CHAPTER 21

The FREQ Procedure

Overview

The FREQ procedure is a descriptive as well as a statistical procedure that produces one-way to *n*-way frequency and crosstabulation tables. *Frequency tables* concisely describe your data by reporting the distribution of variable values.*Crosstabulation tables*, also known as *contingency tables*, summarize data for two or more classification variables by showing the number of observations for each combination of variable values.

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 the existence of an association, PROC FREQ computes statistics that test the null hypothesis of no association. To determine 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 their maximums (or minimums) when there is perfect association. The statistics for contingency tables include

- chi-square tests and measures
- measures of association and tests of these measures
- risks (or binomial proportions) and risk differences for 2×2 tables
- odds ratios and relative risks for 2×2 tables
- tests for trend
- tests and measures of agreement
- Cochran-Mantel-Haenszel statistics.

PROC FREQ computes asymptotic standard errors, confidence limits, and tests for measures of association and measures of agreement. Exact *p*-values and confidence limits 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.

Output 21.1 on page 501 is the simplest form of PROC FREQ output. The one-way frequency tables of hair and eye color show the distributions of these variables. PROC FREQ lists each variable value along with the frequencies and percentages. The statements that produce the output follow:

```
proc freq data=color;
run;
```

Output 21.1 One-Way Frequency Tables Produced with PROC FREQ

```
                          The SAS System                        1

                        The FREQ Procedure

                            Eye Color

                                          Cumulative    Cumulative
   Eyes     Frequency     Percent     Frequency      Percent
   ----------------------------------------------------------
   blue          222       29.13           222         29.13
   brown         341       44.75           563         73.88
   green         199       26.12           762        100.00


                            Hair Color

                                          Cumulative    Cumulative
  Hair      Frequency     Percent     Frequency      Percent
  ------------------------------------------------------------
  black          22        2.89            22          2.89
  dark          182       23.88           204         26.77
  fair          228       29.92           432         56.69
  medium        217       28.48           649         85.17
  red           113       14.83           762        100.00
```

In addition to listing the frequency distribution separately for each variable, you can create a crosstabulation table to show the joint frequency distribution for the two variables. Output 21.2 on page 502 shows a two-way crosstabulation table and chi-square statistics that test the association between eye and hair color of children from two regions of Europe. The statements that produce this 3×5 table also

- □ order the variable values according to their appearance in the data set
- □ exclude the row and column percentages for each cell
- □ include the expected frequency for each cell
- □ include each cell's contribution to the total Pearson chi-square statistic.

In addition to displaying the statistics, the program creates an output data set that contains selected chi-square statistics. For an explanation of the program that produces this output, see Example 5 on page 584.

Output 21.2 Chi-Square Statistics Produced with PROC FREQ

```
              Chi-Square Tests for 3 by 5 Table of Eye and Hair Color            1

                                   The FREQ Procedure

                                 Table of Eyes by Hair

       Eyes(Eye Color)     Hair(Hair Color)

       Frequency      |
       Expected       |
       Cell Chi-Square|
       Percent        |fair    |red     |medium  |dark    |black   |  Total
       ---------------+--------+--------+--------+--------+--------+
       blue           |     69 |     28 |     68 |     51 |      6 |    222
                      | 66.425 | 32.921 |  63.22 | 53.024 | 6.4094 |
                      | 0.0998 | 0.7357 | 0.3613 | 0.0772 | 0.0262 |
                      |   9.06 |   3.67 |   8.92 |   6.69 |   0.79 |  29.13
       ---------------+--------+--------+--------+--------+--------+
       green          |     69 |     38 |     55 |     37 |      0 |    199
                      | 59.543 |  29.51 | 56.671 |  47.53 | 5.7454 |
                      | 1.5019 | 2.4422 | 0.0492 | 2.3329 | 5.7454 |
                      |   9.06 |   4.99 |   7.22 |   4.86 |   0.00 |  26.12
       ---------------+--------+--------+--------+--------+--------+
       brown          |     90 |     47 |     94 |     94 |     16 |    341
                      | 102.03 | 50.568 | 97.109 | 81.446 | 9.8451 |
                      | 1.4187 | 0.2518 | 0.0995 |  1.935 | 3.8478 |
                      |  11.81 |   6.17 |  12.34 |  12.34 |   2.10 |  44.75
       ---------------+--------+--------+--------+--------+--------+
       Total               228      113      217      182       22      762
                         29.92    14.83    28.48    23.88     2.89   100.00


                            Statistics for Table of Eyes by Hair

              Statistic                     DF       Value      Prob
              ------------------------------------------------------
              Chi-Square                     8     20.9248    0.0073
              Likelihood Ratio Chi-Square    8     25.9733    0.0011
              Mantel-Haenszel Chi-Square     1      3.7838    0.0518
              Phi Coefficient                       0.1657
              Contingency Coefficient               0.1635
              Cramer's V                            0.1172

                                   Sample Size = 762
```

```
                    Chi-Square Statistics for Eye and Hair Color                2
                      Output Data Set from the FREQ Procedure

  N    NMISS    _PCHI_    DF_PCHI       P_PCHI    _LRCHI_   DF_LRCHI      P_LRCHI

 762     0     20.9248       8      .007349898   25.9733       8      .001061424
```

Several SAS procedures produce frequency counts; only PROC FREQ computes chi-square tests, measures of association, and measures of agreement for contingency tables. Other procedures to consider for counting are PROC TABULATE for more general table layouts; PROC REPORT for tables and customized summaries, PROC CHART for bar charts and other graphical representations; and PROC UNIVARIATE with the FREQ option for one-way frequency tables. When you want to fit models to categorical data, use a SAS/STAT procedure such as CATMOD, GENMOD, LOGISTIC, PHREG, or PROBIT. For more information on selecting the appropriate statistical analyses, refer to *An Introduction to Categorical Data Analysis* (Agresti, 1996) or *Categorical Data Analysis Using the SAS System* (Stokes, et al. 1995).

Procedure Syntax

Tip: Supports the Output Delivery System (see Chapter 2, “Fundamental Concepts for Using Base SAS Procedures”)

Reminder: You can use the FORMAT, LABEL, and WHERE statements. See Chapter 3, "Statements with the Same Function in Multiple Procedures," for details. You can also use any global statements as well. See Chapter 2, "Fundamental Concepts for Using Base SAS Procedures," for a list.

PROC FREQ <*option(s)*>;
  BY <DESCENDING> *variable-1* <*...*<DESCENDING> *variable-n*> <NOTSORTED>;
  EXACT *statistic-keyword(s)* <*/ option(s)*>;
  OUTPUT *statistic-keyword(s)* <OUT=*SAS-data-set*>;
  TABLES *request(s)* <*/ option(s)*>;
  TEST *statistic-keyword(s)*;
  WEIGHT *variable*;

To do this	Use this statement
Calculate separate frequency or crosstabulation tables for each BY group	BY
Request exact tests for specified statistics	EXACT
Create an output data set that contains specified statistics	OUTPUT
Specify frequency or crosstabulation tables and request tests and measures of association	TABLES
Request asymptotic tests for measures of association and agreement	TEST
Identify a variable whose values weight each observation	WEIGHT

PROC FREQ Statement

PROC FREQ <*option(s)*>;

To do this	Use this option
Specify the input data set	DATA=
Control printed output	
Begin the next one-way table on the current page even if the entire table does not fit on that page	COMPRESS

To do this	Use this option
Specify the outline and cell divider characters for the cells of the crosstabulation tables	FORMCHAR=
Suppress all displayed output	NOPRINT
Specify the order to list the variable values	ORDER=
Display one table per page	PAGE

Options

COMPRESS

begins to display the next one-way frequency table on the same page as the preceding one-way table when 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.

Restriction: not valid with PAGE

Tip: COMPRESS saves paper and screen space.

DATA=*SAS-data-set*

specifies the input SAS data set.

Main discussion: "Procedure Concepts" on page 18

FORMCHAR <(*position(s)*)>='*formatting-character(s)*'

defines the characters to use for constructing the outlines and dividers for the cells of crosstabulation tables.

position(s)

identifies the position of one or more characters in the SAS formatting-character string. A space or a comma separates the positions.

Default: Omitting (*position(s)*), is the same as specifying all 20 possible SAS formatting characters, in order.

Range: PROC FREQ uses formatting characters 1, 2, and 7. Table 21.1 on page 506 shows the formatting characters that PROC FREQ uses.

formatting-character(s)

lists the characters to use for the specified positions. PROC FREQ assigns characters in *formatting-character(s)* to *position(s)*, in the order that they are listed. For instance, the following option assigns the asterisk (*) to the second formatting character, the pound sign (#) to the seventh character, and does not alter the remaining characters:

```
formchar(2,7)='*#'
```

Interaction: The SAS system option FORMCHAR= specifies the default formatting characters. The system option defines the entire string of formatting characters. Specifying the FORMCHAR= option in a procedure can redefine selected characters.

Tip: You can use any character in *formatting-characters*, including hexadecimal characters. If you use hexadecimal characters, you must put an **x** after the closing quote. For example the following option assigns the hexadecimal character 2D to the second formatting character, the hexadecimal character 7C to the seventh character, and does not alter the remaining characters:

```
formchar(2,7)='2D7C'x
```

Tip: Specifying all blanks for *formatting-character(s)* produces tables with no outlines or dividers:

```
formchar (1,2,7)=''
```

(3 blanks)

See also: For information on which hexadecimal codes to use for which characters, consult the documentation for your hardware.

Table 21.1 Formatting Characters Used by PROC FREQ

Position	**Default**	**Used to draw**
1	|	Vertical separators
2	-	Horizontal separators
7	+	Intersections of vertical and horizontal separators

NOPRINT

suppresses all displayed output from PROC FREQ.

Interaction: NOPRINT in the PROC statement disables the Output Delivery System for the entire PROC step.

Tip: Use NOPRINT when you want to create only an output data set with the OUTPUT statement or with the OUT= option in the TABLES statement.

Note: NOPRINT is also available in the TABLES statement where it suppresses the tables, but displays the requested statistics. △

ORDER=DATA | FORMATTED | FREQ | INTERNAL

orders the values of the frequency and crosstabulation table variables according to the specified order, where

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 as PROC SORT. This order is operating environment-dependent.

Default: INTERNAL

Restriction: ORDER= does not apply to missing values, which always appear first.

Featured in: Example 2 on page 575 and Example 3 on page 578

PAGE

displays only one table per page.

Default: displays multiple tables per page as space permits

Restriction: not valid with COMPRESS

BY Statement

Calculates separate analysis for each BY group.

Main discussion: "Statements" on page 68

Featured in: Example 2 on page 575

BY <DESCENDING> *variable-1* <...<DESCENDING> *variable-n*> <NOTSORTED>;

Required Arguments

variable
specifies the variable that the procedure uses to form BY groups. You can specify more than one variable. If you do not use the NOTSORTED option in the BY statement, the observations in the data set must either be sorted by all the variables that you specify, or they must be indexed appropriately.

Options

DESCENDING
specifies that the observations are sorted in descending order by the variable that immediately follows the word DESCENDING in the BY statement.

NOTSORTED
specifies that observations are not necessarily sorted in alphabetic or numeric order. The observations are grouped in another way, for example, chronological order.

The requirement for ordering or indexing observations according to the values of BY variables is suspended for BY-group processing when you use the NOTSORTED option. In fact, the procedure does not use an index if you specify NOTSORTED. The procedure defines a BY group as a set of contiguous observations that have the same values for all BY variables. If observations with the same values for the BY variables are not contiguous, the procedure treats each contiguous set as a separate BY group.

EXACT Statement

Requests exact tests or confidence limits for the specified statistics. Optionally requests Monte Carlo estimates of the exact *p*-values.

Requirements: TABLES statement

Main discussion: "Exact Statistics" on page 563

Featured in: Example 4 on page 580

EXACT *statistic-keyword(s)* <*/ option(s)*>;

Required Arguments

statistic-keyword(s)

specifies the statistics for which to provide exact tests or confidence limits. PROC FREQ can compute exact *p*-values for the following hypothesis tests: chi-square goodness-of-fit 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 hypotheses that the following statistics are equal to zero: Pearson correlation coefficient, Spearman correlation coefficient, simple kappa coefficient, and weighted kappa coefficient. PROC FREQ can compute exact *p*-values for the binomial proportion test, as well as exact confidence limits for the binomial proportion. Additionally, PROC FREQ can compute exact confidence limits for odds ratios for 2×2 tables.

The statistic keywords are identical to options in the TABLES statement and keywords in the OUTPUT statement. You can request exact computations for groups of statistics by using keywords that are identical to the following TABLES statement options: CHISQ, MEASURES, and AGREE. For example, when you specify CHISQ in the EXACT statement, PROC FREQ computes exact *p*-values for the available CHISQ statistics (Pearson chi-square, likelihood-ratio chi-square, and Mantel-Haenszel chi-square). You request exact *p*-values for an individual statistic by specifying a keyword shown in Table 21.2 on page 508.

Note: PROC FREQ computes exact tests by using fast and efficient algorithms that are superior to direct enumeration. This technique is appropriate when a data set is small, sparse, skewed, or heavily tied. For some large problems, exact computations may require a large amount of time or memory. Consider using the asymptotic tests for such problems. Alternatively, when asymptotic methods may not be sufficient for such large problems, consider using Monte Carlo estimation of exact *p*-values. See "Exact Statistics" on page 563 for more information. △

Table 21.2 EXACT Statement Statistic-keywords and Required TABLES Statement Options

Keyword	Exact statistics computed	Required TABLES statement option
AGREE	McNemar's test for 2×2 tables and tests for the simple kappa coefficient and the weighted kappa coefficient	AGREE
BINOMIAL	binomial proportion test for one-way tables	BINOMIAL
CHISQ	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	ALL, CHISQ
FISHER	Fisher's exact test	ALL*, CHISQ*
JT	Jonckheere-Terpstra test	JT
KAPPA	test for the simple kappa coefficient	AGREE
LRCHI	likelihood-ratio chi-square test	ALL, CHISQ
MCNEM	McNemar's test for 2×2 tables	AGREE

Keyword	Exact statistics computed	Required TABLES statement option
MEASURES	tests for the Pearson correlation coefficient and the Spearman correlation and the odds ratio confidence limits for 2×2 tables	ALL, MEASURES
MHCHI	Mantel-Haenszel chi-square test	ALL, CHISQ
OR	odds ratio confidence limits for 2×2 tables	ALL, MEASURES, RELRISK
PCHI	chi-square goodness-of-fit test for one-way tables, Pearson chi-square test for 2×2 tables	ALL, CHISQ
PCORR	test for the Pearson correlation coefficient	ALL, MEASURES
SCORR	test for the Spearman correlation coefficient	ALL, MEASURES
TREND	Cochran-Armitage test for trend	TREND
WTKAP	test for the weighted kappa coefficient	AGREE

* ALL and CHISQ compute Fisher's exact test only for 2×2 tables.

Options

ALPHA=*p*

specifies the confidence level for the confidence limits for the Monte Carlo *p*-value estimates. A confidence level of *p* results in (1–*p*)×100 percent confidence limits. Using ALPHA=.01 results in 99 percent confidence limits. If *p* is between 0 and 1 but is outside the range, PROC FREQ uses the closest range endpoint. For example, if *p*= 0.000001, PROC FREQ uses 0.0001 to determine confidence limits.

Default: 0.01

Range: 0.000<=*p*<=0.0001

Interaction: ALPHA= invokes the MC option.

MAXTIME=*value*

specifies the maximum clock time (in seconds) that PROC FREQ uses to compute an exact *p*-value directly or with Monte Carlo estimation. If the procedure does not complete the computation within the specified time, the computation terminates.

Range: a positive number

See also: "Computational Resources" on page 565

Featured in: Example 7 on page 590

MC

requests Monte Carlo estimation of exact *p*-values, instead of direct exact *p*-value computation. Monte Carlo estimation can be useful for large problems that require a large amount of time and memory for exact computations, but for which asymptotic approximations may not be sufficient.

Restriction: The MC option is available for all statistic keywords except BINOMIAL, MCNEM, and OR. PROC FREQ computes only exact tests or confidence limits for those statistics.

Tip: If the procedure does not complete the computation within the specified time, use MAXTIME= to increase the amount of clock time that PROC FREQ uses to compute the exact *p*-values.

Interaction: ALPHA=, N=, and SEED= automatically invoke the MC option.

Tip: If the procedure does not complete the computation within the specified time, use MAXTIME= to increase the amount of clock time PROC FREQ can use to compute the Monte Carlo estimates.

Main Discussion: "Monte Carlo Estimation" on page 566

N=*n*

specifies the number of samples for Monte Carlo estimation.

Default: 10000

Range: a positive integer

Interaction: N= invokes the MC option.

Tip: Larger values of N= produce more precise estimates of exact *p*-values. Because larger values of N= generate more samples, the computation time increases. If you need more computation time, use MAXTIME= to increase the clock time.

SEED=*n*

specifies the initial seed for random number generation for Monte Carlo estimation.

Default: the time of day from the computer's clock

Range: a positive integer

Interaction: SEED= invokes the MC option.

Using TABLES Statement Options with the EXACT Statement

Table 21.2 on page 508 lists the available statistic keywords and the exact statistics that are computed. If you use only one TABLES statement, you do not need to specify options in the TABLES statement to compute the statistics that the EXACT statement requests. PROC FREQ automatically invokes the corresponding TABLES statement option when you request exact computations. However, when you use multiple TABLES statements, and you want exact computations, you must specify options in the TABLES statement to compute the desired statistics. Then PROC FREQ performs exact computations for all statistics that are also specified in the EXACT statement.

OUTPUT Statement

Creates a SAS data set with the statistics that PROC FREQ computes for the last TABLES statement request. The variables contain statistics for each two-way table or stratum, as well as summary statistics across all strata.

Requirements: TABLES statement

Restriction: Only one OUTPUT statement is allowed.

Tip: Use the Output Delivery System to create a SAS data set from any piece of PROC FREQ output.

Main discussion: "Output Data Sets" on page 570

Featured in: Example 5 on page 584

OUTPUT *statistic-keyword(s)* <OUT=*SAS-data-set*>;

Options

OUT=*SAS-data-set*

names the output data set that contains statistics for the last TABLES statement request. If you omit OUT=, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

Default: DATA*n*

statistic-keyword(s)

specifies the statistics that you want in the new data set. Available statistics are those produced by PROC FREQ for each one-way or two-way table, as well as summary statistics across all strata. When you request a statistic, the OUTPUT data set contains that estimate or test statistic, as well as any associated standard error, degrees of freedom, confidence limits, and *p*–values.

You can save statistics by using keywords that are identical to group options in the TABLES statement: AGREE, ALL, CHISQ, CMH, and MEASURES. Alternatively, you can request an individual statistic by specifying a keyword shown in Table 21.3 on page 511.

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 with a 2×2 table, you cannot specify the keyword OR (odds ratio) in the OUTPUT statement without also specifying ALL, MEASURES, or RELRISK in the TABLES 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.

Table 21.3 OUTPUT Statement Statistic-keywords and Required TABLES Statement Options

Keyword	Output data set statistics	Required TABLES statement option
AGREE	McNemar's test for 2×2 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.	AGREE
AJCHI	continuity-adjusted chi-square for 2×2 tables	ALL, CHISQ
ALL	all statistics under CHISQ, MEASURES, CMH, and the number of nonmissing subjects	ALL
BDCHI	Breslow-Day test	ALL, CMH, CMH1, CMH2
BINOMIAL	binomial proportion statistics for one-way tables	BINOMIAL
CHISQ	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 2×2 tables, Mantel-Haenszel chi-square, Fisher's exact test for 2×2 tables, phi coefficient, contingency coefficient, and Cramer's *V*	ALL, CHISQ

Keyword	Output data set statistics	Required TABLES statement option
CMH	Cochran-Mantel-Haenszel correlation, row mean scores (*ANOVA*), and general association statistics; for 2×2 tables, logit and Mantel-Haenszel adjusted odds ratios, relative risks, and Breslow-Day test	ALL, CMH
CMH1	same as CMH, but excludes general association and row mean scores (*ANOVA*) statistics	ALL, CMH, CMH1, CMH2
CMH2	same as CMH, but excludes the general association statistic	ALL, CMH, CMH2
CMHCOR	Cochran-Mantel-Haenszel correlation statistic	ALL, CMH, CMH1, CMH2
CMHGA	Cochran-Mantel-Haenszel general association statistic	ALL, CMH
CMHRMS	Cochran-Mantel-Haenszel row mean scores (*ANOVA*) statistic	ALL, CMH, CMH2
COCHQ	Cochran's *Q*	AGREE
CONTGY	contingency coefficient	ALL, CHISQ
CRAMV	Cramer's *V*	ALL, CHISQ
EQKAP	test for equal simple kappas	AGREE
EQWKP	test for equal weighted kappas	AGREE
FISHER | EXACT	Fisher's exact test	ALL*, CHISQ*, FISHER, EXACT
GAMMA	gamma	ALL, MEASURES
JT	Jonckheere-Terpstra test	JT
KAPPA	simple kappa coefficient	AGREE
KENTB	Kendall's tau-*b*	ALL, MEASURES
LAMCR	lambda asymmetric (*C* | *R*)	ALL, MEASURES
LAMDAS	lambda symmetric	ALL, MEASURES
LAMRC	lambda asymmetric (*R* | *C*)	ALL, MEASURES
LGOR	adjusted logit odds ratio	ALL, CMH, CMH1, CMH2
LGRRC1	adjusted logit column 1 relative risk	ALL, CMH, CMH1, CMH2
LGRRC2	adjusted logit column 2 relative risk	ALL, CMH, CMH1, CMH2
LRCHI	likelihood ratio chi-square	ALL, CHISQ
MCNEM	McNemar's test	AGREE
MEASURES	gamma, Kendall's tau-*b*, Stuart's tau-*c*, Somers' *D* (*C* | *R*), Somers' *D* (*R* | *C*), Pearson correlation coefficient, Spearman correlation coefficient, lambda asymmetric (*C* | *R*), lambda asymmetric (*R* | *C*), lambda symmetric, uncertainty coefficient (*C* | *R*), uncertainty coefficient (*R* | *C*), and symmetric uncertainty coefficient; for 2×2 tables, odds ratio and relative risks	ALL, MEASURES
MHCHI	Mantel-Haenszel chi-square	ALL, CHISQ
MHOR	adjusted Mantel-Haenszel odds ratio	ALL, CMH, CMH1, CMH2
MHRRC1	adjusted Mantel-Haenszel column 1 relative risk	ALL, CMH, CMH1, CMH2
MHRRC2	adjusted Mantel-Haenszel column 2 relative risk	ALL, CMH, CMH1, CMH2

Keyword	**Output data set statistics**	**Required TABLES statement option**
N	number of nonmissing subjects for the stratum	
NMISS	number of missing subjects for the stratum	
OR	odds ratio	ALL, MEASURES, RELRISK
PCHI	chi-square goodness-of-fit test for one-way tables; for 2-way tables, Pearson chi-square	ALL, CHISQ
PCORR	Pearson correlation coefficient	ALL, MEASURES
PHI	phi coefficient	ALL, CHISQ
PLCORR	polychoric correlation coefficient	PLCORR
RDIF1	column 1 risk difference (row 1 – row 2)	RISKDIFF
RDIF2	column 2 risk difference (row 1 – row 2)	RISKDIFF
RELRISK	odds ratio and relative risks for 2×2 tables	ALL, MEASURES, RELRISK
RISKDIFF	risks and risk differences	RISKDIFF
RISKDIFF1	column 1 risks and risk difference	RISKDIFF
RISKDIFF2	column 2 risks and risk difference	RISKDIFF
RRC1	column 1 relative risk	ALL, MEASURES, RELRISK
RRC2	column 2 relative risk	ALL, MEASURES, RELRISK
RSK1	column 1 risk (overall)	RISKDIFF
RSK11	column 1 risk, for row 1	RISKDIFF
RSK12	column 2 risk, for row 1	RISKDIFF
RSK2	column 2 risk (overall)	RISKDIFF
RSK21	column 1 risk, for row 2	RISKDIFF
RSK22	column 2 risk, for row 2	RISKDIFF
SCORR	Spearman correlation coefficient	ALL, MEASURES
SMDCR	Somers' D $(C\,\vert\,R)$	ALL, MEASURES
SMDRC	Somers' D $(R\,\vert\,C)$	ALL, MEASURES
STUTC	Stuart's tau-*c*	ALL, MEASURES
TREND	Cochran-Armitage test for trend	TREND
TSYMM	Bowker's test of symmetry	AGREE
U	symmetric uncertainty coefficient	ALL, MEASURES
UCR	uncertainty coefficient $(C\,\vert\,R)$	ALL, MEASURES
URC	uncertainty coefficient $(R\,\vert\,C)$	ALL, MEASURES
WTKAP	weighted kappa coefficient	AGREE

* ALL and CHISQ compute Fisher's exact test only for 2×2 tables. Use the FISHER option to compute Fisher's exact test for general $r \times c$ tables.

TABLES Statement

Requests one-way to *n*-way frequency and crosstabulation tables and computes the statistics for these tables.

Default: 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.

Featured in: Example 1 on page 572

TABLES *request(s)* </ *option(s)*>;

Required Arguments

request(s)

specifies the frequency and crosstabulation tables to produce. A request is composed of one variable name or several variable names that are 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*>2), separate the desired variables with asterisks. The unique values of these variables form 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, the TABLES statement request A*B*C*D 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:

Request	Equivalent to
`tables a*(b c);`	`tables a*b a*c;`
`tables (a b)*(c d);`	`tables a*c b*c a*d b*d;`
`tables (a b c)*d;`	`tables a*d b*d c*d;`
`tables a--c;`	`tables a b c;`
`tables (a--c)*d;`	`tables a*d b*d c*d;`

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 *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

To do this	**Use this option**
Control statistical analysis	
Request tests and measures of classification agreement	AGREE
Request tests and measures of association produced by CHISQ, MEASURES, and CMH	ALL
Set the confidence level for confidence limits	ALPHA=
Request binomial proportion, confidence limits, and test for one-way tables	BINOMIAL
Request chi-square tests and measures of association based on chi-square	CHISQ
Request confidence limits for the MEASURES statistics	CL
Request all Cochran-Mantel-Haenszel statistics, adjusted relative risks, and odds ratios	CMH
Request adjusted relative risks and odds ratios and CMH correlation statistic	CMH1
Request adjusted relative risks and odds ratios, CMH correlation, and row mean scores (ANOVA) statistic	CMH2
Specify convergence criterion to compute polychoric correlation	CONVERGE=
Request Fisher's exact test for tables larger than 2×2	FISHER
Request Jonckheere-Terpstra test	JT
Specify maximum number of iterations to compute polychoric correlation	MAXITER=
Request measures of association and their asymptotic standard errors	MEASURES
Treat missing values as nonmissing	MISSING
Request polychoric correlation	PLCORR
Request relative risk measures for 2×2 tables	RELRISK
Request risks and risk differences for 2×2 tables	RISKDIFF
Specify the type of row and column scores	SCORES=
Specify expected frequencies for a one-way table chi-square test	TESTF=
Specify expected proportions for a one-way table chi-square test	TESTP=
Request Cochran-Armitage test for trend	TREND
Control additional table information	
Report each cell's contribution to the total Pearson chi-square statistic	CELLCHI2

To do this	Use this option
Display the cumulative column percentage in each cell	CUMCOL
Display the deviation of the cell frequency from the expected value for each cell	DEVIATION
Display the expected cell frequency for each cell	EXPECTED
Display missing value frequencies	MISSPRINT
List all possible combinations of variable levels even when a combination does not occur	SPARSE
Display percentage of total frequency on *n*-way tables when *n*>2	TOTPCT
Control displayed output	
Suppress the column percentage for each cell	NOCOL
Suppress the cumulative frequencies and the cumulative percentages in one-way frequency tables and in list format	NOCUM
Suppress the frequency count for each cell	NOFREQ
Suppress the percentage, row percentage, and column percentage in crosstabulation tables, or percentages and cumulative percentages in one-way frequency tables and in list format	NOPERCENT
Suppress the display of tables but report the statistics	NOPRINT
Suppress the row percentage for each cell	NOROW
Display two-way to *n*-way tables in list format	LIST
Display the kappa coefficient weights	PRINTKWT
Display the row and the column scores	SCOROUT
Use a field 8 positions wide to display the cell frequencies between 1.E7 and 1.E8	V5FMT
Create an output data set	
Specify an output data set to contain variable values and frequency counts	OUT=
Include the expected frequency of each cell in the output data set	OUTEXPECT
Include the percentage of column frequency, row frequency, and two-way table frequency in the output data set	OUTPCT

AGREE <(WT=*type*)>

requests tests and measures of classification agreement for square tables. The AGREE option provides McNemar's test for 2×2 tables and Bowker's test of symmetry for tables with more than two response categories. The AGREE option also produces the simple kappa statistic, the weighted kappa statistic, their asymptotic standard errors, and the corresponding confidence limits. When there are multiple strata, PROC FREQ computes overall simple and weighted kappa statistics, 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.

(WT=*type*)

specifies the type of weights that PROC FREQ uses to compute the weighted kappa coefficient, where *type* is the following:

CA	Cicchetti-Allison weights
FC	Fleiss-Cohen weights

Default: CA

Main discussion: “Weighted Kappa Coefficient” on page 554

Restriction: The table must be square.

Tip: You can specify PRINTKWT to display the kappa coefficient weights.

Main discussion: “Tests and Measures of Agreement” on page 551

Featured in: Example 9 on page 596

ALL

requests all tests and measures that are computed by the CHISQ, MEASURES, and CMH options.

Interaction: CMH1 and CMH2 control which CMH statistics PROC FREQ computes.

ALPHA=*p*

sets the confidence level for confidence limits. The percentage for the confidence limits is (1–*p*)×100. Using ALPHA=.05 results in 95 percent confidence limits. If *p* is between 0 and 1 but is outside the range, PROC FREQ uses the closest range endpoint. For example, if *p*=0.000001, PROC FREQ uses 0.0001 to determine confidence limits.

Default: 0.05

Range: 0.0001<=*p*<=0.9999

BINOMIAL <(P=*value*)>

computes the binomial proportion for one-way tables. This is the proportion of observations for the first variable level that appears in the output. BINOMIAL also computes the asymptotic standard error, asymptotic and exact confidence limits, and the asymptotic test for the binomial proportion. To specify the null hypothesis proportion value for the test, use P=.

Default: P=0.5

Restriction: for one-way tables

Interaction: To request an exact test for the binomial proportion, specify BINOMIAL in the EXACT statement.

Main Discussion: “Binomial Proportion” on page 543

Featured in: Example 3 on page 578

CELLCHI2

displays each cell’s contribution to the total Pearson chi-square statistic, which is computed as $(frequency - expected)^2/expected$.

Interaction: CELLCHI2 is valid for contingency tables but has no effect on tables that are produced with LIST.

CHISQ

computes chi-square tests of homogeneity or independence for two-way tables, and computes measures of association based on chi-square for two-way tables. The tests include 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 2×2 tables, CHISQ includes Fisher’s exact test and the continuity-adjusted chi-square. For one-way tables, CHISQ computes a chi-square goodness-of-fit test for equal proportions. If you specify the null hypothesis proportions with the TESTP= option, then CHISQ computes a chi-square goodness-of-fit test for the specified proportions. If you specify null hypothesis frequencies with the TESTF= option, CHISQ computes a chi-square goodness-of-fit test for the specified frequencies.

Main discussion: "Chi-Square Tests and Measures" on page 529

Featured in: Example 4 on page 580 and Example 5 on page 584

CL

requests confidence limits for the MEASURES statistics.

Interaction: If you omit MEASURES, CL invokes MEASURES.

Interaction: PROC FREQ determines the confidence coefficient using ALPHA= , which by default equals 0.05 and produces 95 percent confidence limits.

Main discussion: "Measures of Association" on page 533

Featured in: Example 7 on page 590

CMH

computes 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 2×2 tables, PROC FREQ computes adjusted Mantel-Haenszel and logit estimates of the odds ratio and relative risks as well as the corresponding confidence limits. For the stratified 2×2 case, PROC FREQ computes the Breslow-Day test for homogeneity of odds ratios.

Interaction: CMH1 and CMH2 control the number of CMH statistics that PROC FREQ computes.

Main discussion: "Cochran-Mantel-Haenszel Statistics" on page 557

Featured in: Example 6 on page 588

CMH1

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

CMH2

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

Featured in: Example 8 on page 593

CONVERGE=*c*

specifies the convergence criterion for computing the polychoric correlation using the PLCORR option. Iterative computation of the polychoric correlation stops when the convergence measure falls below the value of CONVERGE=, or when the number of iterations that is specified by the MAXITER= option is exceeded, whichever happens first.

Alias: CONV=

Default: 0.0001

Range: a positive number

Main discussion: “Polychoric Correlation” on page 541

CUMCOL

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

DEVIATION

displays the deviation of the cell frequency from the expected frequency for each cell of the crosstabulation table.

Interaction: DEVIATION is valid for crosstabulation tables but has no effect on tables produced with LIST.

Featured in: Example 5 on page 584

EXPECTED

displays the expected cell frequencies under the hypothesis of independence (or homogeneity).

Interaction: EXPECTED is valid for contingency tables but has no effect on tables produced with LIST.

Featured in: Example 5 on page 584

FISHER

computes Fisher’s exact test even when tables are larger than 2×2. You can also request Fisher’s exact test by specifying FISHER in the EXACT statement.

Alias: EXACT

Interaction: If you omit CHISQ, FISHER invokes CHISQ.

Interaction: ALL does not invoke this option.

Main discussion: “Fisher’s Exact Test” on page 532

CAUTION:

For large tables, PROC FREQ may require a large amount of time or memory to compute exact *p*-values. See “Computational Resources” on page 565 for more information. △

JT

performs the Jonckheere-Terpstra test.

Main discussion: “Jonckheere-Terpstra Test” on page 549

LIST

displays two-way to *n*-way tables in a list format rather than as crosstabulation tables.

Restriction: PROC FREQ ignores LIST when you request statistical tests or measures of association.

MAXITER=*n*

specifies the maximum number of iterations for computing the polychoric correlation using the PLCORR option. Iterative computation of the polychoric correlation stops when the number of iterations that is specified by MAXITER= is exceeded, or when the convergence measure falls below the value of the CONVERGE= option, whichever happens first.

Default: 20

Range: an integer between 0 and 32767

Main discussion: “Polychoric Correlation” on page 541

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*, Pearson and Spearman correlation coefficients, lambda (asymmetric and symmetric), uncertainty coefficients (asymmetric and symmetric) and, for 2×2 tables, the odds ratio, column 1 relative risk, column 2 relative risk, and the corresponding confidence limits.

Interaction: CL requests confidence limits.

Main discussion: "Measures of Association" on page 533

Featured in: Example 7 on page 590

MISSING

treats missing values as nonmissing and includes them in calculations of percentages and other statistics.

Main discussion: "Missing Values" on page 567

MISSPRINT

displays missing value frequencies for all tables, even though PROC FREQ does not use the frequencies in the calculation of statistics.

Main discussion: "Missing Values" on page 567

NOCOL

suppresses the column percentages in cells of the crosstabulation table.

Featured in: Example 5 on page 584

NOCUM

suppresses the cumulative frequencies and cumulative percentages for one-way frequency tables and for frequencies in list format.

Featured in: Example 2 on page 575

NOFREQ

suppresses the cell frequencies for a crosstabulation table. This also suppresses frequencies for row totals.

NOPERCENT

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

NOPRINT

suppresses the frequency and crosstabulation tables, but displays all requested tests and statistics.

Featured in: Example 6 on page 588

NOROW

suppresses the row percentages in cells of the crosstabulation table.

Featured in: Example 5 on page 584

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.

Main discussion: "Output Data Sets" on page 570

See also: OUTEXPECT and OUTPCT

Featured in: Example 1 on page 572

OUTEXPECT

includes the expected frequency in the output data set when you specify the OUT= option. The variable EXPECTED contains the expected frequency for each table cell.

Main discussion: "Output Data Sets" on page 570

Featured in: Example 1 on page 572

OUTPCT

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

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.

Main discussion: "Output Data Sets" on page 570

PLCORR

computes the polychoric correlation coefficient. For 2×2 tables, this statistic is more commonly known as the tetrachoric correlation coefficient, and is labeled as such in the displayed output.

Interaction: If you omit MEASURES, PLCORR invokes MEASURES.

Main discussion: "Polychoric Correlation" on page 541

See also: CONVERGE= and MAXITER=

PRINTKWT

requests that PROC FREQ display the kappa coefficient weights.

Interaction: You must specify AGREE to compute the kappa coefficients. The WT= option controls how PROC FREQ computes the kappa coefficient weights.

Main discussion: "Weighted Kappa Coefficient" on page 554

RELRISK

requests relative risk measures for 2×2 tables. These measures include the odds ratio, column 1 relative risk, and column 2 relative risk.

Main discussion: "Odds Ratio and Relative Risks for 2×2 Tables" on page 546

Featured in: Example 4 on page 580

RISKDIFF

requests column 1 and 2 risks (or binomial proportions), risk differences, and their confidence limits for 2×2 tables.

Alias: PDIFF, RDIFF

Main discussion: "Risks and Risk Differences" on page 545

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

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.

Default: TABLE

Main discussion: "Scores" on page 528

Featured in: Example 8 on page 593

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.

Restriction: SCOROUT displays the row and column scores only when statistics are computed for two-way tables.

Tip: To store the scores in an output data set, use the Output Delivery System.

Main discussion: "Scores" on page 528

See also: SCORES= on page 521

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. SPARSE has no effect unless you use the LIST or OUT= option. When you use SPARSE and LIST, PROC FREQ lists any combination of values with a frequency count of zero. When you use SPARSE and OUT= , PROC FREQ includes empty crosstabulation table cells in the output data set.

See also: "Missing Values" on page 567

Featured in: Example 1 on page 572

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.

Range: The sum of the frequency values must equal the total frequency for the one-way table.

Restriction: The number of TESTF= values must equal the number of variable levels in the one-way table. List these values in the order that the corresponding variable levels appear in the output.

Interaction: If you omit CHISQ, TESTF= invokes CHISQ.

Main discussion: "Chi-Square Test for One-Way Tables" on page 530

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.

Range: 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.

Restriction: The number of TESTP= values must equal the number of variable levels in the one-way table. List these values in the order that the corresponding variable levels appear in the output.

Interaction: If you omit CHISQ, TESTP= invokes CHISQ.

Main discussion: "Chi-Square Test for One-Way Tables" on page 530

Featured in: Example 2 on page 575

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.

Restriction: The table must be $2 \times c$ or $r \times 2$.

Main discussion: "Cochran-Armitage Test for Trend" on page 548

Featured in: Example 7 on page 590

V5FMT
uses a field that is 8 positions wide to display the cell frequencies between 1.E7 and 1.E8 so that PROC FREQ does not use scientific notation to display frequencies in this range. By default, PROC FREQ uses a maximum of 7 positions to display cell frequencies. In Version 5 of the SAS System, PROC FREQ used a maximum of 8 positions.

TEST Statement

Computes asymptotic tests for the specified measures of association and measures of agreement.

Requirement: TABLES statement

Main discussion: "Asymptotic Tests" on page 534

Featured in: Example 7 on page 590

TEST *statistic-keyword(s)*;

Required Arguments

statistic-keyword(s)
specifies the statistics for which to provide asymptotic tests. The available statistics are the measures of association and agreement listed in Table 21.4 on page 523. You can use an individual keyword to request a test, or you can use a group keyword (MEASURES or AGREE) to request all available tests in that group.

For each measure of association or agreement that you specify, the TEST statement provides an asymptotic test that the measure equal 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 limits for that measure. The ALPHA= option in the TABLES statement determines the confidence level, which by default equals .05 and provides 95 percent confidence limits. In addition to these asymptotic tests, exact tests for selected measures of association and agreement are available with the EXACT statement. See "EXACT Statement" on page 507 for more information.

Table 21.4 TEST Statement Statistic-keywords and Required TABLES Statement Options

Keyword	Asymptotic tests computed	Required TABLES statement option
AGREE	simple kappa coefficient and weighted kappa coefficient	AGREE
GAMMA	gamma	ALL, MEASURES
KAPPA	simple kappa coefficient	AGREE

Keyword	Asymptotic tests computed	Required TABLES statement option
KENTB	Kendall's tau-*b*	ALL, MEASURES
MEASURES	gamma, Kendall's tau-*b*, Stuart's tau-*c*, Somers' D ($C \mid R$), Somers' D ($R \mid C$), Pearson correlation coefficient, and Spearman correlation coefficient	ALL, MEASURES
PCORR	Pearson correlation coefficient	ALL, MEASURES
SCORR	Spearman correlation coefficient	ALL, MEASURES
SMDCR	Somers' D ($C \mid R$)	ALL, MEASURES
SMDRC	Somers' D ($R \mid C$)	ALL, MEASURES
STUTC	Stuart's tau-*c*	ALL, MEASURES
WTKAP	weighted kappa coefficient	AGREE

WEIGHT Statement

Treats observations as if they appear multiple times in the input data set.

Tip: Use to input the cell counts of an existing table.

Featured in: Example 1 on page 572

WEIGHT *variable*;

Required Arguments

variable

specifies a numeric variable whose value represents the frequency of the observation. 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 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.

Using Negative Weights

If any value of the weight variable is negative, PROC FREQ displays the frequencies (as measured by the weighted values), but does not compute and display percentages and other statistics. If you create an output data set using OUT= 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.

Concepts

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 in tabular form. A table lists all possible combinations of the 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:

```
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  15 2 2
;
```

You can store the same data as cell counts using the following DATA step statements:

```
data counts;
   input R C CellCount @@;
   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 variable CellCount contains the cell count for each row and column combination.

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

```
proc freq data=counts;
   weight CellCount;
   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 on a frequency table. If you decide to round each value to a single digit, include the statement

```
   format x 1.;
```

in the PROC FREQ step. 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 PROC FORMAT 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 answers as follows: 1 as yes, 2 as no, and 8 as no answer. The following PROC FORMAT step creates a user-written format:

```
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 MISSING or MISSPRINT in the TABLES statement to list the frequency for no answer. The frequency count for this level will include 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, see *SAS Language Reference: Dictionary*.

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 variables*10 levels/variable*84 bytes/level
```

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

```
1000 cells * 8 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 CMH or AGREE 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 use 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 “Grouping with Formats” on page 525.

Statistical Computations

This section gives the formulas PROC FREQ uses to compute the following:

□ chi-square tests and statistics (CHISQ option)

□ measures of association (MEASURES option)

□ binomial proportion (BINOMIAL option)

□ risks (or binomial proportions) and risk differences for 2×2 tables (RISKDIFF option)

□ odds ratios and relative risks for 2×2 tables (MEASURES or RELRISK option)

□ Jonckheere-Terpstra test (JT option)

□ Cochran-Armitage test for trend (TREND option)

□ tests and measures of agreement (AGREE option)

□ Cochran-Mantel-Haenszel statistics (CMH option)

Furthermore, this section describes the computation of exact *p*-values.

When selecting statistics to analyze your data, 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 the statistics detect, and the assumptions for valid interpretation of the statistics. For example, the Mantel-Haenszel chi-square statistic requires an ordinal scale for both variables and detects a linear association. On the other hand, the Pearson chi-square is appropriate for all variables and can detect any kind of association, but is less powerful for detecting a linear association. Select tests and measures carefully, choosing those that are appropriate for your data. For more information on when to use a statistic and how to interpret the results, refer to Agresti (1996) and Stokes et al. (1995).

Definitions and Notation

In this chapter, a two-way table represents the crosstabulation of two 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 ith row and the jth column and define the following:

$n_{i\cdot} = \sum_j n_{ij}$ (row totals)

$n_{\cdot j} = \sum_i n_{ij}$ (column totals)

$n = \sum_i \sum_j n_{ij}$ (overall total)

$p_{ij} = n_{ij}/n$ (cell percentages)

$p_{i\cdot} = n_{i\cdot}/n$ (row percentages)

$p_{\cdot j} = n_{\cdot j}/n$ (column percentages)

$R_i =$ score for row i

$C_j =$ score for column j

$\overline{R} = \sum_i n_{i\cdot} R_i/n$ (average row score)

$$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl}$$

$\overline{C} = \sum_j n_{\cdot j} C_j/n$ (average column score)

$$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl}$$

$$D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl}$$

$P = \sum_i \sum_j n_{ij} A_{ij}$ (twice the number of concordances)

$Q = \sum_i \sum_j n_{ij} D_{ij}$ (twice the number of discordances)

Scores

PROC FREQ uses row and column scores 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

$$\begin{aligned}&\text{Row scores :}\\&\qquad R1_i = \sum_{k<i} n_{k\cdot} + (n_{i\cdot} + 1)/2 \quad i = 1, 2, \ldots, R\\&\text{Column scores :}\\&\qquad C1_j = \sum_{l<j} n_{\cdot l} + (n_{\cdot j} + 1)/2 \quad j = 1, 2, \ldots, C\end{aligned}$$

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

$$\begin{aligned}R2_i &= R1_i/n\\C2_j &= C1_j/n\end{aligned}$$

Modified ridit (MODRIDIT) scores (van Elteren 1960 and 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

$$\begin{aligned}R3_i &= R1_i/(n+1)\\C3_j &= C1_j/(n+1)\end{aligned}$$

Chi-Square Tests and Measures

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 2×2 tables, likelihood-ratio chi-square, Mantel-Haenszel chi-square, and Fisher's exact test for 2×2 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. See "Chi-Square Test for One-Way Tables" on page 530 for information. 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 "Exact Statistics" on page 563 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 et al. (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,

$$e_i = n/C \qquad \text{for} \quad i = 1, 2, \ldots, C$$

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

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

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

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

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 option or CHISQ option in the EXACT statement.

For a 2×2 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 option or the CHISQ option in the EXACT statement.

Continuity-Adjusted Chi-Square Test

The continuity-adjusted chi-square statistic for 2×2 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

$$Q_C = \sum_i \sum_j \frac{\left[\max\left(0, |n_{ij} - e_{ij}| - 0.5\right)\right]^2}{e_{ij}}$$

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 "Pearson Correlation Coefficient" on page 538. The Pearson correlation, and thus the Mantel-Haenszel chi-square statistic, use the scores 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 option or the CHISQ option in the EXACT statement.

Refer to Mantel and Haenszel (1959) and Landis et al. (1978).

Fisher's Exact Test

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 that 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 the first cell is greater than that 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 less than 1, and the right-sided alternative is equivalent to an odds ratio greater than 1, where the odds ratio equals $(n_{11}n_{22}/n_{12}n_{21})$. 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).

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 2×2 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 "Exact Statistics" on page 563 for more information.

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 2×2 tables. Otherwise, the range is $0 \leq \phi \leq \min\left(\sqrt{R-1}, \sqrt{C-1}\right)$ (Liebetrau, 1983). The phi coefficient is computed as

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

$$\phi = \sqrt{Q_P/n} \qquad \text{otherwise.}$$

Refer to Fleiss (1981, pp 59-60).

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 limit is always 1. It has the range $-1 \leq V \leq 1$ for 2×2 tables; otherwise, the range is $0 \leq V \leq 1$. Cramer's *V* is computed as

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

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

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 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 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 are appropriate for nominal variables: lambda asymmetric and symmetric, and the 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, which is the square root of the asymptotic variance denoted by *var* in the following sections.

Confidence limits

If you specify the CL option in the TABLES statement, PROC FREQ computes asymptotic confidence limits 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 percent confidence limits. The confidence limits are computed as

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

where 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: gamma=0 equal the p-values for the test of H_0: 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

$$\begin{aligned} P_1 = &\ \text{Prob}\,(Z > z) \quad \text{if} \quad z > 0 \\ P_1 = &\ \text{Prob}\,(Z < z) \quad \text{if} \quad z \le 0 \end{aligned}$$

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 "Exact Statistics" on page 563 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 $-1 \le \Gamma \le 1$. 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

$$var = \frac{16}{(P+Q)^4} \sum_i \sum_j n_{ij} \left(QA_{ij} - PD_{ij}\right)^2$$

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

$$var_0(G) = \frac{4}{(P+Q)^2}\left(\sum_i\sum_j n_{ij}(A_{ij}-D_{ij})^2 - (P-Q)^2/n\right)$$

For 2×2 tables, gamma is equivalent to Yule's Q. Refer to Goodman and Kruskal (1963; 1972), Brown and Benedetti (1977), and Agresti (1990).

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

$$t_b = \frac{(P-Q)}{\sqrt{w_r w_c}}$$

with

$$var = \frac{1}{w^4}\left(\sum_i\sum_j n_{ij}(2wd_{ij} + t_b v_{ij})^2 - n^3 t_b^2 (w_r + w_c)^2\right)$$

where

$$\begin{aligned}
w &= \sqrt{w_r w_c} \\
w_r &= n^2 - \sum_i n_{i\cdot}^2 \\
w_c &= n^2 - \sum_j n_{\cdot j}^2 \\
d_{ij} &= A_{ij} - D_{ij} \\
v_{ij} &= n_{i\cdot} w_c + n_{\cdot j} w_r
\end{aligned}$$

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

$$var_0(t_b) = \frac{4}{w_r w_c}\left(\sum_i\sum_j n_{ij}(A_{ij}-D_{ij})^2 - (P-Q)^2/n\right)$$

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

$$t_c = \frac{m\left(P - Q\right)}{n^2\left(m - 1\right)}$$

with

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

where

$$m = \min\left(R, C\right)$$
$$d_{ij} = A_{ij} - D_{ij}$$

The variance of the estimator under the null hypothesis that tau-*c* equals zero is the same as var in the above equation.

$$var_0\left(t_c\right) = var$$

Refer to Brown and Benedetti (1977).

Somers' *D*

Somers' $D\left(C|R\right)$ and Somers' $D\left(R|C\right)$ 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' $D\left(R|C\right)$ are obtained by interchanging the indices:

$$D\left(C|R\right) = \frac{\left(P - Q\right)}{w_r}$$

with

$$var = \frac{4}{\mathrm{w}_r^4} \sum_i \sum_{\mathrm{j}} n_{ij}\left(w_r d_{ij} - \left(P - Q\right)\left(n - n_{i\cdot}\right)\right)^2$$

where

$$w_r = n^2 - \sum_i n_{i\cdot}^2$$
$$d_{ij} = A_{ij} - D_{ij}$$

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

$$var_0\left(D\left(C|R\right)\right) = \frac{4}{w_r^2}\left(\sum_i \sum_j n_{ij}\left(A_{ij} - D_{ij}\right)^2 - \left(P - Q\right)^2/n\right)$$

Refer to Somers (1962) and Goodman and Kruskal (1972).

Pearson Correlation Coefficient

PROC FREQ computes the Pearson correlation coefficient using the scores specified in the 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

$$r = \frac{v}{w} = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}$$

with

$$var = \frac{1}{w^4}\sum_i\sum_j n_{ij}\left(w\left(R_i - \overline{R}\right)\left(C_j - \overline{C}\right) - \frac{b_{ij}v}{2w}\right)^2$$

The row scores R_i and the column scores C_j are determined by the SCORES= option in the TABLES statement. Then

$$ss_r = \sum_i\sum_j n_{ij}\left(R_i - \overline{R}\right)^2$$

$$ss_c = \sum_i\sum_j n_{ij}\left(C_j - \overline{C}\right)^2$$

$$ss_{rc} = \sum_i\sum_j n_{ij}\left(R_i - \overline{R}\right)\left(C_j - \overline{C}\right)$$

$$b_{ij} = \left(R_i - \overline{R}\right)^2 ss_c + \left(C_j - \overline{C}\right)^2 ss_r$$

$$v = ss_{rc}$$

$$w = \sqrt{ss_r ss_c}$$

where $\overline{R}$ and $\overline{C}$ are the average row and columns scores as defined in “Definitions and Notation” on page 527. 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. The standardized test statistic is computed as

$$r^* = \frac{r}{\sqrt{var_0\left(r\right)}}$$

where $var_0\left(r\right)$ is the variance of the correlation under the null hypothesis.

$$var_0\left(r\right) = \frac{\sum_i \sum_j n_{ij}\left(R_i - \overline{R}\right)^2 \left(C_j - \overline{C}\right)^2 - ss_{rc}^2/n}{ss_r ss_c}$$

This 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 "Exact Statistics" on page 563 for more information on exact tests.

Spearman Rank Correlation Coefficient

The Spearman correlation coefficient is computed using rank scores $R1_i$ and $C1_j$, defined in "Scores" on page 528. 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

$$var = \frac{1}{n^2 w^4} \sum_i \sum_j n_{ij}\left(z_{ij} - \overline{z}\right)^2$$

where

$$v = \sum_i \sum_j n_{ij} R(i)\, C(j)$$

$$w = \frac{1}{12}\sqrt{FG}$$

$$F = n^3 - \sum_i n_{i\cdot}^3$$
$$G = n^3 - \sum_j n_{\cdot j}^3$$
$$R(i) = R1_i - \frac{n}{2}$$
$$C(j) = C1_j - \frac{n}{2}$$

$$\overline{z} = \frac{1}{n}\sum_i \sum_j n_{ij} z_{ij}$$
$$z_{ij} = wv_{ij} - vw_{ij}$$
$$\begin{aligned} v_{ij} = n(R(i)\, C(j) &+ \frac{1}{2}\sum_l n_{il} C(l) + \frac{1}{2}\sum_k n_{kj} R(k) \\ &+ \sum_l \sum_{k>i} n_{kl} C(l) + \sum_k \sum_{l>j} n_{kl} R(k)) \end{aligned}$$

$$w_{ij} = \frac{-n}{96w}\left(Fn_{\cdot j}^2 + Gn_{i\cdot}^2\right)$$

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. The standardized test statistic is computed as

$$r_s^* = \frac{r_s}{\sqrt{var_0\,(r_s)}}$$

where $var_0\,(r_s)$ is the variance of the correlation under the null hypothesis.

$$var_0\,(r_s) = \frac{1}{n^2 w^2} \sum_i \sum_j n_{ij}\left(v_{ij} - \overline{v}\right)^2$$

where

$$\overline{v} = \sum_i \sum_j n_{ij} v_{ij} / n$$

This 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. 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 "Exact Statistics" on page 563 for more information.

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 2×2 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. 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. 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 is 0.0001. The MAXITER= option sets the maximum number of iterations, and the default is 20.

Lambda Asymmetric

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

$$\lambda\left(C|R\right) = \frac{\sum_i r_i - r}{n - r}$$

with

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

where

$$r_i = \max_j \left(n_{ij}\right)$$
$$r = \max_j \left(n_{\cdot j}\right)$$

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 $r_i = n_{\cdot j}$.

Because of the uniqueness assumptions, ties in the frequencies or inthe 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 $r = n_{\cdot j}$. 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 (1963).

Lambda Symmetric

The nondirectional lambda is the average of the two asymmetric lambdas. Lambda symmetric has the range $0 \le \lambda \le 1$. Lambda symmetric is defined as

$$\lambda = \frac{\left(\sum_i r_i + \sum_j c_j - r - c\right)}{(2n - r - c)} = \frac{(w - v)}{w}$$

with

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

where

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

Refer to Goodman and Kruskal (1963).

Uncertainty Coefficient Asymmetric

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 $0 \le U(C|R) \le 1$. The formulas for $U(R|C)$ are obtained by interchanging the indices.

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

with

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

where

$$\begin{aligned}
&v = H\left(X\right) + H\left(Y\right) - H\left(XY\right) \\
&w = H\left(Y\right) \\
&H\left(X\right) = -\sum_i \left(\frac{n_{i\cdot}}{n} \right) \ln \left(\frac{n_{i\cdot}}{n} \right) \\
&H\left(Y\right) = -\sum_j \left(\frac{n_{\cdot j}}{n} \right) \ln \left(\frac{n_{\cdot j}}{n} \right) \\
&H\left(XY\right) = -\sum_i \sum_j \left(\frac{n_{ij}}{n} \right) \ln \left(\frac{n_{ij}}{n} \right)
\end{aligned}$$

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

Uncertainty Coefficient Symmetric

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

$$U = \frac{2\left(H\left(X\right) + H\left(Y\right) - H\left(XY\right)\right)}{H\left(X\right) + H\left(Y\right)}$$

with

$$var = 4 \sum_i \sum_j \frac{n_{ij} \left(H\left(XY\right) \ln \left(\frac{n_{i\cdot} n_{\cdot j}}{n^2} \right) - \left[H(X) + H(Y) \right] \ln \left(\frac{n_{ij}}{n} \right) \right)^2}{n^2 \left(H\left(X\right) + H\left(Y\right) \right)^4}$$

Refer to Goodman and Kruskal (1972).

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 for the first variable level, or class, that appears in the output.

$$\hat{p} = n_1 / n$$

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

$$se\left(\hat{p}\right) = \sqrt{\hat{p}\left(1-\hat{p}\right)/n}$$

Using the normal approximation to the binomial distribution, PROC FREQ constructs asymptotic confidence limits for p according to

$$\hat{p} \pm z_{\alpha/2} \cdot se\left(\hat{p}\right)$$

where $z_{\alpha/2}$ is the $100\left(1-\alpha/2\right)$ percentile of the standard normal distribution. The confidence level α is determined by the ALPHA= option, which by default equals .05 and produces 95 percent confidence limits. Additionally, PROC FREQ computes exact confidence limits 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 P=, PROC FREQ uses $p_0 = 0.5$ by default. The asymptotic test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0\left(1-p_0\right)/n}}$$

PROC FREQ computes one-sided and two-sided *p*-values for this test. 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 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

$$\begin{aligned} P_1 &= \text{Prob}\ \left(Z > z\right) \quad \text{if}\ \ z > 0 \\ P_1 &= \text{Prob}\ \left(Z < z\right) \quad \text{if}\ \ z \leq 0 \end{aligned}$$

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

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

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

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

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

$$P_1 = \min\left(\mathrm{Prob}\left(X \leq n_1 | p_0\right), \mathrm{Prob}\left(X \geq n_1 | p_0\right)\right)$$

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 2×2 tables. This analysis may be appropriate when you are 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 et al. (1995).

Let the frequencies of the 2×2 table be represented as follows:

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

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

$$p_{1|1} = n_{11}/n_{1\cdot}$$

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,

$$p_{1|2} = n_{21}/n_{2\cdot}$$

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

$$p_{\cdot 1} = n_{\cdot 1}/n$$

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,

$$(pdiff)_1 = p_{1|1} - p_{1|2}$$

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

$$se\left(p_{1|i}\right) = \sqrt{p_{1|i}\left(1 - p_{1|i}\right)/n_{i\cdot}}$$

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

$$se\left(p_{\cdot 1}\right) = \sqrt{p_{\cdot 1}\left(1 - p_{\cdot 1}\right)/n}$$

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

$$se\left((pdiff)_1\right) = \sqrt{var\left(p_{1|1}\right) + var\left(p_{1|2}\right)}$$

The standard errors are computed similarly for the column 2 risks and risk difference.

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

$$est \ \pm \ z_{\alpha/2} \cdot \ se\left(est\right)$$

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

PROC FREQ computes exact confidence limits 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 limits for the risk differences. Refer to Agresti (1992) for a discussion of issues involved in constructing exact confidence limits 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 et al. (1995) and Agresti (1996).

The odds of a positive response (column 1) in row 1 is n_{11}/n_{12}. Similarly, the odds of 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

$$\mathrm{OR} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

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 $G = (\mathrm{OR} - 1)\,/\,(\mathrm{OR} + 1)$ transforms the odds ratio to the range $(-1, 1)$ such that $G = 0$ when $\mathrm{OR} = 1$, $G = -1$ when $\mathrm{OR} = 0$, and G is close to 1 for very large values of OR. G is the gamma statistic, which PROC FREQ computes when you specify the MEASURES option.

The asymptotic $100\,(1-\alpha)$ percent confidence limits for the odd ratio are

$$\left(\mathrm{OR}\cdot\ \exp\left(-z\sqrt{v}\right),\ \mathrm{OR}\cdot\ \exp\left(z\sqrt{v}\right)\right)$$

where

$$v = var\,(\ln\ \mathrm{OR}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

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 the OR option in the EXACT statement PROC FREQ computes exact confidence limits 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 limits is not exactly $1-\alpha$, but is at least $1-\alpha$. Thus, these confidence limits 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 variables are observed simultaneously. Refer to Stokes et al. (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,

$$p_{1|1} = n_{11}/n_{1\cdot}$$

Similarly, the column 1 risk for row 2 is

$$p_{1|2} = n_{21}/n_{2\cdot}$$

The column 1 relative risk is then computed as

$$\mathrm{RR}_1 = \frac{p_{1|1}}{p_{1|2}}$$

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 that is 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 $100\,(1-\alpha)$ percent confidence limits for the column 1 relative risk are

$$\left(\mathrm{RR}_1\cdot\exp\left(-z\sqrt{v}\right),\ \mathrm{RR}_1\cdot\exp\left(z\sqrt{v}\right)\right)$$

where

$$v = var\,(\ln\ \mathrm{RR}_1) = \frac{1-p_{1|1}}{n_{11}} + \frac{1-p_{1|2}}{n_{21}}$$

and z is the $100\,(1-\alpha/2)$ percentile of the standard normal distribution. If either n_{11} or n_{21} is zero, PROC FREQ does not compute the relative risks.

The column 2 relative risks are computed similarly.

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. When the table has two rows and *C* columns, PROC FREQ tests for trend across the *C* levels of the column variable.

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

$$T = \frac{\sum\limits_{i=1}^{R} n_{i1}\left(R_i - \overline{R}\right)}{\sqrt{p_{\cdot 1}\left(1-p_{\cdot 1}\right)s^2}}$$

where

$$s^2 = \sum_{i=1}^{R} n_{i\cdot} \left(R_i - \overline{R}\right)^2$$

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 the binomial proportion $p_{i1} = n_{i1}/n_{i\cdot}$ 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 test for trend, which you request by specifying the TREND option in the EXACT statement. See the "EXACT Statement" on page 507 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 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 column 1 probability from row 1 to row *R*. When the 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 trend. The one-sided *p*-value P_1 can be expressed as

$$\begin{array}{ll} P_1 = \text{Prob (Trend} \;\; \text{Statistic} > T) & \text{if } T > 0 \\ P_1 = \text{Prob (Trend} \;\; \text{Statistic} < T) & \text{if } T \leq 0 \end{array}$$

The two-sided *p*-value P_2 is computed as

$$P_2 = \text{Prob } (|\text{Trend} \;\; \text{Statistic}| > |T|)$$

Jonckheere-Terpstra Test

The JT option in the TABLES statement requests the Jonckheere-Terpstra test, which is a nonparametric test for ordered differences among classes. It tests the null hypothesis that the distribution of the response variable does not differ among classes. It is designed to detect alternatives of ordered class differences, which can be expressed as $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_R$ (or $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_R$) with at least one of the inequalities being strict, where τ_i denotes the effect of class i. For such ordered alternatives, the Jonckheere-Terpstra test can be preferable to tests of more general class difference alternatives, such as the Kruskal-Wallis test (requested by the WILCOXON option in the NPAR1WAY procedure). Refer to Pirie (1983) and Hollander and Wolfe (1973) for more information about the Jonckheere-Terpstra test.

The Jonckheere-Terpstra test is appropriate for a contingency table where an ordinal column variable represents the response. The row variable, which can be nominal or ordinal, represents the classification variable. The levels of the row variable should be ordered according to the ordering you want the test to detect. The order of variable levels is determined by the ORDER= option in the PROC FREQ statement. The default is ORDER=INTERNAL, which orders by unformatted value. If you specify ORDER=DATA, PROC FREQ orders values according to their order in the input data set. For more information on how to order variable levels, see the ORDER= option on page 506.

The Jonckheere-Terpstra test statistic is computed by first forming $R\left(R-1\right)/2$ Mann-Whitney counts $M_{i,i'}$, where $i < i'$, for pairs of rows in the contingency table,

$$\begin{aligned} M_{i,i'} = &\left\{\text{number of times } X_{i,j} < X_{i',j'}, j = 1,\ldots,n_{i\cdot};\ j' = 1,\ldots,n_{i'\cdot}\right\} + \\ &\frac{1}{2}\left\{\text{number of times } X_{i,j} = X_{i',j'}, j = 1,\ldots,n_{i\cdot};\ j' = 1,\ldots,n_{i'\cdot}\right\} \end{aligned}$$

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

$$J = \sum_{1 \leq i <} \sum_{i' \leq \mathrm{R}} M_{i,i'}$$

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

$$J^* = \frac{J - E_0\left(J\right)}{\sqrt{var_0\left(J\right)}}$$

where E_0 and $var_0\left(J\right)$ are the expected value and variance of the test statistic under the null hypothesis.

$$E_0\left(J\right) = \left(n^2 - \sum_i n_{i\cdot}^2\right)/4$$

$$var_0\left(J\right) = A/72 + B/\left[36n\left(n-1\right)\left(n-2\right)\right] + C/\left[8n\left(n-1\right)\right]$$

where

$$
\begin{aligned}
A &= n\,(n-1)\,(2n+5) - \sum_i n_{i\cdot}\,(n_{i\cdot}-1)\,(2n_{i\cdot}+5) \\
&\quad - \sum_j n_{\cdot j}\,(n_{\cdot j}-1)\,(2n_{\cdot j}+5) \\
B &= \left[\sum_i n_{i\cdot}\,(n_{i\cdot}-1)\,(n_{i\cdot}-2)\right]\left[\sum_j n_{\cdot j}\,(n_{\cdot j}-1)\,(n_{\cdot j}-2)\right] \\
C &= \left[\sum_i n_{i\cdot}\,(n_{i\cdot}-1)\right]\left[\sum_j n_{\cdot j}\,(n_{\cdot j}-1)\right]
\end{aligned}
$$

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 "EXACT Statement" on page 507 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 expected value of 0, 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 0, 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

$$
\begin{aligned}
P_1 &= \text{Prob (Std JT Statistic} > J^*) \quad \text{if } J^* > 0 \\
P_1 &= \text{Prob (Std JT Statistic} < J^*) \quad \text{if } J^* \leq 0
\end{aligned}
$$

The two-sided *p*-value, P_2, is computed as

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

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 2×2 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. For multiple strata of 2×2 tables, PROC FREQ computes Cochran's ***Q***.

PROC FREQ computes the kappa coefficients (simple and weighted), their asymptotic standard errors, and their confidence limits 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 WTKAP in the TEST statement, PROC FREQ computes the asymptotic test for weighted kappa.

In addition to the asymptotic tests that are described in this section, PROC FREQ also computes the exact p-value for McNemar's test when you specify the keyword MCNEM in the EXACT statement. For the kappa statistic, PROC FREQ computes an exact test of the hypothesis that kappa (or weighted kappa) equals zero when you specify KAPPA (or WTKAP) in the EXACT statement. See "Exact Statistics" on page 563 for more information about these tests.

The discussion of each test and measure of agreement provides the formulas that PROC FREQ uses to compute the AGREE statistics. For information about the use and interpretation of these statistics, refer to Agresti (1990), Agresti (1996), Fleiss (1981), and the references that follow.

McNemar's Test

PROC FREQ computes McNemar's test for 2×2 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 for marginal homogeneity, or a null hypothesis of $p_{1\cdot} = p_{\cdot 1}$. McNemar's test is computed as

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

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 on page 552 in the preceding section. PROC FREQ also computes an exact p-value for McNemar's test when you specify MCNEM in the EXACT statement.

Bowker's Test of Symmetry

PROC FREQ computes Bowker's test of symmetry for square two-way tables that are larger than 2×2. (For 2×2 tables, Bowker's test is identical to McNemar's test.) 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 for each variable, Bowker's test of symmetry is calculated as

$$Q_B = \sum_{i<j}\sum \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}}$$

For large samples, Q_B has an asymptotic chi-square distribution with $R\left(R-1\right)/2$ degrees of freedom under the null hypothesis of symmetryof 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:

$$\widehat{\kappa} = \frac{P_0 - P_e}{1 - P_e}$$

where $P_0 = \Sigma_i p_{ii}$ and $P_e = \Sigma_i p_{i\cdot} p_{\cdot i}$. Viewing the two response variables as two independentratings of the n subjects, the kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds chance agreement, the kappa coefficient is positive, with its magnitude reflecting the strength of agreement. Although 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 is estimated by the following, according to Fleiss et al. (1969):

$$var = \frac{A + B - C}{(1 - P_e)^2\, n}$$

where

$$A = \sum_i p_{ii} \left[1 - (p_{i\cdot} + p_{\cdot i})(1 - \widehat{\kappa})\right]^2$$

$$B = (1 - \widehat{\kappa})^2 \sum_{i \neq j} \sum p_{ij} (p_{\cdot i} + p_{j\cdot})^2$$

and

$$C = \left[\widehat{\kappa} - P_e (1 - \widehat{\kappa})\right]^2$$

PROC FREQ computes confidence limits for the simple kappa coefficient according to

$$\widehat{\kappa} \pm z_{\alpha/2} \cdot \sqrt{var}$$

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

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

$$\widehat{\kappa}^* = \frac{\widehat{\kappa}}{\sqrt{var_0\,(\widehat{\kappa})}}$$

where $var_0\,(\widehat{\kappa})$ is the variance of the kappa coefficient under the nullhypothesis.

$$var_0\,(\widehat{\kappa}) = \frac{P_e + P_e^2 - \sum_i p_{i\cdot} p_{\cdot i} (p_{i\cdot} + p_{\cdot i})}{(1 - P_e)^2\, n}$$

Refer to Fleiss (1981).

In addition to the asymptotic test for kappa, PROC FREQ computes an exact test when you specify the KAPPA option or the AGREE option in the EXACT statement. See "Exact Statistics" on page 563 for more 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. PROC FREQ computes the weights from the column scores, using either the Cicchetti-Allison weight type or the Fleiss-Cohen weight type, which are described below. The weights w_{ij} are constructed so that $0 \leq w_{ij} < 1$ for all $i \neq j$, $w_{ii} = 1$ for all i, and $w_{ij} = w_{ji}$. The weighted kappa coefficient is defined as

$$\widehat{\kappa}_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}}$$

where

$$P_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}$$

and

$$P_{e(w)} = \sum_i \sum_j w_{ij} p_{i\cdot} p_{\cdot j}$$

For 2×2 tables, the weighted kappa coefficient is identical to the simple kappa coefficient. Therefore, PROC FREQ displays only the simple kappa coefficient for 2×2 tables. The asymptotic variance of the weighted kappa coefficient is estimated by the following, according to Fleiss et al. (1969):

$$var = \frac{\sum_i \sum_j p_{ij} \left[w_{ij} - (\overline{w}_{i\cdot} + \overline{w}_{\cdot j})(1 - \widehat{\kappa}_w) \right]^2 - \left[\widehat{\kappa}_w - P_{e(w)}(1 - \widehat{\kappa}_w) \right]^2}{\left(1 - P_{e(w)}\right)^2 n}$$

where

$$\overline{w}_{i\cdot} = \sum_j p_{\cdot j} w_{ij}$$

and

$$\overline{w}_{\cdot j} = \sum_i p_{i\cdot} w_{ij}$$

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

$$\widehat{\kappa}_w \pm z_{\alpha/2} \cdot \sqrt{var}$$

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

To compute an asymptotic test for the weighted kappa coefficient, PROC FREQ uses a standardized test statistic $\widehat{\kappa}_w^*$, which has an asymptotic standard normal distribution underthe null hypothesis. The standardized test statistic is computed as

$$\widehat{\kappa}_w^* = \frac{\widehat{\kappa}_w}{\sqrt{var_0\left(\widehat{\kappa}_w\right)}}$$

where $var_0\left(\widehat{\kappa}_w\right)$ is the variance of the kappa coefficient under the null hypothesis.

$$var_0\left(\widehat{\kappa}_w\right) = \frac{\sum_i \sum_j p_{i\cdot} p_{\cdot j} \left[w_{ij} - (\overline{w}_{i\cdot} + \overline{w}_{\cdot j})\right]^2 - P_{e(w)}^2}{\left(1 - P_{e(w)}\right)^2 n}$$

Refer to Fleiss (1981).

In addition to the asymptotic test for weighted kappa, PROC FREQ computes the exact test when you specify the WTKAP option or the AGREE option in the EXACT statement. See “Exact Statistics” on page 563 for more information on exact tests.

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. By default, PROC FREQ uses the Cicchetti-Allison type. If you specify WT=FC in the AGREE option, then PROC FREQ uses the Fleiss-Cohen weight type to construct kappa weights. To display the kappa weights, specify the PRINTKWT option in the TABLES statement.

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. You can specify the type of score 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 numeric values of the variable levels. 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: $w_{12} = .8, w_{13} = .6, w_{14} = 0, w_{23} = .8, w_{24} = .2$, and $w_{34} = .4$.

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).

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

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 $var\left(\widehat{\kappa}_h\right)$ denote the variance of $\widehat{\kappa}_h$. Then the estimate of the overall kappa, according to Fleiss (1981), is computed as follows:

$$\widehat{\kappa}_{overall} = \sum_{h=1}^{q} \frac{\widehat{\kappa}_h}{var\left(\widehat{\kappa}_h\right)} \Big/ \sum_{h=1}^{q} \frac{1}{var\left(\widehat{\kappa}_h\right)}$$

An estimate of the overall weighted kappa is computed similarly.

Tests for Equal Kappa Coefficients

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

$$Q_\kappa = \sum_{h=1}^{q} \frac{\left(\widehat{\kappa}_h - \widehat{\kappa}_{overall}\right)^2}{var\left(\widehat{\kappa}_h\right)}$$

A similar test is done for weighted kappa coefficients.

Cochran's *Q* Test

When there are multiple strata and two response categories, 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 follows:

$$Q_C = (m-1)\,\frac{m \sum\limits_{j=1}^{m} T_j^2 - T^2}{mT - \sum\limits_{k=1}^{N} S_k^2}$$

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 two variables ($m = 2$), Cochran's *Q* simplifies to McNemar's statistic. 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 or 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. However, PROC FREQ does compute AGREE statistics for tables with a zero row or a zero column. Therefore, 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 "WEIGHT Statement" on page 524). And, for a two-way table where there is no stratification, the analysis includes only those levels that occur with nonzero weight.

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 for 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 a weight of 1E-8. This forms a 5×5 square table for the analysis.

Cochran-Mantel-Haenszel Statistics

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

```
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, 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 2×2 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 $n_{hi\cdot}$ and $n_{h\cdot j}$ 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 (′).

$$
\begin{aligned}
\mathbf{n}'_{hi} &= (n_{hi1}, n_{hi2}, \ldots, n_{hi\mathrm{C}}) && (1{\times}C) \\
\mathbf{n}'_{h} &= (\mathbf{n}'_{h1}, \mathbf{n}'_{h2}, \ldots, \mathbf{n}'_{h\mathrm{R}}) && (1{\times}RC) \\
p_{hi\cdot} &= \frac{n_{hi\cdot}}{n_h} && (1{\times}1) \\
p_{h\cdot j} &= \frac{n_{h\cdot j}}{n_h} && (1{\times}1) \\
\mathbf{P}'_{h*\cdot} &= (p_{h1\cdot}, p_{h2\cdot}, \ldots, p_{h\mathrm{R}\cdot}) && (1{\times}R) \\
\mathbf{P}'_{h\cdot *} &= (p_{h\cdot 1}, p_{h\cdot 2}, \ldots, p_{h\cdot\mathrm{C}}) && (1{\times}C)
\end{aligned}
$$

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, which implies that under H_0, the expected value and covariance matrix of the frequencies are, respectively,

$$\mathbf{m}_h = \mathbf{E}\left[\mathbf{n}_h | \mathrm{H}_0\right] = n_h \left(\mathbf{P}_{h\cdot *} \otimes \mathbf{P}_{h*\cdot}\right)$$

and

$$\mathbf{var}\left[\mathbf{n}_h | \mathrm{H}_0\right] = c\left[\left(\mathbf{D}_{\mathbf{P}h\cdot *} - \mathbf{P}_{h\cdot *}\mathbf{P'}_{h\cdot *}\right) \otimes \left(\mathbf{D}_{\mathbf{P}h*\cdot} - \mathbf{P}_{h*\cdot}\mathbf{P}'_{h*\cdot}\right)\right]$$

where

$$c = \frac{n_h^2}{n_h - 1}$$

and where $\otimes$ denotes Kronecker product multiplication and $\mathbf{D}_{\mathbf{a}}$ is a diagonal matrix with elements of $\mathbf{a}$ on the main diagonal.

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

$$Q_{\mathrm{CMH}} = \mathbf{G}'\mathbf{V}_{\mathbf{G}}^{-1}\mathbf{G}$$

where

$$\mathbf{G} = \sum_h \mathbf{B}_h \left(\mathbf{n}_h - \mathbf{m}_h\right)$$
$$\mathbf{V}_{\mathbf{G}} = \sum_h \mathbf{B}_h \left(\mathbf{Var}\left(\mathbf{n}_h | \mathrm{H}_0\right)\right) \mathbf{B}'_h$$

and where

$$\mathbf{B}_h = \mathbf{C}_h \otimes \mathbf{R}_{\mathrm{h}}$$

is a matrix of fixed constants based on column scores $\mathbf{C}_h$ and row scores $\mathbf{R}_{\mathrm{h}}$. When the null hypothesis is true, the CMH statistic has an asymptotic chi-square distribution with degrees of freedom equal to the rank of $\mathbf{B}_h$. If $\mathbf{V}_{\mathbf{G}}$ is found to be singular, PROC FREQ displays 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:

CMH statistics have low power for detecting an association when 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, with one degree of freedom, was popularized by Mantel and Haenszel (1959) and Mantel (1963) and is therefore known as the Mantel-Haenszel statistic.

The alternative hypothesis 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 meaningless.

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 "Scores" on page 528 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 you specify nonparametric (RANK, RIDIT, or MODRIDIT) scores, 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 the 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 $1 \times C$, and the scores, one for each column, are specified in the SCORES= option. The matrix $\mathbf{R}_h$ has dimension $(R-1) \times R$ which PROC FREQ creates internally as

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

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, then this CMH statistic is identical to Friedman's chi-square. See Example 8 on page 593 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 is 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 because it has only $(R-1)\,(C-1)$ degrees of freedom.

Refer to Cochran (1954); Mantel and Haenszel (1959); Mantel (1963); Birch (1965); and Landis et al. (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,

```
proc freq;
   tables a*b*c*d / cmh;
run;
```

In this example, if the row and column 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 "Odds Ratio and Relative Risks for 2×2 Tables" on page 546 for more information on these measures.

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

Odds Ratio (Case-control Studies): Mantel-Haenszel Adjusted

The Mantel-Haenszel adjusted odds ratio estimator is given by

$$\mathrm{OR}_{\mathrm{MH}} = \frac{\sum_h n_{h11} n_{h22} / n_h}{\sum_h n_{h12} n_{h21} / n_h}$$

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

Using the estimated variance for $\log(\mathrm{OR}_{\mathrm{MH}})$ given by Robins et al. (1986), PROC FREQ computes the corresponding $100(1-\alpha)$ percent confidence limits for the odds ratio as

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

where

$$\begin{aligned}\hat{\sigma}^2 &= var\ [\ln \mathrm{OR}_{\mathrm{MH}}] \\ &= \frac{\sum\limits_h (n_{h11}+n_{h22})(n_{h11}n_{h22})/n_h^2}{2(\sum\limits_h n_{h11}n_{h22}/n_h)^2} \\ &+ \frac{\sum\limits_h [(n_{h11}+n_{h22})(n_{h12}n_{h21}) + (n_{h12}+n_{h21})(n_{h11}n_{h22})]/n_h^2}{2(\sum\limits_h n_{h11}n_{h22}/n_h)(\sum\limits_h n_{h12}n_{h21}/n_h)} \\ &+ \frac{\sum\limits_h (n_{h12}+n_{h21})(n_{h12}n_{h21})/n_h^2}{2(\sum\limits_h n_{h12}n_{h21}/n_h)^2}\end{aligned}$$

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

Odds Ratio (Case-control Studies): Adjusted Logit

The adjusted logit odds ratio estimator (Woolf 1955) is given by

$$\mathrm{OR}_{\mathrm{L}} = \exp\left(\frac{\sum\limits_h w_h \ln \mathrm{OR}_h}{\sum\limits_h w_h}\right)$$

and the corresponding $100(1-\alpha)$ percent confidence limits are

$$\left(\mathrm{OR}_{\mathrm{L}} \cdot \exp\left(-z/\sqrt{\textstyle\sum_h w_h}\right), \mathrm{OR}_{\mathrm{L}} \cdot \exp\left(z/\sqrt{\textstyle\sum_h w_h}\right)\right)$$

where OR_h is the odds ratio for stratum h, and

$$w_h = \frac{1}{var(\ln \mathrm{OR}_h)}$$

Refer to Woolf (1955)

If any cell frequency in a stratum h is zero, then PROC FREQ adds 0.5 to each cell of the stratum before computing OR_h and w_h (Haldane 1955), and displays a warning.

Relative Risks (Cohort Studies)

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

$$\mathrm{RR}_{\text{MH}} = \frac{\sum_h n_{h11} n_{h2\cdot} / n_h}{\sum_h n_{h21} n_{h1\cdot} / n_h}$$

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

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

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

where

$$\begin{aligned}\hat{\sigma}^2 &= v\hat{a}r^2\left[\ln \mathrm{RR}_{\text{MH}}\right] \\ &= \frac{\sum_h \left(n_{h1\cdot} n_{h2\cdot} n_{h\cdot 1} - n_{h11} n_{h21} n_h\right)/n_h^2}{\left(\sum_h n_{h11} n_{h2\cdot}/n_h\right)\left(\sum_h n_{h21} n_{h1\cdot}/n_h\right)}\end{aligned}$$

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

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

and the corresponding $100\left(1-\alpha\right)$ percent confidence limits are

$$\left(\mathrm{RR}_{\text{L}} \cdot \exp\left(-z/\sqrt{\sum_h w_h}\right),\ \mathrm{RR}_{\text{L}} \cdot \exp\left(z/\sqrt{\sum_h w_h}\right)\right)$$

where RR_h is the column 1 relative risk estimator for stratum h, and

$$w_h = \frac{1}{var\left(\ln \mathrm{RR}_h\right)}$$

If n_{h11} or n_{h21} is zero, then PROC FREQ adds 0.5 to each cell of the stratum before computing RR_h and w_h, and displays a warning.

Refer to Kleinbaum, Kupper, and Morgenstern (1982, Sections 17.4, 17.5) and Breslow and Day (1994).

Breslow-Day Test for Homogeneity of the Odds Ratios

When you specify the CMH option, PROC FREQ computes the Breslow-Day test for the stratified analysis of 2×2 tables. It tests the null hypothesis that the odds ratios from 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

$$Q_{\mathrm{BD}} = \frac{\sum_h \left(n_{h11} - \mathrm{E}\left(n_{h11} | \mathrm{OR}_{\mathrm{MH}}\right)\right)^2}{var\left(n_{h11} | \mathrm{OR}_{\mathrm{MH}}\right)}$$

where *E* and *var* denote expected value and variance, respectively. The summation does not include any tables with a zero row or column. If $\mathrm{OR}_{\mathrm{MH}}$ equals zero or if it is undefined, then PROC FREQ does not compute the statistic, and displays 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, and therefore, the Breslow-Day test should never be used as an indicator of validity. △

Refer to Breslow and Day (1993).

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 et al. (1975). Exact computations are based on the statistical theory of exact conditional inference for contingency tables, reviewed by Agresti (1992).

In addition to computation of exact *p*-values, PROC FREQ provides the option of estimating exact *p*-values by Monte Carlo simulation. This can be useful for problems that are so large that exact computations require a great amount of time and memory, but for which asymptotic approximations may not be sufficient.

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 can also compute exact *p*-values for tests of hypotheses that the following statistics are equal to zero: Pearson correlation coefficient, Spearman correlation coefficient, simple kappa coefficient, and weighted kappa coefficient. Additionally, PROC FREQ can compute exact confidence limits for the odds ratio for 2×2 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 limits for the binomial proportion, and an exact test for the binomial proportion value.

If the procedure does not complete the computation within the specified time, use MAXTIME= to increase the amount of clock time that PROC FREQ can use to compute the exact p-values directly or with Monte Carlo estimation.

The following sections summarize the computational algorithms, define the 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 et al. (1984, 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 et al. (1990). For statistics of other forms, PROC FREQ computes an upper limit for the longest path and a lower limit for the shortest path following the approach of Valz and Thompson (1994).

The longest and shortest path distances or limits 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 because it can result in reduced accuracy of the p-values.

PROC FREQ computes exact confidence limits 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 is at least $1 - \alpha$. Thus, these confidence limits 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 limits 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 "Binomial Proportion" on page 543 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 the Pearson chi-square, the likelihood-ratio chi-square, the Mantel-Haenszel chi-square, Fisher's exact test for tables larger than 2×2 tables, McNemar's test, and the one-way 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 zero $(T = 0)$, the alternative of interest may be one-sided $(T < 0, \text{ or } T > 0)$ 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 computes 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 computes 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

$$\begin{aligned} P_1 &= \text{Prob } (\text{Test Statistic } \geq t) \quad &\text{if } t > E_0(T) \\ P_1 &= \text{Prob } (\text{Test Statistic} \leq t) \quad &\text{if } t \leq E_0(T) \end{aligned}$$

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

$$P_2 = \text{Prob } (|\text{Test Statistic} - E_0(T)| \geq |t - E_0(T)|)$$

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. When asymptotic methods may not be sufficient for such large problems, consider using Monte Carlo estimation of exact p-values, as described in "Monte Carlo Estimation" on page 566.

A formula does not exist that can determine in advance how much time or memory that PROC FREQ needs to compute an exact p-value for a certain problem. The time and memory requirements depend on several factors which include the test that is 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, because 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 et al. (1992) and Gail and Mantel (1977).

At any time while PROC FREQ computes exact p-values, you can terminate the computations by pressing the system interrupt key sequence (refer to the *SAS Companion* for your operating environment) and choosing to stop computations. After you terminate exact computations, PROC FREQ completes all other remaining tasks that the procedure specifies. The procedure produces the requested output, reporting 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 clock 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.

Monte Carlo Estimation

If you specify the option MC in the EXACT statement, PROC FREQ computes Monte Carlo estimates of the exact p-values, instead of directly computing the exact p-values. Monte Carlo estimation can be useful for large problems that require a great amount of time and memory for exact computations, but for which asymptotic approximations may not be sufficient. To describe the precision of each Monte Carlo estimate, PROC FREQ provides the asymptotic standard error and $(1 - \alpha) \times 100$ percent confidence limits. The confidence level α is determined by the ALPHA= option in the EXACT statement, which by default equals .01 and produces 99 percent confidence limits. The N= option in the EXACT statement specifies the number of samples that PROC FREQ uses for Monte Carlo estimation, and the default is 10000 samples. You can specify a larger value for N= to improve the precision of the Monte Carlo estimates. Because larger values of N= generate more samples, the computation time increases. Alternatively, you can specify a smaller value of N= to reduce the computation time.

To compute a Monte Carlo estimate of an exact p-value, PROC FREQ generates a random sample of tables with the same total sample size, row totals, and column totals as the observed table. PROC FREQ uses the algorithm of Agresti et al. (1979), which generates tables in proportion to their hypergeometric probabilities, conditional on the marginal frequencies. For each sample table, PROC FREQ computes the value of the test statistic and compares it to the value for the observed table. When estimating a right-sided p-value, PROC FREQ counts all sample tables for which the test statistic is greater than or equal to the observed test statistic. Then the p-value estimate equals the number of these tables divided by the total number of tables sampled.

$$\begin{aligned}
&\hat{P}_{MC} = M/N \\
&M = \text{number of samples with } (\text{Test Statistic} \geq t) \\
&N \ = \text{number of samples} \\
&T = \text{observed Test Statistic}
\end{aligned}$$

PROC FREQ computes left-sided and two-sided p-value estimates similarly. For left-sided p-values, PROC FREQ evaluates whether the test statistic for each sampled table is less than or equal to the observed test statistic. For two-sided p-values, PROC FREQ examines the sample test statistics according to the expression for P_2 given in “Definition of p-Values” on page 565. The variable M above is a binomially distributed variable with N trials and success probability p. It follows that the asymptotic standard error of the Monte Carlo estimate is

$$se\left(\hat{P}_{MC}\right) = \sqrt{\hat{P}_{MC}\left(1 - \hat{P}_{MC}\right) / (N-1)}$$

PROC FREQ constructs asymptotic confidence limits for the p-values according to

$$\hat{P}_{MC} \pm z_{\alpha/2} \cdot se\left(\hat{P}_{MC}\right)$$

where $z_{\alpha/2}$ is the $100\,(1 - \alpha/2)$ percentile of the standard normal distribution, and the confidence level α is determined by the ALPHA= option in the EXACT statement.

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 0, then PROC FREQ computes the confidence limits for the p-value as

$$\left(0, 1 - \alpha^{(1/N)}\right)$$

When the Monte Carlo estimate $\hat{P}_{MC}$ equals 1, then PROC FREQ computes the confidence limits as

$$\left(\alpha^{(1/N)}, 1\right)$$

Results

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, PROC FREQ displays the total frequency of observations with missing values below each table. The following options in the TABLES statement change how 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 keyword in the OUTPUT statement creates a variable in the output data set that contains the number of missing values.

Output 21.3 on page 568 shows three ways that PROC FREQ handles missing values. The first table uses the default method; the second table uses MISSPRINT; and the third table uses MISSING.

Output 21.3 Missing Values in Frequency Tables

```
                         *** Default ***

                       The FREQ Procedure

                                    Cumulative    Cumulative
      A    Frequency     Percent     Frequency      Percent
      ------------------------------------------------------
      1           2       50.00             2        50.00
      2           2       50.00             4       100.00

                     Frequency Missing = 2

                    *** MISSPRINT Option ***

                       The FREQ Procedure

                                    Cumulative    Cumulative
      A    Frequency     Percent     Frequency      Percent
      ------------------------------------------------------
      .           2        .                .          .
      1           2       50.00             2        50.00
      2           2       50.00             4       100.00

                     Frequency Missing = 2

                    *** MISSING Option ***

                       The FREQ Procedure

                                    Cumulative    Cumulative
      A    Frequency     Percent     Frequency      Percent
      ------------------------------------------------------
      .           2       33.33             2        33.33
      1           2       33.33             4        66.67
      2           2       33.33             6       100.00
```

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 with a TABLES statement. To include the missing combinations, use SPARSE with LIST or OUT= 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.

Procedure Output

By default, a one-way table lists the variable name, variable values, frequency counts, percentages, cumulative frequency counts, cumulative percentages, and the number of missing values. Unless you use LIST in the TABLES statement, a two-way table appears as a crosstabulation table. An *n*-way table appears as multiple crosstabulation tables with one table for each combination of values for the stratification variables. By default, each cell of a crosstabulation table lists the frequency count, percentage of the total frequency count, row percentage, and column percentage.

Use the following TABLES statement options to report additional information for each table cell:

CELLCHI2
includes the cell's contribution to the total chi-square statistic

CUMCOL
includes the cumulative column percentage of the cell

DEVIATION
includes the deviation of the cell frequency from the expected value

EXPECTED
includes the expected cell frequency under the hypothesis of independence.

You can also use the SCOROUT option to display the type of score, row score, and column score for two-way tables.

By default, PROC FREQ displays the next one-way frequency table on the current page when there is enough space to display the entire table. If you use COMPRESS in the PROC FREQ statement, the next one-way table starts to display on the current page even when the entire table will not fit. If you use PAGE in the PROC FREQ statement, each frequency or crosstabulation table always displays on a separate page.

Displaying Large Frequencies

By default, PROC FREQ uses the BEST6. format to display a cell frequency when the frequency is less than 1E6. Otherwise, it uses the BEST7. format so that frequency values with more than seven significant digits display in scientific notation (E format). The V5FMT option in the TABLES statement uses BEST8. format so that frequency values with more than eight significant digits display in scientific notation.

When scientific notation is used, only the first few significant digits are shown. If you need more significant digits than PROC FREQ displays, create an output data set by specifying OUT= in the TABLES statement. Then use PROC PRINT and assign an appropriate format to the variable COUNT. For example, the statement

```
format count 10.;
```

displays exact integer counts up to 9999999999. For more information about formats, see the section on components of the SAS language in *SAS Language Reference: Concepts*.

Suppressing the Displayed Output

The NOPRINT option in the PROC FREQ statement and NOPRINT, NOCOL, NOCUM, NOFREQ, NOPERCENT, and NOROW in the TABLES statement suppress displayed output. Use NOPRINT in the PROC FREQ statement to suppress all displayed output as well as the Output Delivery System. Use NOPRINT in the TABLES statement to suppress frequency and crosstabulation tables but still display the requested statistics. Use NOCOL, NOCUM, NOFREQ, NOPERCENT, and NOROW to suppress various frequencies and percentages in the frequency and crosstabulation tables.

CAUTION:

Multiway tables can generate a great deal of displayed output. For example, if the variables A, B, C, D, and E each have ten levels, the table request A*B*C*D*E may generate 1000 or more pages of output. If you are primarily interested in the tests and measures of association, use NOPRINT in the TABLES statement to suppress the tables but display the statistics. Or use NOPRINT in the PROC FREQ statement to suppress all displayed output, and use the OUTPUT statement to store the statistics in an output data set. If you are interested in frequency counts and percentages use LIST in the TABLES statement. △

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:

TABLES statement, OUT= option
creates an output data set that contains frequency or crosstabulation table counts and percentages.

OUTPUT statement
creates an output data set that contains statistics.

PROC FREQ does not display the output data set. Use PROC PRINT, PROC REPORT, or any other SAS reporting tool to display the 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 in the last table request. By default, each observation contains the frequency and percentage for each combination of variable values. When the input data set contains missing values, the output data set 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 variable containing the cell frequency
- PERCENT variable containing the cell percentage.

If you use OUTEXPECT and OUTPCT, the output data set also contains expected frequencies and row, column, and table percentages, respectively. The additional variables are

- EXPECTED variable containing the expected frequency
- PCT_TABL variable containing the percentage of two-way table frequency, for n-way tables where $n > 2$
- PCT_ROW variable containing the percentage of row frequency
- PCT_COL variable containing the percentage of column frequency.

When you submit the following statements

```
proc freq;
   tables a a*b / out=d;
run;
```

the output data set D contains frequencies and percentages for the last table request, A*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 also includes the variables COUNT and PERCENT. The value of COUNT is the number of observations that have 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

```
   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 that contains 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 for these statistics. Additionally, you can output statistics for one-way tables, such as chi-square or binomial proportion statistics. If you use a BY statement, the output data set contains observations for each BY group.

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

□ variables that contain the specified statistics.

The output data set also includes variables with the *p*-value and degrees of freedom, asymptotic standard error (ASE), or confidence limits 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 keywords that are enclosed in underscores. PROC FREQ forms variable names for the corresponding *p*-values, degrees of freedom, or confidence limits by combining the name of the keyword with one of the following prefixes

DF_	degrees of freedom
E_	asymptotic standard error (ASE)
E0_	asymptotic standard error under the null hypothesis
L_	lower confidence limit
P_	*p*-value
P2_	two-sided *p*-value
PL_	left-sided *p*-value
PR_	right-sided *p*-value
U_	upper confidence limit

XP_	exact *p*-value
XP2_	exact two-sided *p*-value
XPR_	exact right-sided *p*-value
XPL_	exact left-sided *p*-value
XL_	exact lower confidence limit
XU_	exact upper confidence limit
Z_	standardized value

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

Examples

Example 1: Creating an Output Data Set with Table Cell Frequencies

Procedure features:

TABLES statement, multiple requests

TABLES statement options:

OUT=
OUTEXPECT
SPARSE

WEIGHT statement

Other features:

PRINT procedure

This example

□ creates two frequency tables and a crosstabulation table using existing cell counts

□ creates an output data set for the last table request with frequencies, percentages, and expected cell frequencies

□ includes zero cell counts in the output data set

□ displays the output data set.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

The data set COLOR contains information on eye and hair color of children from two regions of Europe. The data are recorded as cell counts instead of as one observation per child. Count contains the frequencies of the 15 eye and hair color combinations for each region. Missing eye and hair color combinations are excluded from the data set.

```
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 WEIGHT statement uses Count to weight the observations.

```
proc freq data=color;
   weight count;
```

The TABLES statement requests three tables: Eyes and Hair frequencies and an Eyes by Hair crosstabulation. OUT= creates the FREQCNT data set that contains crosstabulation table frequencies. OUTEXPECT stores expected cell frequencies and SPARSE stores zero cell counts in FREQCNT.

```
   tables eyes hair eyes*hair/out=freqcnt outexpect
                                          sparse;
```

The TITLE statement specifies a title.

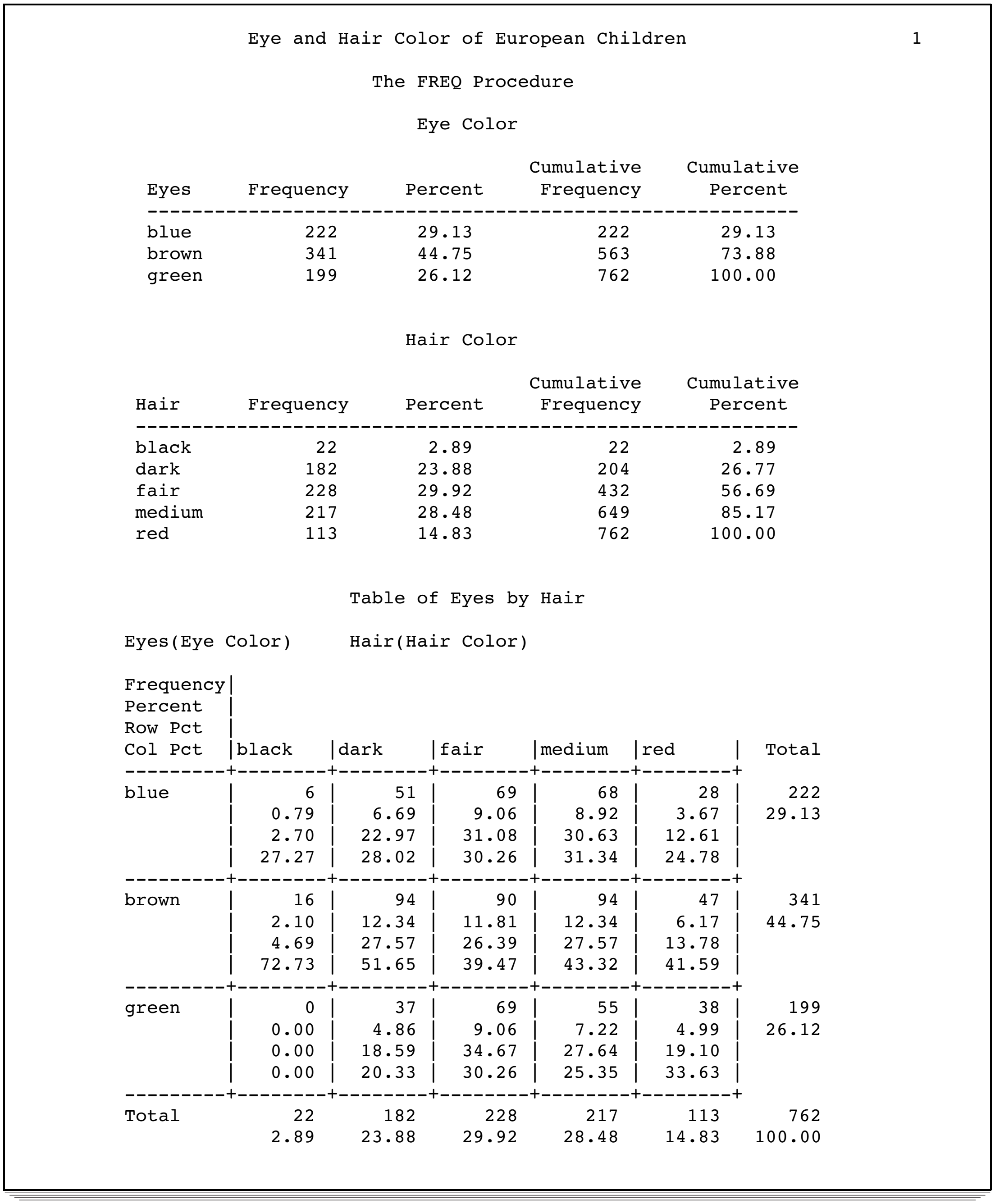
```
   title 'Eye and Hair Color of European Children';
run;
```

PROC PRINT displays the FREQCNT data set. The TITLE statement specifies a title.

```
proc print data=freqcnt noobs;
   title2 'Output Data Set from PROC FREQ';
run;
```

Output

By default, PROC FREQ lists the variable values in alphabetical order. Because Eyes*Hair requests a crosstabulation table, the table rows are eye color and the table columns are hair color. A zero cell count for green eyes and black hair indicates that this eyes and hair combination does not occur in the data.



```
                  Eye and Hair Color of European Children                 1

                             The FREQ Procedure

                                 Eye Color

                                               Cumulative    Cumulative
       Eyes     Frequency     Percent     Frequency      Percent
       ---------------------------------------------------------
       blue          222       29.13           222        29.13
       brown         341       44.75           563        73.88
       green         199       26.12           762       100.00


                                Hair Color

                                               Cumulative    Cumulative
      Hair      Frequency     Percent     Frequency      Percent
      ----------------------------------------------------------
      black           22        2.89            22         2.89
      dark           182       23.88           204        26.77
      fair           228       29.92           432        56.69
      medium         217       28.48           649        85.17
      red            113       14.83           762       100.00


                           Table of Eyes by Hair

     Eyes(Eye Color)     Hair(Hair Color)

     Frequency|
     Percent  |
     Row Pct  |
     Col Pct  |black   |dark    |fair    |medium  |red     |  Total
     ---------+--------+--------+--------+--------+--------+
     blue     |      6 |     51 |     69 |     68 |     28 |    222
              |   0.79 |   6.69 |   9.06 |   8.92 |   3.67 |  29.13
              |   2.70 |  22.97 |  31.08 |  30.63 |  12.61 |
              |  27.27 |  28.02 |  30.26 |  31.34 |  24.78 |
     ---------+--------+--------+--------+--------+--------+
     brown    |     16 |     94 |     90 |     94 |     47 |    341
              |   2.10 |  12.34 |  11.81 |  12.34 |   6.17 |  44.75
              |   4.69 |  27.57 |  26.39 |  27.57 |  13.78 |
              |  72.73 |  51.65 |  39.47 |  43.32 |  41.59 |
     ---------+--------+--------+--------+--------+--------+
     green    |      0 |     37 |     69 |     55 |     38 |    199
              |   0.00 |   4.86 |   9.06 |   7.22 |   4.99 |  26.12
              |   0.00 |  18.59 |  34.67 |  27.64 |  19.10 |
              |   0.00 |  20.33 |  30.26 |  25.35 |  33.63 |
     ---------+--------+--------+--------+--------+--------+
     Total          22      182      228      217      113      762
                  2.89    23.88    29.92    28.48    14.83   100.00
```

The output data set contains frequency counts and percentages for the last table. The data set also includes an observation for the zero cell count and a variable with the expected cell frequency for each table cell.

```
                    Eye and Hair Color of European Children                    2
                         Output Data Set from PROC FREQ

                Eyes     Hair      COUNT    EXPECTED    PERCENT

                blue     black        6       6.409     0.7874
                blue     dark        51      53.024     6.6929
                blue     fair        69      66.425     9.0551
                blue     medium      68      63.220     8.9239
                blue     red         28      32.921     3.6745
                brown    black       16       9.845     2.0997
                brown    dark        94      81.446    12.3360
                brown    fair        90     102.031    11.8110
                brown    medium      94      97.109    12.3360
                brown    red         47      50.568     6.1680
                green    black        0       5.745     0.0000
                green    dark        37      47.530     4.8556
                green    fair        69      59.543     9.0551
                green    medium      55      56.671     7.2178
                green    red         38      29.510     4.9869
```

Example 2: Computing Chi-Square Tests for One-Way Frequency Tables

Procedure features:

PROC FREQ statement option:

ORDER=

BY statement

TABLES statement options:

NOCUM
TESTP=

WEIGHT statement

Other features:

SORT procedure

Data set: COLOR on page 573

This example

- □ sorts a data set by geographic region
- □ creates a one-way frequency table for each BY group
- □ orders the values of the frequency table by their appearance in the input data set
- □ suppresses the cumulative frequencies and percentages
- □ computes a chi-square goodness-of-fit test for specified proportions.

The chi-square goodness-of-fit test examines whether the children's hair color has a specified multinomial distribution for two regions. The hypothesized distribution for hair color is 30 percent fair, 12 percent red, 30 percent medium, 25 percent dark, and 3 percent black.

Program

```
options nodate pageno=1 linesize=80 pagesize=60;
```

PROC SORT sorts the observations by the variable Region.

```
proc sort data=color;
   by region;
run;
```

ORDER=DATA orders the frequency table values (hair color) by their order in the data set. The WEIGHT statement uses Count to weight the observations.

```
proc freq data=color order=data;
   weight count;
```

The TABLES statement requests a frequency table for hair color. NOCUM suppresses the cumulative frequencies and percentages. TESTP= specifies hypothesized percentages for the chi-square test. The number of percentages equals the number of table levels and the percentages sum to 100.

```
   tables hair/nocum testp=(30 12 30 25 3);
```

The BY statement produces a separate table for each BY group and displays a heading above each one.

```
   by region;
```

The TITLE statement specifies a title.

```
   title 'Hair Color of European Children';
run;
```

Output

The frequency table lists the variable values (hair color) in the order that they appear in the data set. The last 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 .05 level for region 2 (**p**≤.05) but not for region 1, indicating a significant departure from the hypothesized percentages in region 2.

```
                        Hair Color of European Children                         1

----------------------------- Geographic Region=1 ------------------------------

                               The FREQ Procedure

                                  Hair Color

                                                        Test
                  Hair      Frequency     Percent     Percent
                  -------------------------------------------
                  fair            76       30.89       30.00
                  red             19        7.72       12.00
                  medium          83       33.74       30.00
                  dark            65       26.42       25.00
                  black            3        1.22        3.00


                                Chi-Square Test
                           for Specified Proportions
                           -------------------------
                           Chi-Square         7.7602
                           DF                      4
                           Pr > ChiSq         0.1008
```

```
                        Hair Color of European Children                         2

----------------------------- Geographic Region=2 ------------------------------

                               The FREQ Procedure

                                  Hair Color

                                                        Test
                  Hair      Frequency     Percent     Percent
                  -------------------------------------------
                  fair           152       29.46       30.00
                  red             94       18.22       12.00
                  medium         134       25.97       30.00
                  dark           117       22.67       25.00
                  black           19        3.68        3.00


                                Chi-Square Test
                           for Specified Proportions
                           -------------------------
                           Chi-Square        21.3824
                           DF                      4
                           Pr > ChiSq         0.0003
```

Example 3: Computing Binomial Proportions for One-Way Frequency Tables

Procedure features:

PROC FREQ statement option:

ORDER=

TABLES statement options:

ALPHA=
BINOMIAL

WEIGHT statement

Data set: COLOR on page 573

This example

- creates a one-way frequency tables using existing cell counts
- orders the values of the frequency table by their frequency in the input data set
- computes the binomial proportion and the corresponding test statistic
- specifies the null hypothesis proportion for the asymptotic test of the binomial proportion
- specifies the confidence level for the confidence limits.

Program

```
options nodate pageno=1 linesize=80 pagesize=40;
```

ORDER=FREQ orders the frequency table values by their frequency in the data set. The WEIGHT statement uses Count to weight the observations.

```
proc freq data=color order=freq;
   weight count;
```

The TABLES statement requests a frequency table for eye color. BINOMIAL computes the binomial proportion and confidence limits, and also tests the hypothesis that the proportion for the first eye color level equals 0.5. ALPHA= specifies 90 percent confidence limits.

```
   tables eyes/binomial alpha=.1;
```

The TABLES statement requests a frequency table for hair color. BINOMIAL computes the binomial proportion and confidence limits, and also tests the hypothesis that the proportion for the first hair color level equals 0.28.

```
   tables hair/binomial(p=.28);
```

The TITLE statement specifies a title.

```
   title 'Hair and Eye Color of European Children';
run;
```

Output

The frequency table lists the variable values in the order of the descending frequency count. PROC FREQ computes the binomial proportion for the first variable level. The report includes the asymptotic standard error (ASE), and asymptotic and exact confidence limits for the binomial proportion. The specified confidence level of .1 results in 90 percent confidence limits.

Because the value of **Z** is less than zero for eye color, PROC FREQ computes a left-sided **p**–value. The small **p**–value supports the alternative hypothesis that the true value of the proportion of children with brown eyes is less than 50 percent.

```
                   Hair and Eye Color of European Children                   1

                              The FREQ Procedure

                                  Eye Color

                                            Cumulative    Cumulative
          Eyes     Frequency     Percent     Frequency      Percent
          ----------------------------------------------------------
          brown         341       44.75           341        44.75
          blue          222       29.13           563        73.88
          green         199       26.12           762       100.00


                             Binomial Proportion
                               for Eyes = brown
                       --------------------------------
                       Proportion                0.4475
                       ASE                       0.0180
                       90% Lower Conf Limit      0.4179
                       90% Upper Conf Limit      0.4771

                       Exact Conf Limits
                       90% Lower Conf Limit      0.4174
                       90% Upper Conf Limit      0.4779

                         Test of H0: Proportion = 0.5

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

Because the value of **Z** is greater than zero for hair color, PROC FREQ computes a right-sided **p**–value. The large **p**–value provides insufficient evidence to reject the null hypothesis that the proportion of children with fair hair equals 28 percent.

```
          Hair and Eye Color of European Children                    2

                     The FREQ Procedure

                        Hair Color

                                   Cumulative    Cumulative
Hair      Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
fair           228       29.92           228        29.92
medium         217       28.48           445        58.40
dark           182       23.88           627        82.28
red            113       14.83           740        97.11
black           22        2.89           762       100.00


                    Binomial Proportion
                      for Hair = fair
              --------------------------------
              Proportion                0.2992
              ASE                       0.0166
              95% Lower Conf Limit      0.2667
              95% Upper Conf Limit      0.3317

              Exact Conf Limits
              95% Lower Conf Limit      0.2669
              95% Upper Conf Limit      0.3331

               Test of H0: Proportion = 0.28

              ASE under H0              0.0163
              Z                         1.1812
              One-sided Pr >  Z         0.1188
              Two-sided Pr > |Z|        0.2375
```

Example 4: Analyzing a 2×2 Contingency Table

Procedure features:

PROC FREQ statement option:

ORDER=

EXACT statement

TABLES statement options:

CHISQ
RELRISK

WEIGHT statement

Other features:

FORMAT procedure

SORT procedure

This example

- □ creates a two-way contingency table using existing cell counts
- □ sorts the data in descending order so that the first table cell contains the frequency of positive exposure and positive response
- □ computes chi-square tests, exact Pearson chi-square test, and Fisher's exact test to compare the probability of coronary heart disease for two types of diet
- □ computes estimates of the relative risk and 95 percent exact confidence limits for the odds ratio.

Program

```
options nodate pageno=1 linesize=84 pagesize=64;
```

PROC FORMAT creates user-written 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;
```

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 instead of as one observation per subject. The variable Count contains the frequencies for each exposure and response combination.

```
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 sorts the observations in descending order by the variables Exposure and Response.

```
proc sort data=fatcomp;
   by descending exposure descending response;
run;
```

ORDER=DATA orders the contingency table values by their order in the data set. The WEIGHT statement uses Count to weight the observations.

```
proc freq data=fatcomp order=data;
   weight count;
```

The TABLES statement requests a two-way table. CHISQ requests chi-square tests. RELRISK requests relative risk measures.

```
    tables exposure*response / chisq relrisk;
```

The EXACT statement requests the exact Pearson chi-square test and exact confidence limits for the odds ratio.

```
    exact pchi or;
```

The FORMAT statement assigns formats to the variables Exposure and Response. The TITLE statement specifies a title.

```
    format exposure expfmt. response rspfmt.;
    title 'Case-Control Study of High Fat/Cholesterol Diet';
 run;
```

Output

The contingency table lists the variable values so that the first table cell contains the frequency of positive exposure and response. PROC FREQ does not truncate the formatted variable values that are more than 16 characters but uses multiple lines to show Exposure levels.

PROC FREQ displays a warning message that sample size requirements may not be met for the asymptotic chi-square tests. The exact tests are appropriate when sample size is small.

Because the alternative hypothesis for this analysis states that coronary heart disease was more likely to be associated with a high-fat diet, a one-sided test is needed. Fisher's exact test (right-sided) tests that the probability of heart disease in the high-fat group exceeds the probability of heart disease in the low-fat group.

The odds ratio, which provides an estimate of the relative risk when an event is rare, indicates that the odds of heart disease are 8.25 times higher in the high fat diet group. However, the wide confidence limits indicate that this estimate has low precision.

```
                 Case-Control Study of High Fat/Cholesterol Diet                1

                              The FREQ Procedure

                         Table of Exposure by Response

                   Exposure          Response(Heart Disease)

                   Frequency        |
                   Percent          |
                   Row Pct          |
                   Col Pct          |Yes     |No      |  Total
                   -----------------+--------+--------+
                   High Cholesterol |     11 |      4 |     15
                    Diet            |  47.83 |  17.39 |  65.22
                                    |  73.33 |  26.67 |
                                    |  84.62 |  40.00 |
                   -----------------+--------+--------+
                   Low Cholesterol  |      2 |      6 |      8
                    Diet            |   8.70 |  26.09 |  34.78
                                    |  25.00 |  75.00 |
                                    |  15.38 |  60.00 |
                   -----------------+--------+--------+
                   Total                  13       10       23
                                       56.52    43.48   100.00


                   Statistics for Table of Exposure by Response

              Statistic                     DF       Value      Prob
              ------------------------------------------------------
              Chi-Square                     1      4.9597    0.0259
              Likelihood Ratio Chi-Square    1      5.0975    0.0240
              Continuity Adj. Chi-Square     1      3.1879    0.0742
              Mantel-Haenszel Chi-Square     1      4.7441    0.0294
              Phi Coefficient                       0.4644
              Contingency Coefficient               0.4212
              Cramer's V                            0.4644

           WARNING: 50% of the cells have expected counts less than 5.
                    (Asymptotic) Chi-Square may not be a valid test.
                   Case-Control Study of High Fat/Cholesterol Diet                  2

                                 The FREQ Procedure


                              Pearson Chi-Square Test
                          ----------------------------------
                          Chi-Square                  4.9597
                          DF                               1
                          Asymptotic Pr >  ChiSq      0.0259
                          Exact      Pr >= ChiSq      0.0393


                                Fisher's Exact Test
                          ----------------------------------
                          Cell (1,1) Frequency (F)        11
                          Left-sided Pr <= F          0.9967
                          Right-sided Pr >= F         0.0367

                          Table Probability (P)       0.0334
                          Two-sided Pr <= P           0.0393

                     Statistics for Table of Exposure by Response

                     Estimates of the Relative Risk (Row1/Row2)

          Type of Study                   Value       95% Confidence Limits
          -----------------------------------------------------------------
          Case-Control (Odds Ratio)      8.2500        1.1535       59.0029
          Cohort (Col1 Risk)             2.9333        0.8502       10.1204
          Cohort (Col2 Risk)             0.3556        0.1403        0.9009


                           Odds Ratio (Case-Control Study)
                         -----------------------------------
                         Odds Ratio                   8.2500

                         Asymptotic Conf Limits
                         95% Lower Conf Limit         1.1535
                         95% Upper Conf Limit        59.0029

                         Exact Conf Limits
                         95% Lower Conf Limit         0.8677
                         95% Upper Conf Limit       105.5488

                                  Sample Size = 23
```

Example 5: Creating an Output Data Set Containing Chi-Square Statistics

Procedure features:

PROC FREQ statement option:

ORDER=

OUTPUT statement options:

OUT=

statistic-keywords

TABLES statement options:

CHISQ

DEVIATION

EXPECTED

NOCOL
NOROW

WEIGHT statement

Other features:

PRINT procedure

Data set: COLOR on page 573

This example

- creates a 3×5 contingency table showing the joint frequency distribution for two variables
- suppresses the row and column percentages for each cell
- displays the expected frequency for each cell
- displays each cell's contribution to the total Pearson chi-square statistic
- creates an output data set with Pearson chi-square and likelihood-ratio chi-square statistics
- displays the output data set.

Program

ORDER=DATA orders the table values (eye and hair color) by their order in the data set. The WEIGHT statement uses Count to weight the observations.

```
options nodate pageno=1 pagesize=60;

proc freq data=color order=data;
   weight count;
```

The TABLES statement requests a two-way table. CHISQ requests chi-square tests. EXPECTED displays the expected cell frequency, and CELLCHI2 displays the cell contribution to chi-square. NOROW and NOCOL suppress the row and column percents for each cell.

```
   tables eyes*hair /chisq expected cellchi2
                     norow nocol;
```

The OUTPUT statement creates the CHISQDAT data set with eight variables. N stores the number of nonmissing observations, NMISS stores the number of missing observations, PCHI stores Pearson chi-square statistics, and LRCHI stores likelihood-ratio chi-square statistics. The TITLE statement specifies a title.

```
   output out=chisqdat pchi lrchi n nmiss;
   title 'Chi-Square Tests for 3 by 5 Table of Eye and Hair Color';
run;
```

PROC PRINT displays the CHISQDAT data set. The TITLE statement specifies a title.

```
proc print data=chisqdat noobs;
   title 'Chi-Square Statistics for Eye and Hair Color';
   title2 'Output Data Set from the FREQ Procedure';
run;
```

Output

The contingency table lists eye and hair color in the order that they appear in the data set. The first column label explains the contents of each table cell. The Pearson chi-square provides evidence of an association between eye and hair color (p=.007). 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. Exactly the opposite occurs with the brown-eyed children.

```
                     Chi-Square Tests for 3 by 5 Table of Eye and Hair Color                     1

                                        The FREQ Procedure

                                      Table of Eyes by Hair

                   Eyes(Eye Color)     Hair(Hair Color)

                   Frequency      |
                   Expected       |
                   Cell Chi-Square|
                   Percent        |fair    |red     |medium  |dark    |black   |  Total
                   ---------------+--------+--------+--------+--------+--------+
                   blue           |     69 |     28 |     68 |     51 |      6 |    222
                                  | 66.425 | 32.921 |  63.22 | 53.024 | 6.4094 |
                                  | 0.0998 | 0.7357 | 0.3613 | 0.0772 | 0.0262 |
                                  |   9.06 |   3.67 |   8.92 |   6.69 |   0.79 |  29.13
                   ---------------+--------+--------+--------+--------+--------+
                   green          |     69 |     38 |     55 |     37 |      0 |    199
                                  | 59.543 |  29.51 | 56.671 |  47.53 | 5.7454 |
                                  | 1.5019 | 2.4422 | 0.0492 | 2.3329 | 5.7454 |
                                  |   9.06 |   4.99 |   7.22 |   4.86 |   0.00 |  26.12
                   ---------------+--------+--------+--------+--------+--------+
                   brown          |     90 |     47 |     94 |     94 |     16 |    341
                                  | 102.03 | 50.568 | 97.109 | 81.446 | 9.8451 |
                                  | 1.4187 | 0.2518 | 0.0995 |  1.935 | 3.8478 |
                                  |  11.81 |   6.17 |  12.34 |  12.34 |   2.10 |  44.75
                   ---------------+--------+--------+--------+--------+--------+
                   Total               228      113      217      182       22      762
                                     29.92    14.83    28.48    23.88     2.89   100.00


                                  Statistics for Table of Eyes by Hair

                          Statistic                     DF       Value      Prob
                          ------------------------------------------------------
                          Chi-Square                     8     20.9248    0.0073
                          Likelihood Ratio Chi-Square    8     25.9733    0.0011
                          Mantel-Haenszel Chi-Square     1      3.7838    0.0518
                          Phi Coefficient                       0.1657
                          Contingency Coefficient               0.1635
                          Cramer's V                            0.1172

                                         Sample Size = 762
```

The output data set has one observation that contains the sample size, number of missing observations, and chi-square statistics with the corresponding degrees of freedom and probability values.

```
                    Chi-Square Statistics for Eye and Hair Color                  2
                     Output Data Set from the FREQ Procedure

   N    NMISS    _PCHI_   DF_PCHI       P_PCHI    _LRCHI_   DF_LRCHI     P_LRCHI

  762     0     20.9248      8      .007349898   25.9733       8      .001061424
```

Example 6: Computing Cochran-Mantel-Haenszel Statistics for a Stratified Table

Procedure features:

TABLES statement options:

CMH
NOPRINT

WEIGHT statement

This example

□ creates stratified two-way contingency tables using existing cell counts

□ suppresses the display of the contingency tables

□ computes Cochran-Mantel-Haenszel statistics adjusting for the effects of a stratification variable.

Program

The data set MIGRAINE contains hypothetical data for a clinical trial of migraine treatment. Subjects of both genders either receive new drug therapy or a placebo. Their response to treatment is coded as better or the same. The data are recorded as cell counts instead of as one observation per subject. The variable Frequency contains the frequencies for each treatment and response combination.

```
options nodate pageno=1 linesize=80 pagesize=60;

data migraine;
   input Gender $ Treatment $ Improve $ Frequency @@;
   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 WEIGHT statement uses Frequency to weight the observations.

```
proc freq data=migraine;
   weight frequency;
```

The TABLES statement requests a three-way table stratified by Gender where Treatment forms the rows and Improve forms the columns. CMH requests the Cochran-Mantel-Haenszel statistics. NOPRINT suppresses the display of contingency tables.

```
   tables gender*treatment*improve/cmh noprint;
```

The TITLE statement specifies a title.

```
   title1 'Clinical Trial for Treatment of Migraine Headaches';
run;
```

Output

PROC FREQ computes Cochran-Mantel-Haenszel statistics, controlling for Gender. For stratified 2×2 contingency tables, these statistics include estimates of the common relative risk and the Breslow-Day test for homogeneity of the odds ratios. For a stratified 2×2 table, the three CMH statistics test the same hypothesis. The significant **p**-value (.004) indicates that the association between treatment and response remains strong after adjusting for gender.

The large **p**-value for the Breslow-Day test (.222) indicates no significant gender difference in the odds ratios. Because this is a prospective study, the relative risk estimate assesses the effectiveness of the new drug. The probability of migraine improvement with the new drug is just over two times the probability of improvement with the placebo.

```
            Clinical Trial for Treatment of Migraine Headaches              1

                            The FREQ Procedure

               Summary Statistics for Treatment by Improve
                          Controlling for Gender

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

     Statistic    Alternative Hypothesis    DF       Value      Prob
     ---------------------------------------------------------------
         1        Nonzero Correlation        1      8.3052    0.0040
         2        Row Mean Scores Differ     1      8.3052    0.0040
         3        General Association        1      8.3052    0.0040


            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                    Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel          3.3132       1.4456       7.5934
  (Odds Ratio)    Logit                    3.2941       1.4182       7.6515

Cohort            Mantel-Haenszel          2.1636       1.2336       3.7948
  (Col1 Risk)     Logit                    2.1059       1.1951       3.7108

Cohort            Mantel-Haenszel          0.6420       0.4705       0.8761
  (Col2 Risk)     Logit                    0.6613       0.4852       0.9013


                           Breslow-Day Test for
                      Homogeneity of the Odds Ratios
                      ------------------------------
                      Chi-Square              1.4929
                      DF                           1
                      Pr > ChiSq              0.2218


                         Total Sample Size = 106
```

Example 7: Computing the Cochran-Armitage Trend Test

Procedure features:

- EXACT statement options:
  - *statistic-keywords*
  - MAXTIME=
- TABLES statement options:
  - CL
  - MEASURES
  - TREND
- TEST statement
- WEIGHT statement

This example

- creates a two-way table using existing cell counts
- computes measures of association and asymptotic 95% confidence limits
- computes asymptotic and exact p-values for the Cochran-Armitage trend test
- specifies the maximum time to compute an exact p-value
- computes asymptotic tests for Somers' $D(C\,|\,R)$.

The Cochran-Armitage test checks for trend in binomial proportions across levels of a single factor. Use this test for a contingency table with a two-level response variable and an explanatory variable with any number of ordered levels. The binomial proportion is defined as the proportion in the first level of the response variable. PROC FREQ uses explanatory variable scores to compute the Cochran-Armitage test, which you can set to meaningful values that reflect the degree of difference among the levels.

Program

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 coded No or Yes. The data are recorded as cell counts instead of as one observation per subject. The variable Count contains the frequencies for each drug dose and response combination.

```
options nodate pageno=1 linesize=80 pagesize=72;

data pain;
   input Dose Adverse $ Count @@;
   cards;
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 WEIGHT statement uses Count to weight the observations.

```
proc freq data=pain;
   weight count;
```

The TABLES statement requests a two-way table. TREND requests the Cochran-Armitage trend test. MEASURES requests measures of associations and CL computes confidence limits.

```
   tables dose*adverse /trend measures cl;
```

The TEST statement computes an asymptotic test for Somers' **D**(**C** | **R**).

```
   test smdcr;
```

The EXACT statement requests exact trend test. MAXTIME= specifies that PROC FREQ terminate the computations after 60 seconds (1 minute).

```
   exact trend /maxtime=60;
```

The TITLE statement specifies a title.

```
   title1 'Clinical Trial for Treatment of Pain';
run;
```

Output

The Row Pct values show the expected increasing trend in the proportion of adverse effects (from 18.75% to 71.88%).

```
                       Clinical Trial for Treatment of Pain                    1

                               The FREQ Procedure

                            Table of Dose by Adverse

                       Dose      Adverse

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


                     Statistics for Table of Dose by Adverse

                                                                      95%
   Statistic                              Value       ASE     Confidence Limits
   ----------------------------------------------------------------------------
   Gamma                                 0.5313    0.0935      0.3480      0.7146
   Kendall's Tau-b                       0.3373    0.0642      0.2114      0.4631
   Stuart's Tau-c                        0.4111    0.0798      0.2547      0.5675

   Somers' D C|R                         0.2569    0.0499      0.1592      0.3547
   Somers' D R|C                         0.4427    0.0837      0.2786      0.6068

   Pearson Correlation                   0.3776    0.0714      0.2378      0.5175
   Spearman Correlation                  0.3771    0.0718      0.2363      0.5178

   Lambda Asymmetric C|R                 0.2373    0.0837      0.0732      0.4014
   Lambda Asymmetric R|C                 0.1250    0.0662      0.0000      0.2547
   Lambda Symmetric                      0.1604    0.0621      0.0388      0.2821

   Uncertainty Coefficient C|R           0.1261    0.0467      0.0346      0.2175
   Uncertainty Coefficient R|C           0.0515    0.0191      0.0140      0.0890
   Uncertainty Coefficient Symmetric     0.0731    0.0271      0.0199      0.1262
```

Somers' **D** (**C|R**)measures the association. The column variable (Adverse) is the response and the row variable (Dose) is a predictor. Because the asymptotic 95% confidence limit does not contain zero, this indicates a strong positive association. Similarly, Pearson and Spearman correlation coefficients show evidence of a strong positive association as hypothesized.

The Cochran-Armitage test supports the trend hypothesis. The small left-sided **p**-values indicate that the probability of the Column 1 level (Adverse=No) decreases as Dose increases, or equivalently, that the probability of the Column 2 level (Adverse=Yes) increases as Dose increases. The two-sided **p**-value tests against either the increasing or the 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.

```
                     Clinical Trial for Treatment of Pain                     2

                              The FREQ Procedure

                    Statistics for Table of Dose by Adverse

                                Somers' D C|R
                      --------------------------------
                      Somers' D C|R             0.2569
                      ASE                       0.0499
                      95% Lower Conf Limit      0.1592
                      95% Upper Conf Limit      0.3547

                       Test of H0: Somers' D C|R = 0

                      ASE under H0              0.0499
                      Z                         5.1511
                      One-sided Pr >  Z         <.0001
                      Two-sided Pr > |Z|        <.0001


                        Cochran-Armitage Trend Test
                      -------------------------------
                      Statistic (Z)           -4.7918

                      Asymptotic Test
                      One-sided Pr <  Z        <.0001
                      Two-sided Pr > |Z|       <.0001

                      Exact Test
                      One-sided Pr <=  Z    7.237E-07
                      Two-sided Pr >= |Z|   1.324E-06

                             Sample Size = 161
```

Example 8: Computing Friedman's Chi-Square Statistic

Procedure features:

TABLES statement, multiple requests

TABLES statement options:

CMH2
NOPRINT
SCORES=
SCOROUT

This example

- □ computes the first two Cochran-Mantel-Haenszel statistics
- □ uses rank scores to compute the Cochran-Mantel-Haenszel statistics
- □ suppresses the display of contingency tables for each stratum.

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 they are 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.

Program

The data set HYPNOSIS contains data for a study investigating whether hypnosis has the same effect on skin potential (measured in millivolts) for four emotions. (Lehmann 1975, 264). Eight subjects are asked to display fear, happiness (joy), depression (sadness), and calmness under hypnosis. The data are recorded as one observation per subject for each emotion.

```
options nodate pageno=1 linesize=80;

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
;
```

The TABLES statement requests a three-way table stratified by Subject and a two-way table. Emotion and SkinResponse form the rows and columns of each table. CMH2 requests the first two Cochran-Mantel-Haenszel statistics. SCORES=RANK uses rank scores to compute these statistics. NOPRINT suppresses the display of contingency tables.

```
proc freq data=hypnosis;
   tables subject*emotion*skinresponse emotion*skinresponse
```

```
          /cmh2 scores=rank noprint;
```

The TITLE statement specifies a title.

```
   title1 'Examining the Effect of Hypnosis on Skin Potential';
run;
```

Output

PROC FREQ computes Cochran-Mantel-Haenszel statistics across strata controlling for Subject. Because CMH statistics 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 .09 indicates that differences in skin potential response for different emotions are significant at the 10% level but not at the 5% level.

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**=.904). Thus, adjusting for subject is critical to reducing the background variation due to subject differences.

```
                Examining the Effect of Hypnosis on Skin Potential             1

                               The FREQ Procedure

                  Summary Statistics for Emotion by SkinResponse
                             Controlling for Subject

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

         Statistic    Alternative Hypothesis    DF       Value      Prob
         ---------------------------------------------------------------
             1        Nonzero Correlation        1      0.2400    0.6242
             2        Row Mean Scores Differ     3      6.4500    0.0917


                             Total Sample Size = 32
```

```
                Examining the Effect of Hypnosis on Skin Potential             2

                               The FREQ Procedure

                  Summary Statistics for Emotion by SkinResponse

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

         Statistic    Alternative Hypothesis    DF       Value      Prob
         ---------------------------------------------------------------
             1        Nonzero Correlation        1      0.0001    0.9933
             2        Row Mean Scores Differ     3      0.5678    0.9038


                             Total Sample Size = 32
```

Example 9: Testing Marginal Homogeneity with Cochran's *Q*

Procedure features:

TABLES statement, multiple requests

TABLES statement options:

AGREE
NOCUM
NOPRINT

WEIGHT statement

Other features:

FORMAT procedure

This example

- creates frequency tables for the analysis variables using existing cell counts
- computes tests and measures of agreement, which include Cochran's *Q* statistic for stratified 2×2 contingency tables
- suppresses the cumulative frequencies and cumulative percentages
- suppresses the display of contingency tables.

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 PROC CATMOD in SAS/STAT software to fit a repeated-measures model. Data for this example are from *Categorical Data Analysis* by Alan Agresti. Copyright © 1990. Reprinted by permission of John Wiley and Sons, Inc.

Program

PROC FORMAT creates a user-written format to identify the response to treatment.

```
options nodate pageno=1 linesize=80 pagesize=60;

proc format;
   value $responsefmt 'F'='Favorable'
                      'U'='Unfavorable';
run;
```

The data set DRUGS contains data for a study of three drugs to treat a chronic condition (Agresti, 1990). Forty-six subjects receive drugs A, B, and C. The response to each is coded as favorable (**F**) or unfavorable (**U**). The data are recorded as cell counts instead of as one observation per patient. The variable Count contains the cell count.

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

```
F U U 4   U F F  2   U F U 4
U U F 6   U U U  6
;
```

The WEIGHT statement uses Count to weight the observations.

```
proc freq data=drugs;
   weight count;
```

The TABLES statement requests frequency tables of Drug_A, Drug_B, and Drug_C. NOCUM suppresses the cumulative values.

```
   tables drug_a drug_b drug_c/nocum;
```

The TABLES statement requests a three-way table of Drug_A, Drug_B, and Drug_C. AGREE requests measures of agreement. NOPRINT suppresses the display of contingency tables.

```
   tables drug_a*drug_b*drug_c/agree noprint;
```

The FORMAT statement assigns formats to the levels of Drug_A, Drug_B, and Drug_C. The TITLE statement specifies a title.

```
   format drug_a drug_b drug_c $responsefmt.;
   title 'Study of Three Drug Treatments for a Chronic Disease';
run;
```

Output

The one-way frequency tables provides 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 for drug C.

```
Study of Three Drug Treatments for a Chronic Disease             1

                 The FREQ Procedure

        Drug_A         Frequency     Percent
        ------------------------------------
        Favorable            28       60.87
        Unfavorable          18       39.13


        Drug_B         Frequency     Percent
        ------------------------------------
        Favorable            28       60.87
        Unfavorable          18       39.13


        Drug_C         Frequency     Percent
        ------------------------------------
        Favorable            16       34.78
        Unfavorable          30       65.22
```

McNemar’s test shows strong discordance between drugs B and C when the response to drug A is favorable. A small negative value of simple kappa indicates no agreement between the drug B response and the drug C response.

```
Study of Three Drug Treatments for a Chronic Disease             2

                 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 Limit     -0.2615
          95% Upper Conf Limit      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 Limit     -0.5909
          95% Upper Conf Limit      0.2832

                  Sample Size = 18
```

In this example, the hypothesis of interest is whether the response to treatment is equal for the three drugs. Cochran's **Q** is statistically significant (**p**=.014), which leads to rejection of the null hypothesis that the probability of favorable response is the same for the three drugs

```
Study of Three Drug Treatments for a Chronic Disease             3

                 The FREQ Procedure

      Summary Statistics for Drug_B by Drug_C
               Controlling for Drug_A


             Overall Kappa Coefficient
          --------------------------------
          Kappa                    -0.0588
          ASE                       0.1034
          95% Lower Conf Limit     -0.2615
          95% Upper Conf Limit      0.1439


                Test for Equal Kappa
                    Coefficients
                --------------------
                Chi-Square    0.2314
                DF                 1
                Pr > ChiSq    0.6305


              Cochran's Q, for Drug_A
                by Drug_B by Drug_C
              -----------------------
              Statistic (Q)    8.4706
              DF                    2
              Pr > Q           0.0145


               Total Sample Size = 46
```

References

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7(1), 131–177.

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: John Wiley and Sons, Inc.

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley and Sons, Inc.

Agresti, A., Mehta, C.R. and Patel, N.R. (1990), "Exact Inference for Contingency Tables with Ordered Categories," *Journal of the American Statistical Association*, 85, 453–458.

Agresti, A., Wackerly, D. and Boyett, J.M. (1979), " Exact Conditional Tests for Cross-Classifications: Approximation of Attained Significance Levels," *Psychometrika*, 44, 75-83.

Birch, M.W. (1965), "The Detection of Partial Association, II: The General Case," *Journal of the Royal Statistical Society, B*, 27, 111–124.

Bishop, Y., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Bowker, A.H. (1948), "Bowker's Test for Symmetry," *Journal of the American Statistical Association*, 43, 572–574.

Breslow, N.E. and Day, N.E. (1993), *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies*, IARC Scientific Publications, No. 32, New York: Oxford University Press, Inc.

Breslow, N.E. and Day, N.E. (1994), *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies*, IARC Scientific Publications, No. 82, New York: Oxford University Press, Inc.

Bross, I.D.J. (1958), "How to Use Ridit Analysis," *Biometrics*, 14, 18–38.

Brown, M.B. and Benedetti, J.K. (1977), "Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables," *Journal of the American Statistical Association* 72, 309–315.

Cicchetti, D.V. and Allison, T. (1971), "A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings," *American Journal of EEG Technology*, 11, 101–109.

Cochran, W.G. (1950), "The Comparison of Percentages in Matched Samples," *Biometrika*, 37, 256–266.

Cochran, W.G. (1954), "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451.

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37–46.

Drasgow, F. (1986), "Polychoric and Polyserial Correlations" in *Encyclopedia of Statistical Sciences, Volume 7*, eds. S. Kotz and N. L. Johnson, New York: John Wiley and Sons, Inc., 68–74.

Fienberg, S.E. (1980), *The Analysis of Cross-Classified Data,* 2nd Edition, Cambridge, MA: MIT Press.

Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions*, 2nd Edition, New York: John Wiley and Sons, Inc.

Fleiss, J.L. and Cohen, J. (1973), " The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and Psychological Measurement*, 33, 613–619.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969), "Large-Sample Standard Errors of Kappa and Weighted Kappa," *Psychological Bulletin*, 72, 323–327.

Freeman, G.H. and Halton, J.H. (1951), "Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance," *Biometrika*, 38, 141–149.

Gail, M. and Mantel, N. (1977), "Counting the Number of $r \times c$ Contingency Tables with Fixed Margins," *Journal of the American Statistical Association*, 72, 859-862.

Gart, J.J. (1971), "The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification," *Review of the International Statistical Institute*, 39(2), 148–169.

Goodman, L.A. and Kruskal, W.H. (1954, 1959, 1963, 1972), "Measures of Association for Cross-Classification I, II, III, and IV," *Journal of the American Statistical Association*, 49, 732–764; 54, 123–163; 58, 310–364; 67, 415–421.

Greenland, S. and Robins, J.M. (1985), "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, 42, 311-323.

Haldane, J.B.S. (1955), "The Estimation and Significance of the Logarithm of a Ratio of Frequencies," *Annals of Human Genetics*, 20, 309–314.

Hollander, M. and Wolfe, D.A. (1973), *Nonparametric Statistical Methods*, New York: John Wiley and Sons, Inc.

Kendall, M. (1955), *Rank Correlation Methods*, 2nd Edition, London: Charles Griffin and Co.

Kendall, M. and Stuart, A. (1979), *The Advanced Theory of Statistics, Volume 2*, New York: Macmillan Publishing Company, Inc.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982), *Epidemiologic Research: Principles and Quantitative Methods*, Research Methods Series, New York: Van Nostrand Reinhold.

Landis, R.J., Heyman, E.R., and Koch, G.G. (1978), "Average Partial Association in Three-way Contingency Tables: A Review and Discussion of Alternative Tests," *International Statistical Review*, 46, 237–254.

Leemis, L.M. and Trivedi, K.S. (1996), "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," *The American Statistician*, 50(1), 63–68.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day, Inc.

Liebetrau, A.M. (1983), *Measures of Association*, Quantitative Application in the Social Sciences, Vol. 32, Beverly Hills: Sage Publications, Inc.

Mack, G.A. and Skillings, J.H. (1980), "A Friedman-Type Rank Test for Main Effects in a Two-Factor ANOVA," *Journal of the American Statistical Association*, 75, 947–951.

Mantel, N. (1963), "Chi-square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure," *Journal of the American Statistical Association*, 58, 690–700.

Mantel, N. and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719–748.

Margolin, B.H. (1988), "Test for Trend in Proportions," *Johnson's Encyclopedia of Statistics, Volume 9*, eds. S. Kotz and N.L. Johnson, New York: John Wiley and Sons, Inc., 334–336.

McNemar, Q. (1947), "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, 12, 153–157.

Mehta, C.R. and Patel, N.R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in $r{\times}c$ Contingency Tables," *Journal of the American Statistical Association*, 78, 427–434.

Mehta, C.R., Patel, N.R., and Senchaudhuri, P. (1991), "Exact Stratified Linear Rank Tests for Binary Data," *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, ed. E.M. Keramidas, 200–207.

Mehta, C.R., Patel, N.R., and Tsiatis, A.A. (1984), "Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data," *Biometrics*, 40, 819–825.

Narayanan, A. and Watts, D. (1996), "Exact Methods in the NPAR1WAY Procedure," in *Proceedings of the Twenty-First Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 1290–1294.

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 12, 443–460.

Pirie, W. (1983), "Jonckheere Tests for Ordered Alternatives," in *Encyclopedia of Statistical Sciences, Volume 4*, eds. S. Kotz and N.L. Johnson, New York: John Wiley and Sons, Inc., 315–318.

Radlow, R. and Alf, E.F. (1975), "An Alternate Multinomial Assessment of the Accuracy of the Chi-Square Test of Goodness of Fit," *Journal of the American Statistical Association*, 70, 811-813.

Robins, J.M., Breslow, N., and Greenland, S. (1986), "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, 42, 311-323.

Snedecor, G.W. and Cochran, W.G. (1989), *Statistical Methods*, 8th Edition, Ames, IA: Iowa State University Press.

Somers, R.H. (1962), "A New Asymmetric Measure of Association for Ordinal Variables," *American Sociological Review*, 27, 799–811.

Stokes, M.E., Davis, C.S., and Koch, G.G. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc.

Theil, H. (1972), *Statistical Decomposition Analysis*, Amsterdam: North-Holland Publishing Company.

Thomas, D.G. (1971), "Algorithm AS-36. Exact Confidence Limits for the Odds Ratio in a 2×2 Table," *Applied Statistics*, 20, 105–110.

Valz, P.D. and Thompson, M.E. (1994), "Exact Inference for Kendall's S and Spearman's Rho with Extensions to Fisher's Exact Test in r×c Contingency Tables," *Journal of Computational and Graphical Statistics*, 3(4), 459–472.

van Elteren, P.H. (1960), "On the Combination of Independent Two-Sample Tests of Wilcoxon," *Bulletin of the International Statistical Institute*, 37, 351–361.

Woolf, B. (1955), "On Estimating the Relationship between Blood Group and Disease," *Annals of Human Genetics*, 19, 251–253.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Procedures Guide, Version 8,* Cary, NC: SAS Institute Inc., 1999. 1729 pp.

SAS® Procedures Guide, Version 8

ISBN 1–58025–482–9

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.