the EXACT statement, PROC FREQ displays exact probability values for all three of these chi-square tests.

• If you specify the FISHER option in the TABLES statement (or, equivalently, the FISHER option in the EXACT statement), PROC FREQ displays the probability value for Fisher’s exact test for tables larger than 2 × 2. In addition, PROC FREQ displays the CHISQ output listed earlier, even if you do not also specify the CHISQ option.

• If you specify the MEASURES option, PROC FREQ displays the following statistics and their asymptotic standard errors (ASE) for each two-way table: Gamma, Kendall’s Tau-b, Stuart’s Tau-c, Somers’ D(C|R), Somers’ D(R|C), Pearson Correlation, Spearman Correlation, Lambda Asymmetric (C|R), Lambda Symmetric, Uncertainty Coefficient (C|R), Uncertainty Coefficient (R|C), and Uncertainty Coefficient Symmetric. If you specify the CL option, PROC FREQ also displays confidence bounds for these measures.

• If you specify the PLCORR option, PROC FREQ displays the tetrachoric correlation for 2 × 2 tables or the polychoric correlation for larger tables. In addition, PROC FREQ displays the MEASURES output listed earlier, even if you do not also specify the MEASURES option.

• If you specify the option GAMMA, KENTB, STUTC, SMDCR, SMDRC, PCORR, or SCORR in the TEST statement, PROC FREQ displays asymptotic tests for Gamma, Kendall’s Tau-b, Stuart’s Tau-c, Somers’ D(C|R), Somers’ D(R|C), the Pearson Correlation, or the Spearman Correlation, respectively. If you specify the MEASURES option in the TEST statement, PROC FREQ displays all these asymptotic tests. The test output includes the statistic, its asymptotic standard error (ASE), Confidence Bounds, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the one-sided and two-sided probability values.

• If you specify the PCORR or SCORR option in the EXACT statement, PROC FREQ displays asymptotic and exact tests for the Pearson Correlation or the
Spearman Correlation, respectively. The test output includes the correlation, its asymptotic standard error (ASE), Confidence Bounds, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the asymptotic and exact one-sided and two-sided probability values.

- If you specify the RISKDIFF option for $2 \times 2$ tables, PROC FREQ displays the Column 1 and Column 2 Risk Estimates. For each column, PROC FREQ displays Row 1 Risk, Row 2 Risk, Total Risk, and Risk Difference, together with their asymptotic standard errors (ASE), Asymptotic Confidence Bounds, and Exact Confidence Bounds. Exact confidence bounds are not available for the risk difference.

- If you specify the MEASURES option or the RELRISK option for $2 \times 2$ tables, PROC FREQ displays Estimates of the Relative Risk for Case-Control and Cohort studies, together with their Confidence Bounds. These measures are also known as the Odds Ratio and the Column 1 and 2 Relative Risks. If you specify the OR option in the EXACT statement, PROC FREQ also displays Exact Confidence Bounds for the Odds Ratio.

- If you specify the TREND option, PROC FREQ displays the Cochran-Armitage Trend Test for tables that are $2 \times C$ or $R \times 2$. For this test, PROC FREQ gives the Statistic (Z) and the one-sided and two-sided probability values. If you specify the TREND option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test.

- If you specify the JT option, PROC FREQ displays the Jonckheere-Terpstra Test, showing the Statistic (JT), the standardized test statistic (Z), and the one-sided and two-sided probability values. If you specify the JT option in the EXACT statement, PROC FREQ also displays the exact one-sided and two-sided probability values for this test.

- If you specify the AGREE option and the PRINTKWT option, PROC FREQ displays the Kappa Coefficient Weights for square tables greater than $2 \times 2$.

- If you specify the AGREE option, for two-way tables PROC FREQ displays McNemar’s Test and the Simple Kappa Coefficient for $2 \times 2$ tables. For square tables larger than $2 \times 2$, PROC FREQ displays Bowker’s Test of Symmetry, the Simple Kappa Coefficient, and the Weighted Kappa Coefficient. For McNemar’s Test and Bowker’s Test of Symmetry, PROC FREQ displays the Statistic (S), the degrees of freedom (DF), and the probability value (Pr > S). If you specify the MCNEM option in the EXACT statement, PROC FREQ also displays the exact probability value for McNemar’s test. For the simple and weighted kappa coefficients, PROC FREQ displays the kappa values, asymptotic standard errors (ASE), and Confidence Bounds.

- If you specify the KAPPA or WTKAP option in the TEST statement, PROC FREQ displays asymptotic tests for the simple kappa coefficient or the weighted kappa coefficient, respectively. If you specify the AGREE option in the TEST statement, PROC FREQ displays both these asymptotic tests. The test output includes the kappa coefficient, its asymptotic standard error (ASE), Confidence Bounds, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the one-sided and two-sided probability values.
If you specify the KAPPA or WTKAP option in the EXACT statement, PROC FREQ displays asymptotic and exact tests for the simple kappa coefficient or the weighted kappa coefficient, respectively. The test output includes the kappa coefficient, its asymptotic standard error (ASE), Confidence Bounds, the ASE under the null hypothesis H0, the standardized test statistic (Z), and the asymptotic and exact one-sided and two-sided probability values.

If you specify the AGREE option, for multiple strata PROC FREQ displays Overall Simple and Weighted Kappa Coefficients, with their asymptotic standard errors (ASE) and Confidence Bounds. PROC FREQ also displays Tests for Equal Kappa Coefficients, giving the Chi-Squares, degrees of freedom (DF), and probability values (Pr > ChiSq) for the Simple Kappa and Weighted Kappa. For multiple strata of 2 × 2 tables, PROC FREQ displays Cochran’s Q, giving the Statistic (Q), the degrees of freedom (DF), and the probability value (Pr > Q).

If you specify the CMH option, PROC FREQ displays Cochran-Mantel-Haenszel Statistics for the following three alternative hypotheses: Nonzero Correlation, Row Mean Scores Differ (ANOVA Statistic), and General Association. For each of these statistics, PROC FREQ gives the degrees of freedom (DF) and the probability value (Prob). For 2 × 2 tables, PROC FREQ also displays Estimates of the Common Relative Risk for Case-Control and Cohort studies, together with their confidence bounds. These include both Mantel-Haenszel and Logit stratum-adjusted estimates of the common Odds Ratio, Column 1 Relative Risk, and Column 2 Relative Risk. Also for 2 × 2 tables, PROC FREQ displays the Breslow-Day Test for Homogeneity of the Odds Ratios. For this test, PROC FREQ gives the Chi-Square, the degrees of freedom (DF), and the probability value (Pr > ChiSq).

**ODS Table Names**

PROC FREQ assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. For more information on ODS, see Chapter 14, “Using the Output Delivery System.”

**Table 26.11. ODS Tables Produced in PROC FREQ**

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement / Option*</th>
</tr>
</thead>
<tbody>
<tr>
<td>BinomialProp</td>
<td>Binomial proportion</td>
<td>TABLES / BINOMIAL (for one-way tables)</td>
</tr>
<tr>
<td>BinomialPropTest</td>
<td>Binomial proportion test</td>
<td>TABLES / BINOMIAL (for one-way tables)</td>
</tr>
<tr>
<td>BreslowDayTest</td>
<td>Breslow-Day test</td>
<td>TABLES / CMH (for h × 2 × 2 tables)</td>
</tr>
<tr>
<td>CMH</td>
<td>Cochran-Mantel-Haenszel statistics</td>
<td>TABLES / CMH</td>
</tr>
<tr>
<td>ChiSq</td>
<td>Chi-square tests and measures</td>
<td>TABLES / CHISQ</td>
</tr>
<tr>
<td>CochransQ</td>
<td>Cochran’s Q</td>
<td>TABLES / AGREE (for h × 2 × 2 tables)</td>
</tr>
<tr>
<td>ColumnScores</td>
<td>Column scores</td>
<td>TABLES / SCOROUT</td>
</tr>
<tr>
<td>CommonRelRisks</td>
<td>Common relative risks</td>
<td>TABLES / CMH</td>
</tr>
<tr>
<td>CrossTabFreqs</td>
<td>Cross-tabulation table</td>
<td>TABLES (n-way table request, n &gt; 1)</td>
</tr>
</tbody>
</table>
Table 26.11. (continued)

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement / Option*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EqualKappaTest</td>
<td>Test for equal kappas</td>
<td>TABLES / AGREE (for $h \times 2 \times 2$ tables)</td>
</tr>
<tr>
<td>EqualKappaTests</td>
<td>Tests for equal kappas</td>
<td>TABLES / AGREE (for $h \times r \times r$ tables, $r &gt; 2$)</td>
</tr>
<tr>
<td>Gamma</td>
<td>Gamma</td>
<td>TEST / GAMMA</td>
</tr>
<tr>
<td>GammaTest</td>
<td>Gamma test</td>
<td>TEST / GAMMA</td>
</tr>
<tr>
<td>JTTTest</td>
<td>Jonckheere-Terpstra test</td>
<td>TABLES / JT</td>
</tr>
<tr>
<td>KappaStatistics</td>
<td>Kappa statistics</td>
<td>TABLES / AGREE (for tables larger than $2 \times 2$, and no TEST or EXACT requests for kappas)</td>
</tr>
<tr>
<td>KappaWeights</td>
<td>Kappa weights</td>
<td>TABLES / AGREE and PRINTKWT</td>
</tr>
<tr>
<td>List</td>
<td>List frequencies</td>
<td>TABLES / LIST</td>
</tr>
<tr>
<td>McNemarsTest</td>
<td>McNemar’s test</td>
<td>TABLES / AGREE (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>Measures</td>
<td>Measures of association</td>
<td>TABLES / MEASURES</td>
</tr>
<tr>
<td>OneWayChiSq</td>
<td>One-way chi-square goodness-of-fit test</td>
<td>TABLES / CHISQ (for a one-way table)</td>
</tr>
<tr>
<td>OneWayFreqs</td>
<td>One-way frequencies</td>
<td>PROC (with no TABLES stmt) or TABLES (one-way table request)</td>
</tr>
<tr>
<td>OverallKappa</td>
<td>Overall kappa coefficient</td>
<td>TABLES / AGREE (for $h \times 2 \times 2$ tables)</td>
</tr>
<tr>
<td>OverallKappas</td>
<td>Overall kappa coefficients</td>
<td>TABLES / AGREE (for $h \times r \times r$ tables, $r &gt; 2$)</td>
</tr>
<tr>
<td>PearsonCorr</td>
<td>Pearson correlation coefficient</td>
<td>TEST or EXACT / PCORR</td>
</tr>
<tr>
<td>PearsonCorrTest</td>
<td>Pearson correlation test</td>
<td>TEST or EXACT / PCORR</td>
</tr>
<tr>
<td>RelativeRisks</td>
<td>Relative risk estimates</td>
<td>TABLES / RELRISK or MEASURES (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>RiskDiffCol1</td>
<td>Column 1 risk estimates</td>
<td>TABLES / RISKDIFF (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>RiskDiffCol2</td>
<td>Column 2 risk estimates</td>
<td>TABLES / RISKDIFF (for $2 \times 2$ tables)</td>
</tr>
<tr>
<td>RowScores</td>
<td>Row scores</td>
<td>TABLES / SCOROUT</td>
</tr>
<tr>
<td>SimpleKappa</td>
<td>Simple kappa coefficient</td>
<td>TEST or EXACT / KAPPA</td>
</tr>
<tr>
<td>SimpleKappaTest</td>
<td>Simple kappa test</td>
<td>(TEST or EXACT / KAPPA)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>or (TABLES / AGREE (for $2 \times 2$ tables))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and (TEST or EXACT / WTKAP))</td>
</tr>
<tr>
<td>SomersDCR</td>
<td>Somers’ $D(C</td>
<td>R)$</td>
</tr>
<tr>
<td>SomersDCRTest</td>
<td>Somers’ $D(C</td>
<td>R)$ test</td>
</tr>
<tr>
<td>SomersDRC</td>
<td>Somers’ $D(R</td>
<td>C)$</td>
</tr>
<tr>
<td>SomersDRCTest</td>
<td>Somers’ $D(R</td>
<td>C)$ test</td>
</tr>
<tr>
<td>SpearmanCorr</td>
<td>Spearman correlation coefficient</td>
<td>TEST or EXACT / SCORR</td>
</tr>
<tr>
<td>SpearmanCorrTest</td>
<td>Spearman correlation test</td>
<td>TEST or EXACT / SCORR</td>
</tr>
<tr>
<td>SymmetryTest</td>
<td>Test of symmetry</td>
<td>TABLES / AGREE</td>
</tr>
<tr>
<td>TauB</td>
<td>Kendall’s tau-$b$</td>
<td>TEST / KENTB</td>
</tr>
<tr>
<td>TauBTest</td>
<td>Kendall’s tau-$b$ test</td>
<td>TEST / KENTB</td>
</tr>
<tr>
<td>TauC</td>
<td>Stuart’s tau-$c$</td>
<td>TEST / STUTC</td>
</tr>
<tr>
<td>TauCTest</td>
<td>Stuart’s tau-$c$ test</td>
<td>TEST / STUTC</td>
</tr>
<tr>
<td>TrendTest</td>
<td>Cochran-Armitage test for trend</td>
<td>TABLES / TREND</td>
</tr>
</tbody>
</table>
Table 26.11. (continued)

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement / Option*</th>
</tr>
</thead>
<tbody>
<tr>
<td>WeightedKappa</td>
<td>Weighted kappa coefficient</td>
<td>(TEST or EXACT / WTKAP) or (TABLES / AGREE (for (r \times r) tables, (r &gt; 2)) and TEST or EXACT / KAPPA)</td>
</tr>
<tr>
<td>WeightedKappaTest</td>
<td>Weighted kappa test</td>
<td>TEST or EXACT / WTKAP</td>
</tr>
</tbody>
</table>

* The ALL option in the TABLES statement invokes CHISQ, MEASURES, and CMH.

Examples

Example 26.1. Creating an Output Data Set with Table Cell Frequencies

The eye and hair color of children from two different regions of Europe are recorded in the data set Color. Instead of recording one observation per child, the data are recorded as cell counts, where the variable Count contains the number of children exhibiting each of the 15 eye and hair color combinations. The data set does not include missing combinations.

```sas
data Color;
  input Region Eyes $ Hair $ Count @@;
  label Eyes = 'Eye Color'
  Hair = 'Hair Color'
  Region = 'Geographic Region';
  datalines;
  1 blue fair 23 1 blue red 7 1 blue medium 24
  1 blue dark 11 1 green fair 19 1 green red 7
  1 green medium 18 1 green dark 14 1 brown fair 34
  1 brown red 5 1 brown medium 41 1 brown dark 40
  1 brown black 3 2 blue fair 46 2 blue red 21
  2 blue medium 44 2 blue dark 40 2 blue black 6
  2 green fair 50 2 green red 31 2 green medium 37
  2 green dark 23 2 brown fair 56 2 brown red 42
  2 brown medium 53 2 brown dark 54 2 brown black 13;
```

The following statements read the Color data set and create an output data set containing the frequencies, percentages, and expected cell frequencies of the Eyes by Hair two-way table. The TABLES statement requests three tables: Eyes and Hair frequency tables and an Eyes by Hair crosstabulation table. The OUT= option creates the FreqCnt data set, which contains the crosstabulation table frequencies. The OUTEXPECT option outputs the expected cell frequencies to FreqCnt, and the SPARSE option includes the zero cell counts. The WEIGHT statement specifies that Count contains the observation weights. The following statements create Output 26.1.1 through Output 26.1.3.
Example 26.1. Creating an Output Data Set with Table Cell...

```
proc freq data=Color;
  weight Count;
  tables Eyes Hair Eyes*Hair/out=FreqCnt outexpect sparse;
  title 'Eye and Hair Color of European Children';
run;
proc print data=FreqCnt noobs;
  title2 'Output Data Set from PROC FREQ';
run;
```

Output 26.1.1. Frequency Table

```
<table>
<thead>
<tr>
<th>Eyes</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>222</td>
<td>29.13</td>
<td>222</td>
<td>29.13</td>
</tr>
<tr>
<td>brown</td>
<td>341</td>
<td>44.75</td>
<td>563</td>
<td>73.88</td>
</tr>
<tr>
<td>green</td>
<td>199</td>
<td>26.12</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hair</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>22</td>
<td>2.89</td>
<td>22</td>
<td>2.89</td>
</tr>
<tr>
<td>dark</td>
<td>182</td>
<td>23.88</td>
<td>204</td>
<td>26.77</td>
</tr>
<tr>
<td>fair</td>
<td>228</td>
<td>29.92</td>
<td>432</td>
<td>56.69</td>
</tr>
<tr>
<td>medium</td>
<td>217</td>
<td>28.48</td>
<td>649</td>
<td>85.17</td>
</tr>
<tr>
<td>red</td>
<td>113</td>
<td>14.83</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>
```
Output 26.1.2. Cross Tabulation Table

Eye and Hair Color of European Children

The FREQ Procedure

Table of Eyes by Hair

<table>
<thead>
<tr>
<th>Eyes (Eye Color)</th>
<th>Hair (Hair Color)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>black</td>
<td>6</td>
<td>0.79</td>
<td>2.70</td>
<td>27.27</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>dark</td>
<td>51</td>
<td>6.69</td>
<td>22.97</td>
<td>22.97</td>
<td>29.13</td>
</tr>
<tr>
<td></td>
<td>fair</td>
<td>69</td>
<td>9.06</td>
<td>31.08</td>
<td>31.08</td>
<td>24.78</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>68</td>
<td>8.92</td>
<td>30.63</td>
<td>30.63</td>
<td>24.78</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>28</td>
<td>3.67</td>
<td>12.61</td>
<td>12.61</td>
<td>14.83</td>
</tr>
<tr>
<td>brown</td>
<td>black</td>
<td>16</td>
<td>2.10</td>
<td>4.69</td>
<td>72.73</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td>dark</td>
<td>94</td>
<td>12.34</td>
<td>27.57</td>
<td>51.65</td>
<td>44.75</td>
</tr>
<tr>
<td></td>
<td>fair</td>
<td>90</td>
<td>11.81</td>
<td>26.39</td>
<td>39.47</td>
<td>31.08</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>94</td>
<td>12.34</td>
<td>27.57</td>
<td>43.32</td>
<td>27.57</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>47</td>
<td>6.17</td>
<td>13.78</td>
<td>41.59</td>
<td>13.78</td>
</tr>
<tr>
<td>green</td>
<td>black</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td>dark</td>
<td>37</td>
<td>4.86</td>
<td>18.59</td>
<td>55</td>
<td>26.12</td>
</tr>
<tr>
<td></td>
<td>fair</td>
<td>69</td>
<td>9.06</td>
<td>34.67</td>
<td>59.543</td>
<td>30.26</td>
</tr>
<tr>
<td></td>
<td>medium</td>
<td>55</td>
<td>7.22</td>
<td>27.64</td>
<td>56.671</td>
<td>24.92</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>38</td>
<td>4.98</td>
<td>25.35</td>
<td>29.510</td>
<td>14.03</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>22</td>
<td>2.89</td>
<td>22.82</td>
<td>29.92</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td></td>
<td>182</td>
<td>23.88</td>
<td>29.92</td>
<td>28.48</td>
<td>14.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>228</td>
<td>29.92</td>
<td>28.48</td>
<td>28.48</td>
<td>100.00</td>
</tr>
</tbody>
</table>

By default, PROC FREQ displays the variable values in alphabetical order (Output 26.1.1). The 'Eyes*Hair' specification produces a crosstabulation table (Output 26.1.2) with eye color defining the table rows and hair color defining the table columns. A zero cell count for green eyes and black hair indicates that this eye and hair color combination does not occur in the data.

Output 26.1.3. OUT= Data Set

Output Data Set from PROC FREQ

<table>
<thead>
<tr>
<th>Eyes</th>
<th>Hair</th>
<th>COUNT</th>
<th>EXPECTED</th>
<th>PERCENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>black</td>
<td>6</td>
<td>6.409</td>
<td>0.7874</td>
</tr>
<tr>
<td>blue</td>
<td>dark</td>
<td>51</td>
<td>53.024</td>
<td>6.6929</td>
</tr>
<tr>
<td>blue</td>
<td>fair</td>
<td>69</td>
<td>66.425</td>
<td>9.0551</td>
</tr>
<tr>
<td>blue</td>
<td>medium</td>
<td>68</td>
<td>63.220</td>
<td>8.9239</td>
</tr>
<tr>
<td>blue</td>
<td>red</td>
<td>28</td>
<td>32.921</td>
<td>3.6745</td>
</tr>
<tr>
<td>brown</td>
<td>black</td>
<td>16</td>
<td>9.845</td>
<td>2.0997</td>
</tr>
<tr>
<td>brown</td>
<td>dark</td>
<td>94</td>
<td>81.446</td>
<td>12.3360</td>
</tr>
<tr>
<td>brown</td>
<td>fair</td>
<td>90</td>
<td>102.031</td>
<td>11.8110</td>
</tr>
<tr>
<td>brown</td>
<td>medium</td>
<td>94</td>
<td>97.109</td>
<td>12.3360</td>
</tr>
<tr>
<td>brown</td>
<td>red</td>
<td>47</td>
<td>50.568</td>
<td>6.1680</td>
</tr>
<tr>
<td>green</td>
<td>black</td>
<td>0</td>
<td>5.745</td>
<td>0.0000</td>
</tr>
<tr>
<td>green</td>
<td>dark</td>
<td>37</td>
<td>47.530</td>
<td>4.8556</td>
</tr>
<tr>
<td>green</td>
<td>fair</td>
<td>69</td>
<td>59.543</td>
<td>9.0551</td>
</tr>
<tr>
<td>green</td>
<td>medium</td>
<td>55</td>
<td>56.671</td>
<td>7.2178</td>
</tr>
<tr>
<td>green</td>
<td>red</td>
<td>38</td>
<td>29.510</td>
<td>4.9869</td>
</tr>
</tbody>
</table>

The output data set (Output 26.1.3) contains frequency counts and percentages for the last table. The data set also includes an observation for the zero cell count (SPARSE) and a variable with the expected cell frequency for each table cell (OUTEXPECT).
Example 26.2. Computing Chi-square Tests for One-Way Frequency Tables

This example examines whether the children's hair color (from Example 26.1 on page 1284) has a specified multinomial distribution for the two regions. The hypothesized distribution for hair color is 30% fair, 12% red, 30% medium, 25% dark, and 3% black.

In order to test the hypothesis for each region, the data are first sorted by Region. Then the FREQ procedure uses a BY statement to produce a separate table for each BY group (Region). The option ORDER=DATA orders the frequency table values (hair color) by their order in the data set. The TABLES statement requests a frequency table for hair color, and the option NOCUM suppresses the display of the cumulative frequencies and percentages. The TESTP= option specifies the hypothesized percentages for the chi-square test; the number of percentages specified equals the number of table levels, and the percentages sum to 100. The following statements produce Output 26.2.1.

```sas
proc sort data=Color;
  by Region;
run;
proc freq data=Color order=data;
  weight Count;
  tables Hair/nocum testp=(30 12 30 25 3);
  by Region;
  title 'Hair Color of European Children';
run;
```

Output 26.2.1. One-way Frequency Table with BY Group

<table>
<thead>
<tr>
<th>Hair Color</th>
<th>Frequency</th>
<th>Percent</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fair</td>
<td>76</td>
<td>30.89</td>
<td>30.00</td>
</tr>
<tr>
<td>red</td>
<td>19</td>
<td>7.72</td>
<td>12.00</td>
</tr>
<tr>
<td>medium</td>
<td>83</td>
<td>33.74</td>
<td>30.00</td>
</tr>
<tr>
<td>dark</td>
<td>65</td>
<td>26.42</td>
<td>25.00</td>
</tr>
<tr>
<td>black</td>
<td>3</td>
<td>1.22</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Chi-Square Test for Specified Proportions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>7.7602</td>
</tr>
<tr>
<td>DF</td>
<td>4</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.1008</td>
</tr>
</tbody>
</table>
Output 26.2.1.  (continued)

Hair Color of European Children
----------------------------- Geographic Region=2 -----------------------------

The FREQ Procedure

Hair Color

<table>
<thead>
<tr>
<th>Hair</th>
<th>Frequency</th>
<th>Percent</th>
<th>Test Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fair</td>
<td>152</td>
<td>29.46</td>
<td>30.00</td>
</tr>
<tr>
<td>red</td>
<td>94</td>
<td>18.22</td>
<td>12.00</td>
</tr>
<tr>
<td>medium</td>
<td>134</td>
<td>25.97</td>
<td>30.00</td>
</tr>
<tr>
<td>dark</td>
<td>117</td>
<td>22.67</td>
<td>25.00</td>
</tr>
<tr>
<td>black</td>
<td>19</td>
<td>3.68</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Chi-Square Test for Specified Proportions

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.3824</td>
<td>4</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

The frequency tables list the variable values (hair color) in the order in which they appear in the data set. The “Test Percent” column lists the hypothesized percentages for the chi-square test. Always check that you have ordered the TESTP= percentages to correctly match the order of the variable levels.

PROC FREQ computes a chi-square statistic for each region. The chi-square statistic is significant at the 0.05 level for Region 2 (p=0.0003) but not for Region 1. This indicates a significant departure from the hypothesized percentages in Region 2.

Example 26.3.  Computing Binomial Proportions for One-Way Frequency Tables

The binomial proportion is computed as the proportion of observations for the first level of the variable that you are studying. The following statements compute the proportion of children with brown eyes (from the data set in Example 26.1 on page 1284) and test this value against the hypothesis that the proportion is 50%. Also, these statements test whether the proportion of children with fair hair is 28%.

```sas
proc freq data=Color order=freq;
  weight Count;
  tables Eyes / binomial alpha=.1;
  tables Hair / binomial(p=.28);
  title 'Hair and Eye Color of European Children';
run;
```

The first TABLES statement produces a frequency table for eye color. The BINOMIAL option computes the binomial proportion and confidence bounds, and it tests the hypothesis that the proportion for the first eye color level (brown) is 0.5. The option ALPHA=.1 specifies that 90% confidence bounds should be computed. The
second TABLES statement creates a frequency table for hair color and computes the binomial proportion and confidence bounds, but it tests that the proportion for the first hair color (fair) is 0.28. These statements produce Output 26.3.1 and Output 26.3.2.

**Output 26.3.1.** Binomial Proportion for Eye Color

<table>
<thead>
<tr>
<th>Hair and Eye Color of European Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>The FREQ Procedure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Eye Color</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Eyes</td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>brown</td>
</tr>
<tr>
<td>blue</td>
</tr>
<tr>
<td>green</td>
</tr>
</tbody>
</table>

**Binomial Proportion**
for Eyes = brown

| Proportion | 0.4475 |
| ASE        | 0.0180 |
| 90% Lower Conf Bound | 0.4179 |
| 90% Upper Conf Bound | 0.4771 |

**Exact Conf Bounds**
| 90% Lower Conf Bound | 0.4174 |
| 90% Upper Conf Bound | 0.4779 |

**Test of H0: Proportion = 0.5**

| ASE under H0 | 0.0181 |
| Z            | -2.8981 |
| One-sided Pr < | 0.0019 |
| Two-sided Pr > | Z | 0.0038 |

The frequency table in Output 26.3.1 displays the variable values in order of descending frequency count. Since the first variable level is 'brown', PROC FREQ computes the binomial proportion of children with brown eyes. PROC FREQ also computes its asymptotic standard error (ASE), and asymptotic and exact 90% confidence bounds. If you do not specify the ALPHA= option, then PROC FREQ computes the default 95% confidence bounds.

Because the value of $Z$ is less than zero, PROC FREQ computes a left-sided $p$-value (0.0019). This small $p$-value supports the alternative hypothesis that the true value of the proportion of children with brown eyes is less than 50%.
Output 26.3.2. Binomial Proportion for Hair Color

Hair and Eye Color of European Children

The FREQ Procedure

<table>
<thead>
<tr>
<th>Hair Color</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fair</td>
<td>228</td>
<td>29.92</td>
<td>228</td>
<td>29.92</td>
</tr>
<tr>
<td>medium</td>
<td>217</td>
<td>28.48</td>
<td>445</td>
<td>58.40</td>
</tr>
<tr>
<td>dark</td>
<td>182</td>
<td>23.88</td>
<td>627</td>
<td>82.28</td>
</tr>
<tr>
<td>red</td>
<td>113</td>
<td>14.83</td>
<td>740</td>
<td>97.11</td>
</tr>
<tr>
<td>black</td>
<td>22</td>
<td>2.89</td>
<td>762</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Binomial Proportion for Hair = fair

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>0.2992</td>
</tr>
<tr>
<td>ASE</td>
<td>0.0166</td>
</tr>
<tr>
<td>95% Lower Conf Bound</td>
<td>0.2667</td>
</tr>
<tr>
<td>95% Upper Conf Bound</td>
<td>0.3317</td>
</tr>
</tbody>
</table>

Exact Conf Bounds

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Lower Conf Bound</td>
<td>0.2669</td>
</tr>
<tr>
<td>95% Upper Conf Bound</td>
<td>0.3331</td>
</tr>
</tbody>
</table>

Test of H0: Proportion = 0.28

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE under H0</td>
<td>0.0163</td>
</tr>
<tr>
<td>Z</td>
<td>1.1812</td>
</tr>
<tr>
<td>One-sided Pr &gt; Z</td>
<td>0.1188</td>
</tr>
<tr>
<td>Two-sided Pr &gt;</td>
<td>Z</td>
</tr>
</tbody>
</table>

Output 26.3.2 displays the results from the second TABLES statement. PROC FREQ computes the default 95% confidence limits since the ALPHA= option is not specified. The value of $Z$ is greater than zero, so PROC FREQ computes a right-sided p-value (0.1188). This large p-value provides insufficient evidence to reject the null hypothesis that the proportion of children with fair hair is 28%.

Example 26.4. Analyzing a 2x2 Contingency Table

This example computes chi-square tests and Fisher’s exact test to compare the probability of coronary heart disease for two types of diet. It also estimates the relative risk and computes exact confidence bounds for the odds ratio.

The data set FatComp contains hypothetical data for a case-control study of high fat diet and the risk of coronary heart disease. The data are recorded as cell counts, where the variable Count contains the frequencies for each exposure and response combination. The data is sorted in descending order by the variables Exposure and Response, so that the first cell of the $2 \times 2$ table contains the frequency of positive exposure and positive response. The FORMAT procedure creates formats to identify the type of exposure and response with character values.
proc format;
  value ExpFmt 1='High Cholesterol Diet'
              0='Low Cholesterol Diet';
  value RspFmt 1='Yes'
              0='No';
run;

data FatComp;
  input Exposure Response Count;
  label Response='Heart Disease';
datalines;
  0 0 6
  0 1 2
  1 0 4
  1 1 11;
proc sort data=FatComp;
  by descending Exposure descending Response;
run;

In the following statements, the TABLES statement creates a two-way table, and the option ORDER=DATA orders the contingency table values by their order in the data set. The CHISQ option produces several chi-square tests, while the RELRISK option produces relative risk measures. The EXACT statement creates the exact Pearson chi-square test and exact confidence bounds for the odds ratio. These statements produce Output 26.4.1 through Output 26.4.3.

proc freq data=FatComp order=data;
  weight Count;
  tables Exposure*Response / chisq rellrisk;
  exact pchi or;
  format Exposure ExpFmt. Response RspFmt.;
  title 'Case-Control Study of High Fat/Cholesterol Diet';
run;
**Output 26.4.1.** Contingency Table

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Response (Heart Disease)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Cholesterol Diet</td>
<td>Yes</td>
<td>11</td>
<td>47.83</td>
<td>73.33</td>
<td>84.62</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>4</td>
<td>17.39</td>
<td>26.67</td>
<td>40.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>15</td>
<td>65.22</td>
<td>84.62</td>
<td>100.00</td>
</tr>
<tr>
<td>Low Cholesterol Diet</td>
<td>Yes</td>
<td>2</td>
<td>8.70</td>
<td>25.00</td>
<td>15.38</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>6</td>
<td>26.09</td>
<td>75.00</td>
<td>60.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>8</td>
<td>34.78</td>
<td>84.62</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The contingency table in Output 26.4.1 displays the variable values so that the first table cell contains the frequency for the first cell in the data set, the frequency of positive exposure and positive response.

**Output 26.4.2.** Chi-Square Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob (Asymptotic)</th>
<th>Prob (Exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>4.9597</td>
<td>0.0259</td>
<td>0.0393</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>5.0975</td>
<td>0.0240</td>
<td></td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>3.1879</td>
<td>0.0742</td>
<td></td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>4.7441</td>
<td>0.0294</td>
<td></td>
</tr>
<tr>
<td>Fisher’s Exact Test (Left)</td>
<td></td>
<td></td>
<td></td>
<td>0.9967</td>
</tr>
<tr>
<td>(Right)</td>
<td></td>
<td></td>
<td></td>
<td>0.0367</td>
</tr>
<tr>
<td>(2-Tail)</td>
<td></td>
<td></td>
<td></td>
<td>0.0393</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.4644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.4212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cramer’s V</td>
<td></td>
<td>0.4644</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

WARNING: 50% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test.

Sample Size = 23

Since the expected counts in some of the cells are small, PROC FREQ displays a warning that the asymptotic chi-square tests may not be appropriate. In this case, the exact tests in Output 26.4.2 are appropriate. The alternative hypothesis for this analysis states that coronary heart disease is more likely to be associated with a high
fat diet, so a one-sided test is desired. Fisher’s exact right-sided test analyzes whether the probability of heart disease in the high fat group exceeds the probability of heart disease in the low fat group; since this $p$-value is small, the alternative hypothesis is supported.

**Output 26.4.3. Relative Risk**

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Value</th>
<th>(Asymptotic) 95% Confidence Bounds</th>
<th>(Exact) 95% Confidence Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>8.2500</td>
<td>1.1535 59.0029</td>
<td>0.8677 105.5488</td>
</tr>
<tr>
<td>Cohort (Col1 Risk)</td>
<td>2.9333</td>
<td>0.8502 10.1204</td>
<td></td>
</tr>
<tr>
<td>Cohort (Col2 Risk)</td>
<td>0.3556</td>
<td>0.1403 0.9009</td>
<td></td>
</tr>
</tbody>
</table>

Sample Size = 23

The odds ratio, displayed in Output 26.4.3, provides an estimate of the relative risk when an event is rare. This estimate indicates that the odds of heart disease is 8.25 times higher in the high fat diet group; however, the wide confidence bounds indicate that this estimate has low precision.

**Example 26.5. Creating an Output Data Set Containing Chi-Square Statistics**

This example uses the Color data from Example 26.1 (page 1284) to output the Pearson chi-square and the likelihood-ratio chi-square statistics to a SAS data set. The following statements create a two-way table of eye color versus hair color.

```sas
proc freq data=Color order=data;
  weight Count;
  tables Eyes*Hair / chisq expected cellchi2 norow nocol;
  output out=ChiSqData pchi lrchi n nmiss;
  title 'Chi-Square Tests for 3 by 5 Table of Eye and Hair Color';
  run;
proc print data=ChiSqData noobs;
  title 'Chi-Square Statistics for Eye and Hair Color';
  title2 'Output Data Set from the FREQ Procedure';
  run;
```

The CHISQ option produces chi-square tests, the EXPECTED option displays expected cell frequencies in the table, and the CELLCHI2 option displays the cell contribution to the chi-square. The NOROW and NOCOL options suppress the display of row and column percents in the table.

The OUTPUT statement creates the ChiSqData data set with eight variables: the N option stores the number of nonmissing observations, the NMISS option stores the
number of missing observations, and the PCHI and LRCHI options store Pearson and likelihood-ratio chi-square statistics, respectively, together with their degrees of freedom and \( p \)-values.

The preceding statements produce Output 26.5.1 through Output 26.5.2.

**Output 26.5.1.** Contingency Table

<table>
<thead>
<tr>
<th>Eyes (Eye Color)</th>
<th>Hair (Hair Color)</th>
<th>Frequency</th>
<th>Expected</th>
<th>Cell Chi-Square</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td></td>
<td>69</td>
<td>28</td>
<td>68</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66.425</td>
<td>32.921</td>
<td>63.22</td>
<td>53.024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0998</td>
<td>0.7357</td>
<td>0.3613</td>
<td>0.0772</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.06</td>
<td>3.67</td>
<td>8.92</td>
<td>6.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.06</td>
<td>4.99</td>
<td>7.22</td>
<td>4.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.81</td>
<td>6.17</td>
<td>12.34</td>
<td>12.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.92</td>
<td>14.83</td>
<td>28.48</td>
<td>23.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.92</td>
<td>14.83</td>
<td>28.48</td>
<td>23.88</td>
</tr>
</tbody>
</table>

**Output 26.5.2.** Chi-Square Statistics

<table>
<thead>
<tr>
<th>Statistics for Table of Eyes by Hair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>Chi-Square</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
</tr>
<tr>
<td>Phi Coefficient</td>
</tr>
<tr>
<td>Contingency Coefficient</td>
</tr>
<tr>
<td>Cramer’s V</td>
</tr>
</tbody>
</table>

Sample Size = 762
Example 26.6. Computing Cochran-Mantel-Haenszel Statistics...

The contingency table in Output 26.5.1 displays eye and hair color in the order in which they appear in the Color data set. The Pearson chi-square statistic in Output 26.5.2 provides evidence of an association between eye and hair color ($p=0.0073$). The cell chi-square values show that most of the association is due to more green-eyed children with fair or red hair and fewer with dark or black hair. The opposite occurs with the brown-eyed children.

Output 26.5.3. Output Data Set

The OUT= data set is displayed in Output 26.5.3. It contains one observation with the sample size, the number of missing values, and the chi-square statistics and corresponding degrees of freedom and $p$-values as in Output 26.5.2.

Example 26.6. Computing Cochran-Mantel-Haenszel Statistics for a Stratified Table

The data set Migraine contains hypothetical data for a clinical trial of migraine treatment. Subjects of both genders receive either a new drug therapy or a placebo. Their response to treatment is coded as 'Better' or 'Same'. The data are recorded as cell counts, and the number of subjects for each treatment and response combination is recorded in the variable Count.

```sas
data Migraine;
  input Gender $ Treatment $ Response $ Count @@;
datalines;
  female Active Better 16 female Active Same 11
  female Placebo Better 5 female Placebo Same 20
  male Active Better 12 male Active Same 16
  male Placebo Better 7 male Placebo Same 19
;
```

The following statements create a three-way table stratified by Gender, where Treatment forms the rows and Response forms the columns. The CMH option produces the Cochran-Mantel-Haenszel statistics. For this stratified $2 \times 2$ table, estimates of the common relative risk and the Breslow-Day test for homogeneity of the odds ratios are also displayed. The NOPRINT option suppresses the display of the contingency tables. These statements produce Output 26.6.1 through Output 26.6.3.

```sas
proc freq data=Migraine;
  weight Count;
  tables Gender*Treatment*Response / cmh noprint;
  title1 'Clinical Trial for Treatment of Migraine Headaches';
run;
```

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>3</td>
<td>General Association</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

For a stratified $2 \times 2$ table, the three CMH statistics displayed in Output 26.6.1 test the same hypothesis. The significant $p$-value (0.004) indicates that the association between treatment and response remains strong after adjusting for gender.

Output 26.6.2.  CMH Option: Relative Risks

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Method</th>
<th>Value</th>
<th>95% Confidence Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>Mantel-Haenszel</td>
<td>3.3132</td>
<td>1.4456 7.5934</td>
</tr>
<tr>
<td>(Odds Ratio)</td>
<td>Logit</td>
<td>3.2941</td>
<td>1.4182 7.6515</td>
</tr>
<tr>
<td>Cohort</td>
<td>Mantel-Haenszel</td>
<td>2.1636</td>
<td>1.2336 3.7948</td>
</tr>
<tr>
<td>(Col1 Risk)</td>
<td>Logit</td>
<td>2.1059</td>
<td>1.1951 3.7108</td>
</tr>
<tr>
<td>Cohort</td>
<td>Mantel-Haenszel</td>
<td>0.6420</td>
<td>0.4705 0.8761</td>
</tr>
<tr>
<td>(Col2 Risk)</td>
<td>Logit</td>
<td>0.6613</td>
<td>0.4852 0.9013</td>
</tr>
</tbody>
</table>

The CMH option also produces a table of relative risks, as shown in Output 26.6.2. Because this is a prospective study, the relative risk estimate assesses the effectiveness of the new drug; the “Cohort (Col1 Risk)” values are the appropriate estimates for the first column, or the risk of improvement. The probability of migraine improvement with the new drug is just over two times the probability of improvement with the placebo.
Output 26.6.3. CMH Option: Breslow-Day Test

Clinical Trial for Treatment of Migraine Headaches

The FREQ Procedure

Summary Statistics for Treatment by Response
Controlling for Gender

Breslow-Day Test for
Homogeneity of the Odds Ratios

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1.4929</td>
</tr>
<tr>
<td>DF</td>
<td>1</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.2218</td>
</tr>
</tbody>
</table>

Total Sample Size = 106

The large \( p \)-value for the Breslow-Day test (0.2218) in Output 26.6.3 indicates no significant gender difference in the odds ratios.

Example 26.7. Computing the Cochran-Armitage Trend Test

The data set Pain contains hypothetical data for a clinical trial of a drug therapy to control pain. The clinical trial investigates whether adverse responses increase with larger drug doses. Subjects receive either a placebo or one of four drug doses. An adverse response is recorded as \texttt{Adverse}='Yes'; otherwise, it is recorded as \texttt{Adverse}='No'. The number of subjects for each drug dose and response combination is contained in the variable \texttt{Count}.

```sas
data Pain;
  input Dose Adverse Count @@;
datalines;
0 No  26 0 Yes  6
1 No  26 1 Yes  7
2 No  23 2 Yes  9
3 No  18 3 Yes 14
4 No  9  4 Yes 23
;
```

The TABLES statement in the following program produces a two-way table. The MEASURES option produces measures of association, and the CL option produces confidence bounds for these measures. The TREND option tests for a trend across the ordinal values of the \texttt{Dose} variable with the Cochran-Armitage test. The EXACT statement produces exact \( p \)-values for this test, and the MAXTIME= option terminates the exact computations if they do not complete within 60 seconds. The TEST statement computes an asymptotic test for Somer’s \( D(C|R) \). These statements produce Output 26.7.1 through Output 26.7.3.
Chapter 26. The FREQ Procedure

```
proc freq data=Pain;
  weight Count;
  tables Dose*Adverse / trend measures cl;
  test smdcr;
  exact trend / maxtime=60;
  title1 'Clinical Trial for Treatment of Pain';
run;
```

Output 26.7.1. Contingency Table

<table>
<thead>
<tr>
<th>Dose</th>
<th>Adverse</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>0</td>
<td>26</td>
<td>6</td>
<td>32</td>
<td>16.15</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td>16.15</td>
<td>3.73</td>
<td>19.88</td>
<td>81.25</td>
<td>18.75</td>
</tr>
<tr>
<td></td>
<td>25.49</td>
<td>10.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>7</td>
<td>33</td>
<td>16.15</td>
<td>4.35</td>
</tr>
<tr>
<td></td>
<td>16.15</td>
<td>4.35</td>
<td>20.50</td>
<td>78.79</td>
<td>21.21</td>
</tr>
<tr>
<td></td>
<td>25.49</td>
<td>11.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>9</td>
<td>32</td>
<td>14.29</td>
<td>5.59</td>
</tr>
<tr>
<td></td>
<td>14.29</td>
<td>5.59</td>
<td>19.88</td>
<td>71.88</td>
<td>28.13</td>
</tr>
<tr>
<td></td>
<td>22.55</td>
<td>15.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>14</td>
<td>32</td>
<td>11.18</td>
<td>8.70</td>
</tr>
<tr>
<td></td>
<td>11.18</td>
<td>8.70</td>
<td>19.88</td>
<td>56.25</td>
<td>43.75</td>
</tr>
<tr>
<td></td>
<td>17.65</td>
<td>23.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>23</td>
<td>32</td>
<td>5.59</td>
<td>14.29</td>
</tr>
<tr>
<td></td>
<td>5.59</td>
<td>14.29</td>
<td>19.88</td>
<td>28.13</td>
<td>71.88</td>
</tr>
<tr>
<td></td>
<td>8.82</td>
<td>38.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>102</td>
<td>59</td>
<td>161</td>
<td>63.35</td>
<td>36.65</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The “Row Pct” values in Output 26.7.1 show the expected increasing trend in the proportion of adverse effects due to increasing dosage (from 18.75% to 71.88%).
Output 26.7.2. Measures of Association

Clinical Trial for Treatment of Pain

The FREQ Procedure

Statistics for Table of Dose by Adverse

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
<th>95% Confidence Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.5313</td>
<td>0.0935</td>
<td>0.3480 0.7146</td>
</tr>
<tr>
<td>Kendall’s Tau-b</td>
<td>0.3373</td>
<td>0.0642</td>
<td>0.2114 0.4631</td>
</tr>
<tr>
<td>Stuart’s Tau-c</td>
<td>0.4111</td>
<td>0.0798</td>
<td>0.2547 0.5675</td>
</tr>
<tr>
<td>Somers’ D C</td>
<td>R</td>
<td>0.2569</td>
<td>0.0499</td>
</tr>
<tr>
<td>Somers’ D R</td>
<td>C</td>
<td>0.4427</td>
<td>0.0837</td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>0.3776</td>
<td>0.0714</td>
<td>0.2378 0.5175</td>
</tr>
<tr>
<td>Spearman Correlation</td>
<td>0.3771</td>
<td>0.0718</td>
<td>0.2363 0.5178</td>
</tr>
<tr>
<td>Lambda Asymmetric C</td>
<td>R</td>
<td>0.2373</td>
<td>0.0837</td>
</tr>
<tr>
<td>Lambda Asymmetric R</td>
<td>C</td>
<td>0.1250</td>
<td>0.0662</td>
</tr>
<tr>
<td>Lambda Symmetric</td>
<td>0.1604</td>
<td>0.0621</td>
<td>0.0388 0.2821</td>
</tr>
<tr>
<td>Uncertainty Coefficient C</td>
<td>R</td>
<td>0.1261</td>
<td>0.0467</td>
</tr>
<tr>
<td>Uncertainty Coefficient R</td>
<td>C</td>
<td>0.0515</td>
<td>0.0191</td>
</tr>
<tr>
<td>Uncertainty Coefficient Symmetric</td>
<td>0.0731</td>
<td>0.0271</td>
<td>0.0199 0.1262</td>
</tr>
</tbody>
</table>

Somers’ D C|R

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>ASE</th>
<th>95% Confidence Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somers’ D C</td>
<td>R</td>
<td>0.2569</td>
<td>0.0499</td>
</tr>
<tr>
<td>ASE under H0</td>
<td>0.0499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>5.1511</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-sided Pr &gt; Z</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-sided Pr &gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample Size = 161

Output 26.7.2 displays the measures of association produced by the MEASURES option. Somer’s $D(C|R)$ measures the association treating the column variable (Adverse) as the response and the row variable (Dose) as a predictor. Because the asymptotic 95% confidence bounds do not contain zero, this indicates a strong positive association. Similarly, the Pearson and Spearman correlation coefficients show evidence of a strong positive association, as hypothesized.
The Cochran-Armitage test (Output 26.7.3) supports the trend hypothesis. The small left-sided \( p \)-values for the Cochran-Armitage test indicate that the probability of the Column 1 level (\text{Adverse}='\text{No}') decreases as \text{Dose} increases or, equivalently, that the probability of the Column 2 level (\text{Adverse}='\text{Yes}') increases as \text{Dose} increases. The two-sided \( p \)-value tests against either an increasing or decreasing alternative. This is an appropriate hypothesis when you want to determine whether the drug has progressive effects on the probability of adverse effects but the direction is unknown.

Example 26.8. Computing Friedman’s Chi-Square Statistic

Friedman’s test is a nonparametric test for treatment differences in a randomized complete block design. Each block of the design may be a subject or a homogeneous group of subjects. If blocks are groups of subjects, the number of subjects in each block must equal the number of treatments. Treatments are randomly assigned to subjects within each block. If there is one subject per block, then the subjects are repeatedly measured once under each treatment. The order of treatments is randomized for each subject.

In this setting, Friedman’s test is identical to the ANOVA (row means scores) CMH statistic when the analysis uses rank scores (SCORES=RANK). The three-way table uses subject (or subject group) as the stratifying variable, treatment as the row variable, and response as the column variable. PROC FREQ handles ties by assigning midranks to tied response values. If there are multiple subjects per treatment in each block, the ANOVA CMH statistic is a generalization of Friedman’s test.

The data set \text{Hypnosis} contains data from a study investigating whether hypnosis has the same effect on skin potential (measured in millivolts) for four emotions (Lehmann 1975, p. 264). Eight subjects are asked to display fear, joy, sadness, and calmness under hypnosis. The data are recorded as one observation per subject for each emotion.
Example 26.8. Computing Friedman’s Chi-Square Statistic

```sas
data Hypnosis;
  length Emotion $ 10;
  input Subject Emotion $ SkinResponse @@;
  datalines;
1 fear 23.1 1 joy 22.7 1 sadness 22.5 1 calmness 22.6
2 fear 57.6 2 joy 53.2 2 sadness 53.7 2 calmness 53.1
3 fear 10.5 3 joy 9.7 3 sadness 10.8 3 calmness 8.3
4 fear 23.6 4 joy 19.6 4 sadness 21.1 4 calmness 21.6
5 fear 11.9 5 joy 13.8 5 sadness 13.7 5 calmness 13.3
6 fear 54.6 6 joy 47.1 6 sadness 39.2 6 calmness 37.0
7 fear 21.0 7 joy 13.6 7 sadness 13.7 7 calmness 14.8
8 fear 20.3 8 joy 23.6 8 sadness 16.3 8 calmness 14.8
;
```

In the following statements, the TABLES statement creates a three-way table stratified by Subject and a two-way table; the variables Emotion and SkinResponse form the rows and columns of each table. The CMH2 option produces the first two Cochran-Mantel-Haenszel statistics, the option SCORES=RANK specifies that rank scores are used to compute these statistics, and the NOPRINT option suppresses the contingency tables. These statements produce Output 26.8.1 and Output 26.8.2.

```sas
proc freq data=Hypnosis;
  tables Subject*Emotion*SkinResponse Emotion*SkinResponse
    / cmh2 scores=rank noprint;
run;
```

**Output 26.8.1. CMH Statistics: Stratifying by Subject**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>0.2400</td>
<td>0.6242</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>3</td>
<td>6.4500</td>
<td>0.0917</td>
</tr>
</tbody>
</table>

Total Sample Size = 32

Because the CMH statistics in Output 26.8.1 are based on rank scores, the Row Mean Scores Differ statistic is identical to Friedman’s chi-square ($Q = 6.45$). The $p$-value of 0.0917 indicates that differences in skin potential response for different emotions are significant at the 10% level but not at the 5% level.
Output 26.8.2. CMH Statistics: No Stratification

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Nonzero Correlation</td>
<td>1</td>
<td>1</td>
<td>0.0001</td>
<td>0.9933</td>
</tr>
<tr>
<td>2 Row Mean Scores Differ</td>
<td>3</td>
<td>3</td>
<td>0.5678</td>
<td>0.9038</td>
</tr>
</tbody>
</table>

Total Sample Size = 32

When you do not stratify by subject, the Row Mean Scores Differ CMH statistic is identical to a Kruskal-Wallis test and is not significant (p=0.9038 in Output 26.8.2). Thus, adjusting for subject is critical to reducing the background variation due to subject differences.

Example 26.9. Testing Marginal Homogeneity with Cochran's Q

When a binary response is measured several times or under different conditions, Cochran’s Q tests that the marginal probability of a positive response is unchanged across the times or conditions. When there are more than two response categories, you can use the CATMOD procedure to fit a repeated-measures model.

The data set Drugs contains data for a study of three drugs to treat a chronic disease (Agresti 1990). Forty-six subjects receive drugs A, B, and C. The response to each drug is either favorable ('F') or unfavorable ('U').

```sas
proc format;
   value $ResponseFmt 'F'='Favorable'
       'U'='Unfavorable';

data drugs;
   input Drug_A $ Drug_B $ Drug_C $ Count @@;
   datalines;
   F F F 6  U F F 2  
   F F U 16  U F U 4  
   F U F 2  U U F 6  
   F U U 4  U U U 6  
;
```

The following statements create one-way frequency tables of the responses to each drug. The AGREE option produces Cochran’s Q and other measures of agreement for the three-way table. These statements produce Output 26.9.1 through Output 26.9.3.
Example 26.9. Testing Marginal Homogeneity with Cochran’s Q

```sas
proc freq data=Drugs;
  weight Count;
  tables Drug_A Drug_B Drug_C / nocum;
  tables Drug_A*Drug_B*Drug_C / agree noprint;
  format Drug_A Drug_B Drug_C $ResponseFmt.;
  title 'Study of Three Drug Treatments for a Chronic Disease';
run;
```

Output 26.9.1. One-Way Frequency Tables

<table>
<thead>
<tr>
<th>Drug_A</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>28</td>
<td>60.87</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>18</td>
<td>39.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug_B</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>28</td>
<td>60.87</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>18</td>
<td>39.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug_C</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>16</td>
<td>34.78</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>30</td>
<td>65.22</td>
</tr>
</tbody>
</table>

The one-way frequency tables in Output 26.9.1 provide the marginal response for each drug. For drugs A and B, 61% of the subjects reported a favorable response while 35% of the subjects reported a favorable response to drug C.
Output 26.9.2. Measures of Agreement

Study of Three Drug Treatments for a Chronic Disease

The FREQ Procedure

Statistics for Table 1 of Drug_B by Drug_C
Controlling for Drug_A=Favorable

McNemar’s Test
----------------
Statistic (S) 10.8889
DF 1
Pr > S 0.0010

Simple Kappa Coefficient
------------------------
Kappa -0.0328
ASE 0.1167
95% Lower Conf Bound -0.2615
95% Upper Conf Bound 0.1960

Sample Size = 28

Statistics for Table 2 of Drug_B by Drug_C
Controlling for Drug_A=Unfavorable

McNemar’s Test
---------------
Statistic (S) 0.4000
DF 1
Pr > S 0.5271

Simple Kappa Coefficient
------------------------
Kappa -0.1538
ASE 0.2230
95% Lower Conf Bound -0.5909
95% Upper Conf Bound 0.2932

Sample Size = 18
McNemar’s test (Output 26.9.2) shows strong discordance between drugs B and C when the response to drug A is favorable. The small negative value of the simple kappa indicates no agreement between drug B response and drug C response.

Output 26.9.3. Cochran’s $Q$

Cochran’s $Q$ is statistically significant ($p=0.0144$ in Output 26.9.3), which leads to rejection of the hypothesis that the probability of favorable response is the same for the three drugs.

References


