Paper 69-25

PROC FREQ: It's More Than Counts

Richard Severino, The Queen's Medical Center, Honolulu, HI

ABSTRACT

The FREQ procedure can be used for more than just obtaining a simple frequency distribution or a 2-way cross-tabulation. Multi-dimension tables can be analyzed using proc FREQ. There are many options which control what statistical test is performed as well as what output is produced. Some of the tests require that the data satisfy certain conditions. Some options produce a set of results from which one must select appropriately for the situation at hand. Which of the results produced by using the CHISQ option should one use? What is the WEIGHT statement for? Why would one create an output data set with the OUT= option? This paper (beginning tutorial) will answer these questions as many of the options available in Proc FREQ are reviewed.

INTRODUCTION

The name alone might lead anyone to think that primary use of PROC FREQ is to generate tables of frequencies. According to the SAS® documentation, "the FREQ procedure produces one-way to n-way frequency and cross-tabulation tables". In the second edition of The Little SAS® Book, Delwiche and Slaughter state that the most obvious reason for using PROC FREQ is to create tables showing the distribution of categorical data values. In fact, PROC FREQ is more than just a procedure for counting and cross tabulating. PROC FREQ is capable of producing test statistics and other statistical measures in order to analyze categorical data based on the cell frequencies in 2-way or higher tables.

There are quite a few options one can use in PROC FREQ and the output often includes additional information the user did not request or expect. A first time user trying to obtain a simple chi-square test statistic from a 2-way table may be surprised to see that the CHISQ option gives them more than just the Pearson Chi-Square. What are the different statistical tests and measures available in PROC FREQ? Can the output be controlled? Can you eliminate the unwanted or inappropriate test statistics? These are some of the questions that this paper will address.

OVERVIEW

The general syntax for PROC FREQ is:

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

with the last statement, TEST, being a new addition in version 7. As the options are discussed, any that are new with version 7 and not available in version 6.12 will be identified.

The only required statement is `PROC FREQ;` which will produce a one-way frequency table for each variable in the data set. For example, suppose we are using a data set consisting of the coffee data in chapter 4 of The Little SAS Book. The data consists of two variables: the type of coffee ordered and the window it was ordered from. If we run the following code:

```
proc freq;
run;
```

then the resulting output would look like that in Output 1.

Output 1. Default output for **PROC FREQ**

```
Coffee Data
-----------

Output when running:   PROC FREQ;
                       RUN;
--------------------------------

                                 Cumulative  Cumulative
COFFEE   Frequency   Percent   Frequency    Percent
-----------------------------------------------------
cap              6      20.7           6       20.7
esp              8      27.6          14       48.3
ice              4      13.8          18       62.1
kon             11      37.9          29      100.0

Frequency Missing = 1


                                 Cumulative  Cumulative
WINDOW   Frequency   Percent   Frequency    Percent
-----------------------------------------------------
d               13      43.3          13       43.3
w               17      56.7          30      100.0
```

It is best to use the TABLES statement to specify the variables for which a frequency distribution or cross-tabulation is desired. Failing to do so will result in a frequency distribution which lists all the unique values of any continuous variables in the data set as well as the categorical ones. It is good practice to include the DATA= option especially when using multiple data sets . More than one TABLE statement can be used in PROC FREQ, and more than one table request can be made on each TABLE statement.

We can divide all of the statements and options available in PROC FREQ into three primary categories:

1. Controlling the frequency or cross-tabulation output as far as content and appearance is concerned

2. Requesting statistical tests or measures

and 3. Writing tables and results to SAS data sets.

I will begin by addressing one-way tables. Those readers already familiar with one-way tables and the options that can be used with them may wish to skip to the section on two-way and higher tables.

ONE-WAY TABLES

The simplest output from PROC FREE is a one-way frequency table which lists the unique values of the variable, a count of the number of observations at each value, the percent this count represents, a cumulative count and a cumulative percent.

Suppose that we have data on the pain level experienced 24 hours after one of 3 different surgical procedures used to repair a hernia is performed. The data consists of 3 variables: the

medical center where the procedure was performed, the procedure performed and the level of pain (none, tolerable, intolerable) reported by the patient 24 hours later. The data is shown in the following data step:

```
Data pain ;
input site group pain ;
label  site = 'Test Center'
      group = 'Procedure'
       pain = 'Pain Level' ;

cards;
1  1  2
1  2  0
1  2  1
   .
   .
   .
3  3  1
;
run;
```

To obtain a frequency distribution of the pain level, we would run the following:

```
proc freq data=pain;
 tables pain;
run;
```

which would result in the one-way table in Output 2.

Output 2. One-way Frequency for Pain Data.

Pain Data

PAIN	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	Pain Level			
0	6	22.2	6	22.2
1	15	55.6	21	77.8
2	6	22.2	27	100.0

In the column with the heading 'PAIN' we see the 3 unique values for pain in the data: 0,1 and 2. The 'Frequency' column shows the number of observations for each value of PAIN and the 'Percent' column shows what percent of the data have each value of PAIN. The 'Cumulative Frequency' and 'Cumulative Percent' are simply running totals. For PAIN = 1, the cumulative frequency of 21 means that 21 cases have a PAIN value of 1 or 0. The corresponding cumulative percent of 77.8% is obtained by dividing the cumulative frequency, 21, by the total count, 27, and multiplying by 100.

The cumulative frequency and percent are most meaningful if the variable is at least ordinal. In this example, the variable PAIN is ordinal, i.e. the pain level increases as the value increases, and it makes sense to say "21 cases, or 77.8% of the sample, reported a pain level of 1 or less".

If we use PROC FORMAT to define a format to be used for the variable PAIN, we can then include a format statement in PROC FREE as in the statements below which result in the output shown in Output3.

```
proc format;
value pain 0='None' 1='Tolerable'
           2='Intolerable' ;
run;

proc freq data=pain;
 tables pain;
format pain pain.;
run;
```

Output 3. Using the **FORMAT** statement in **PROC FREE**

Pain Data - with FORMAT statement

PAIN	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	Pain Level			
None	6	22.2	6	22.2
Tolerable	15	55.6	21	77.8
Intolerable	6	22.2	27	100.0

Notice that the numeric values of PAIN have been replaced with some meaningful description.

Now let's say that the variable GROUP is coded such that Procedures A, B and C are coded as 2, 1, and 3 respectively. A format can be defined and used in PROC FREE to produce the one-way table in Output 4.

Output 4. Using the **FORMAT** statement in **PROC FREE**

Pain Data - with FORMAT statement

GROUP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	Procedure			
B	9	33.3	9	33.3
A	9	33.3	18	66.7
C	9	33.3	27	100.0

Notice that procedure 'B' appears before procedure 'A'. By default, PROC FREE orders the data according to the unformatted values of the variable, and since the unformatted value of GROUP for procedure 'A' is 2, it comes after 1 which is the unformatted value for procedure 'B'. We can rectify the situation by using the option `ORDER=FORMATTED` in the PROC FREE statement.

Output 5 shows the one-way table obtained from the following code:

```
proc freq data=pain order=formatted;
 tables group;
format pain pain.;
run;
```

Output 5. Using **ORDER=FORMATTED**

Pain Data - ORDER=FORMATTED

GROUP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	Procedure			
A	9	33.3	9	33.3
B	9	33.3	18	66.7
C	9	33.3	27	100.0

Assume that the differences between procedures A, B and C are qualitative rather than quantitative. The variable GROUP is therefore a nominal variable whose values have no natural order and the cumulative frequencies and percent produced by `PROC FREE` are rather meaningless. Including the option `NOCUM` on the `TABLES` statement will eliminate the cumulative frequency and percent from the table as shown in Output 6.

```
proc freq data=pain order=formatted;
title1 'Pain Data' ;
title2 '---------' ;
 tables group / nocum;
format group group. ;
run;
```

Output 6. Cumulative Frequency and Percent Eliminated form the Output: **NOCUM**

```
Pain Data - using NOCUM Option
------------------------------

         Procedure

GROUP    Frequency    Percent
---------------------------
A                9       33.3
B                9       33.3
C                9       33.3
```

Notice the absence of the cumulative frequency and cumulative percent. Output 7 shows that we can also eliminate the percent column by including the option NOPERCENT in the TABLES statement:

```
proc freq data=pain order=formatted;
 tables group / nocum nopercent;
format group group. ;
run;
```

Output 7. Result of using **NOCUM** and **NOPERCENT**

```
Pain Data - Using NOCUM and NOPERCENT
-------------------------------------

    Procedure

GROUP   Frequency
-----------------
A               9
B               9
C               9
```

In the Pain data example, we have the actual data from which to create the one-way frequency tables. That is, the data above consists of the response for each patient in each group at each site.

The WEIGHT statement is used when we already have the counts. For example, if we are told that in a study to estimate the prevalence of asthma, 300 people from 3 different cities were examined with the following results: of: of the 100 people examined in each city, 35, 40 and 25 were found to have asthma in Los Angeles, New York and St. Louis respectively. The following data step reads the summarized data, and then using the WEIGHT statement in PROC FREE, we produce a one-way table (Output 8) for the overall sample prevalence of asthma in the 3 cities combined.

```
data asthma;
input city asthma count;
cards;
1 1 35
1 0 65
2 1 40
2 0 60
3 1 25
3 0 75;
run;

proc format;
value asth 0='No Asthma' 1 'Asthma' ;
run;

proc freq data=asthma;
 weight count;
 tables asthma;
format asthma asth.;
run;
```

Output 8. Cumulative Frequency and Percent Eliminated form the Output: **NOCUM**

```
Asthma Data - Using WEIGHT statement
------------------------------------

                              Cumulative Cumulative
 ASTHMA    Frequency  Percent  Frequency    Percent
-----------------------------------------------------
No Asthma       200     66.7        200      66.7
Asthma          100     33.3        300     100.0
```

Of the options available on the TABLES statement, only NOCUM, NOPERCENT, NOPRINT and OUT=<SAS data set name> will have any effect on the output for a one-way table. The NOPRINT option suppresses any printed output and is usually used with the OUT= option with which one can specify the name of a SAS data set to which the contents of the table will be written.

The NOPRINT option may be used if we want to use the frequencies or percents from the one-way table as input to a data step, macro or other SAS program for customized processing.

The options for statistical tests or measures do not produce any output for one-way tables. However, one might be interested in computing a 95% confidence interval for a single proportion from a one-way table. Consider the data in example 13.1 of *Common Statistical Methods for Clinical Research with SAS® Examples* by Glenn Walker. The data is from a study in which 25 patients with genital warts are given a combination of treatments which resulted in 14 patients being cured. The standard treatment has a known cure rate of 40% and so the question is whether the success rate of the combination of treatments is consistent with that of the standard treatment. We can compute a 95% confidence interval for the proportion of successful treatments based on the study and then see if the interval includes .4 (40%) or not. To do this we will make use of the WEIGHT statement and the OUT= option in PROC FREE. The DATA step, PROC FREE statements and resulting output follow.

```
data ex13_1;
input cured $ count;
cards;
YES 14
NO  11;
run;

proc freq data=ex13_1;
   weight count;
   tables cured / nocum out=curedci ;
run;
```

The option NOCUM has suppressed the printing of the cumulative counts and frequencies which are not necessary as there are only two categories. The OUT=CUREDCI has written the table as shown to a SAS data set named CUREDCI. The data set CUREDCI has 3 variables (CURED, COUNT and PERCENT) and 2 observations (one for NO, one for YES).

Output 9. One-way Table for Data from Glenn Walker's Example 13.1

```
Data from Example 13.1

CURED    Frequency    Percent
--------------------------
NO            11       44.0
YES           14       56.0
```

The following data step uses this output data set to compute an approximate 95% confidence interval for the cure rate.

```
data curedci; set curedci ;
 P = PERCENT/100 ;
 N = COUNT + LAG(COUNT) ;

 LB = P - ( 1.96*SQRT( P*(1-P)/N ) ) ;
* reset lower bound to 0 if <0 ;
 IF LB < 0 THEN LB = 0 ;

 UB = P + ( 1.96*SQRT( P*(1-P)/N ) ) ;
* reset upper bound to 1 if >1 ;
 IF UB > 1 Then UB = 1 ;

if _N_ = 2;

label   p = 'Proportion Cured'
       LB = 'Lower Bound'
       UB = 'Upper Bound' ;
run;
```

Output 10 shows the results of the above data step when printed using PROC PRINT. While it is not within the scope of this paper to explain the details of the data step, the code is included here for illustrative purposes. From the printout we see that a 40% cure rate is included in the 95% confidence interval and so we would conclude that there is no evidence that the combination of treatments is any different form the standard treatment.

Output 10. Printout of Approximate 95% Confidence Interval for a Single Proportion

```
Approximate 95% Confidence Interval
For Proportion Cured

Proportion     Lower      Upper
  Cured        Bound      Bound

   0.56       0.36542    0.75458
```

In version 8 of SAS, the BINOMIAL option on the TABLES statement will yield similar results to those in Output 10 without any extra steps. Output 11 shows the results of running the following PROC FREE in version 8. Note that the EXACT statement has also been included. This is available in version 8 and is used here to obtain the exact test in addition to the test using the normal approximation

```
proc freq data=ex13_1 order=data;
  weight count;
  EXACT BINOMIAL;
  tables cured / nocum BINOMIAL (P=0.4) ;
run;
```

Notice that the TABLES statement above includes (P=0.4) in addition to BINOMIAL. This specifies that the null hypothesis to be tested is H_0:p=0.4. If (P= *value*) is omitted then the null hypothesis tested will be H_0:p=0.5 .

Output 11. Results with **BINOMIAL** Option and **EXACT** statement

```
The FREE Procedure

cured    Frequency     Percent
-----------------------------
YES            14       56.00
NO             11       44.00


Binomial Proportion for cured = YES
-----------------------------------
Proportion (P)                0.5600
ASE                           0.0993
95% Lower Conf Limit          0.3654
95% Upper Conf Limit          0.7546

Exact Conf Limits
95% Lower Conf Limit          0.3493
95% Upper Conf Limit          0.7560

   Test of H0: Proportion = 0.4

ASE under H0                  0.0980
Z                             1.6330
One-sided Pr >  Z             0.0512
Two-sided Pr > |Z|            0.1025

Exact Test
One-sided Pr >=  P            0.0778
Two-sided = 2 * One-sided     0.1556
```

There are 2 sets of confidence limits given in Output 11. The first set, identical to the limits in Output 10, are an approximation based on the normal distribution. The next set, labeled Exact Conf Limits are based on the Binomial distribution.

Results based on exact tests are especially desirable when the sample size is small.

TWO-WAY TABLES

The 2x2 Table

The simplest cross-tabulation is a 2x2 (said "2 by 2") table. It is called a 2x2 table because it is a cross-tabulation of two categorical variables, each with only two categories. In a cross-tabulation, one variable will be the row variable and the other will be the column variable.

The general form of a 2x2 cross-tabulation is shown in Figure 1. There are 4 cells. A cell is one combination of categories formed from the column and row variables.

Figure 1. A 2x2 table

		Column Variable 1	Column Variable 2	
Row Variable	1	a	b	r1
	2	c	d	r2
		c1	c2	n

In Figure 1, ***a*** is the number of cases that are in category one of the column variable and category one of the row variable while ***b*** is the number of cases in category two of the column variable and

category one of the row variable. The total number of cases in category one of the row variable is ***r1=a+b***, while the total number of cases in the second category of the row variable is ***r2=c+d***. Similarly, the total number of cases in the first and second categories of the column variable are ***c1= a+c*** and ***c2 = b+d*** respectively. The total number of cases is ***n = a+b+c+d = r1+r2 = c1+c2*.**

Consider the RESPIRE data described on page 41 of *Categorical Data Analysis Using the SAS System*. The variables in the data set are TREATMNT, a character variable containing values for treatment (placebo or test), RESPONSE, another character variable containing the value of response defined as respiratory improvement (y=yes, n=no), CENTER, a numeric variable containing the number of the center at which the study was performed, and COUNT, a numeric variable containing the number of observations that have the respective TREATMNT, RESPONSE and CENTER values. To obtain the cross-tabulation of treatment with response shown in Output 12, we use the **PROC FREE** statement:

```
proc freq data=respire;
   weight count;
   tables treatmnt*response;
run;
```

Note that the `WEIGHT` statement was used and that there were no options specified on the `TABLES` statement.

Output 12. Respire Data: 2x2 Table of **TREATMNT*RESPONSE**

```
TABLE OF TREATMNT BY RESPONSE

TREATMNT     RESPONSE

Frequency|
Percent  |
Row Pct  |
Col Pct  |n       |y       |  Total
---------+--------+--------+
placebo  |     52 |     38 |     90
         |  28.89 |  21.11 |  50.00
         |  57.78 |  42.22 |
         |  68.42 |  36.54 |
---------+--------+--------+
test     |     24 |     66 |     90
         |  13.33 |  36.67 |  50.00
         |  26.67 |  73.33 |
         |  31.58 |  63.46 |
---------+--------+--------+
Total          76      104      180
             42.22    57.78   100.00
```

In general, to obtain a 2x2 table using PROC FREE, simply include the statement

`TABLES R*C ;`

where R and C represent the row and column variables respectively.

Percent, Row Percent, Col Percent

The upper left hand corner of the table contains a legend of what the numbers inside each cell represent.

Let us consider the second cell of the table in Output 12 as a means of reviewing what is meant by percent, row percent and column percent (shown as col percent). The second cell is the combination of 'Yes Response' and 'Placebo', i.e. it represents the patients who were given placebo and responded favorably anyway. There were 38 cases (**frequency**) which responded to treatment and in which placebo was given as the treatment . Thirty eight cases is 21.11% (**percent**) of the 180 cases. The **row percent** of 42.22% indicates that 42.22% (38/90) of the cases receiving placebo responded and the **column percent** of 36.54% indicates that 36.54% (38/104) of the cases responding were given placebo.

Suppose we are interested in comparing the response rates between treatments. In this case we may want to ignore the column percent as well as the overall percent. The printing of the percent and column percent in each cell can be suppressed by specifying the `NOPERCENT` and `NOCOL` options on the `TABLES` statement. Output 13 shows the result of running the following statements:

```
proc freq data=respire;
  weight count;
  tables treatmnt*response / nocol nopercent;
run;
```

Output 13. 2x2 Table with Percent and Column Percent Values Suppressed

```
TABLE OF TREATMNT BY RESPONSE

TREATMNT     RESPONSE

Frequency|
Row Pct  |n       |y       |  Total
---------+--------+--------+
placebo  |     52 |     38 |     90
         |  57.78 |  42.22 |
---------+--------+--------+
test     |     24 |     66 |     90
         |  26.67 |  73.33 |
---------+--------+--------+
Total          76      104      180
```

The percent and column percent have been eliminated from the output, making it much more readable. Notice also that the margin percent values have been suppressed.

If the table had been specified with the `TREATMNT` as the column variable then we could use the `NOROW` option instead of the `NOCOL`. The following statements

```
proc freq data=respire;
   weight count;
   tables response*treatmnt /norow nopercent;
run;
```

result in the output shown in Output 14.

Output 14. Percent and Row Percent Values Suppressed

```
TABLE OF RESPONSE BY TREATMNT

RESPONSE     TREATMNT

Frequency|
Col Pct  |placebo |test    |  Total
---------+--------+--------+
n        |     52 |     24 |     76
         |  57.78 |  26.67 |
---------+--------+--------+
y        |     38 |     66 |    104
         |  42.22 |  73.33 |
---------+--------+--------+
Total          90       90      180
```

Using `NOCOL`, `NOROW` and `NOPERCENT` on the `TABLES` statement allows the output to be customized to suit the needs of the situation. Of course, there may be times when a two way table of counts is needed just to check the data (Delwiche and Slaughter),

in which case the three options can be specified together in order suppress all but the actual counts.

The ORDER = Option

Just as it affected the order in which the rows of a one-way table are printed, using the ORDER= option on the PROC FREE statement affects the order in which the columns or rows are printed. The effect is similar to that on the one-way tables except that it affects the column order as well as the row order. For example, if in the RESPIRE data the first observation had a response value of 'y' and treatmnt value of 'test' , then including ORDER=DATA on the PROC FREE statement that produced the table in output 14 would now produce a table in which the row and column order are opposite of what they were (Output 15).

Output 15. 2x2 Table when ORDER=DATA is used (Compare to Output 14)

```
TABLE OF RESPONSE BY TREATMNT

RESPONSE     TREATMNT

Frequency|
Col Pct  |test    |placebo |  Total
---------+--------+--------+
y        |     66 |    38  |    104
         |  73.33 |  42.22 |
---------+--------+--------+
n        |     24 |     52 |     76
         |  26.67 |  57.78 |
---------+--------+--------+
Total          90       90      180
```

If ORDER=INTERNAL is used, which is the default, then regardless of the order the data is in when read by the DATA step, the rows and column will be in ascending order (numerical or alphabetical).

ORDER=FORMATTED will cause PROC FREE to use the format values if any to determine the order of rows or columns and ORDER=FREQUENCY will order the columns and rows by observed column and row frequencies respectively

Use of the ORDER= option can be helpful in forcing a particular order on the table. However, as will be discussed later, the order of the rows and columns has a direct effect on the interpretation of many of the statistical measures and tests produced by other options.

More Output Control

The CUMCOL option will add cumulative column percents to the table output.

Specifying NOFREQ will suppress the printing of the cell counts, and NOPRINT will suppress printing of the table in the output. NOPRINT will not suppress the printing of any requested statistical tests or measures.

The MISSING option causes missing values to be interpreted as nonmissing and hence be included in any calculations of statistics and analysis while the MISSPRINT option will include any missing values in the table but exclude them from any calculation of statistics.

The LIST option causes the table to printed in a list format rather than as a cross-tabulation. Expected values, which will be discussed a little later, will not be printed if LIST is used. This option will be ignored if any statistical tests or measures are requested.

The SPARSE option is used to have all possible combinations of levels of the variables printed when using LIST or saved to an output data set when using OUT= . This option does not affect the printed output unless used with LIST.

The Chi-Square Test

A Chi-Square test can be used to test the null hypothesis that two categorical variables are not associated , i.e. they are independent of each other. The categorical variables need not be ordinal. In general, the Chi-Square test involves the cell frequencies and the expected frequencies. The expected frequencies are the counts we would expect if the null hypothesis of no association is true.

As an example, consider again the RESPIRE data as shown in Output 15. Of the subjects receiving the 'test' treatment, 73.33% responded favorably (i.e. showed respiratory improvement) while only 42.22% of those receiving placebo responded favorably. Is the difference in response rates statistically significant? To answer the question we can perform a chi-square test by using the CHISQ option on the TABLES statement. Output 16 shows the results obtained from running the following:

```
proc freq data=respire;
weight=count;
tables response*treatmnt
 / CHISQ norow nopercent;
run;
```

Only the results specific to the CHISQ option are shown in Output 16 as the resulting two-way table is already shown on Output 15. What is very noticeable is that the CHISQ option has produced more than one Chi-Square test result.

There are 4 headings in the ***STATISTICS*** results of the output. Below the ***Statistic*** heading is the name of the statistic from the associated statistical test. The degrees of freedom for each statistic is shown below the ***DF*** heading. The value of the test statistic and the p-value are printed under the ***Value*** and ***Prob*** headings respectively. These are the ***Chi-Square***, the ***Likelihood Ratio Chi-Square,*** the ***Continuity Adjusted Chi-Square*** and the ***Mantel-Haenszel Chi-Square*** statistics.

The p-values for ***Fisher's Exact Test*** are also given.

Output 16. Results obtained with the CHISQ Option

```
STATISTICS FOR TABLE OF RESPONSE BY TREATMNT

Statistic                    DF   Value       Prob
------------------------------------------------------
Chi-Square                    1   17.854      0.001
Likelihood Ratio Chi-Square   1   18.195      0.001
Continuity Adj. Chi-Square    1   16.602      0.001
Mantel-Haenszel Chi-Square    1   17.755      0.001
Fisher's Exact Test (Left)                    1.000
                    (Right)                1.99E-05
                    (2-Tail)               3.99E-05
Phi Coefficient                   0.315
Contingency Coefficient           0.300
Cramer's V                        0.315

Sample Size = 180
```

The last three statistics, for which there are no degrees of freedom or p-values printed, are measures of association: the ***Phi Coefficient***, the ***Contingency Coefficient*** and ***Cramer's V***.

The statistic labeled 'Chi-Square' on the first line is the Pearson Chi-Square. The null hypothesis is that the response is not associated with the treatment. The large value of the chi-square statistic, 17.854, and the p-value of 0.001 indicate that the null hypothesis should be rejected at the 0.05 level of significance. Thus we would conclude that the improvement rate in the 'test' subjects is significantly greater than that in the 'placebo' subjects.

The ***Pearson Chi-Square*** statistic involves the differences between the observed cell frequencies and the expected frequencies. A rule of thumb is that the expected frequency in every cell of the table should be at least 5. If the expected value is less than 5 in one or more cells, SAS will print a message in the output stating the percent of the cells in which this occurs.

If the expected value of one or more cells is less than 5, the chi-square test may not be valid. In this case, ***Fisher's Exact Test*** is an alternative test which does not depend on the expected values. A criticism of this test is that it fixes the row and column margin totals, which in effect makes an assumption about the distribution of the variables in the population being studied.

The ***Continuity-Adjusted Chi-Square*** test statistic consists of the Pearson Chi-Square modified with an adjustment for continuity. As the sample size increases, the difference between the continuity-adjusted and Pearson Chi-Square decreases. Thus in very large samples the two statistics are almost the same. This test statistic is also an alternative to Pearson's if any of the expected values in a 2x2 table are less than 5 (Cody and Smith, 1997). Some prefer to use the continuity-adjusted chi-square statistic when the sample size is small regardless of the expected values.

The expected frequencies can be obtained by using the `EXPECTED` option on the `TABLES` command. Additionally, the difference between the observed cell count and the expected cell count can be obtained by using the `DEVIATION` option. Finally, the amount that each cell contributes to the value of the test statistic will be printed by specifying `CELLCHI2.`

Output 17 shows the expected values, differences between the expected and observed counts as well as each cell's contribution to the Chi-Square statistic. None of the expected values are less than 5 and the sum of the `Cell Chi-Square` values is equal to the `Pearson Chi-Square` test statistic in Output 16. Some interpretations

The ***Mantel-Haenszel Chi-Square*** is related to the Pearson Chi-Square and, in the 2x2 case, as the sample size gets large these statistics converge (Stokes, Davis and Koch, 1995). In the case of 2xC or Rx2 tables, if the variable with more than 2 categories is ordinal, the Mantel-Haenszel Chi-square is a test for trend (Cody and Smith,1997) while the Pearson Chi-square remains a general test for association.

The ***Likelihood Ratio Chi-Square*** is asymptotically equivalent to the Pearson Chi-Square and Mantel-Haenszel Chi-Square but not usually used when analyzing 2x2 tables. It is used in logistic regression and loglinear modeling which involves contingency tables. Unless PROC FREE is being used as part of a loglinear or logistic regression analysis, the likelihood ratio chi-square can be ignored.

The ***Phi Coefficient*** is a measure of the degree of association between two categorical variables and is interpretable as a correlation coefficient. It is derived from the Chi-Square statistic, but is free of the influence of the total sample size (Fleiss, 1981). Being independent of the sample size is a desirable quality because the Chi-Square statistic itself is sensitive to sample size. As the sample size increases, the Chi-Square value will increase even if the cell proportions remain unchanged.

Output 17. Printout with `EXPECTED`, `DEVIATION` and `CELLCHI2.`

```
TABLE OF RESPONSE BY TREATMNT

RESPONSE        TREATMNT

Frequency       |
Expected        |
Deviation       |
Cell Chi-Square|
Col Pct         |placebo |test    |  Total
----------------+--------+--------+
n               |     52 |     24 |     76
                |     38 |     38 |
                |     14 |    -14 |
                | 5.1579 | 5.1579 |
                |  57.78 |  26.67 |
----------------+--------+--------+
y               |     38 |     66 |    104
                |     52 |     52 |
                |    -14 |     14 |
                | 3.7692 | 3.7692 |
                |  42.22 |  73.33 |
----------------+--------+--------+
Total                90       90      180
```

The ***Contingency Coefficient*** is another measure of association derived from the chi-square and is similar to the Phi coefficient in interpretation.

Cramer's V is also derived from the chi-square and in the 2x2 table it is identical to the Phi coefficient.

These three measures of degree of association are well suited for nominal variables in which the order of the levels is meaningless.

In Output 17 the expected values are all greater than 5 and the sample size of 180 is not small. The Pearson Chi-Square is appropriate and adequate for the analysis of response by treatment. We may want to produce an output with only the Pearson Chi-Square statistics. This can be accomplished by using the `OUTPUT` statement to specify a data set to which PROC FREE will save only the statistics we specify and then including the `OUT=` option on the `TABLES` statement so that the table is output to a data set. The following code

```
proc freq data=respire;
  weight count;
  tables response*treatmnt
    / norow nopercent chisq OUT=resptabl;
  OUTPUT OUT=stats PCHI ;
run;
```

will generate the same output as Output 15 and Output 16 combined while creating a data set named ***stats*** which will contain the value, degrees of freedom and p-value for the Pearson Chi-Square and another data set ***resptabl*** which will contain the table itself. With these two data sets, it is possible to create a printout which consisted of the table in Output 15 and the Pearson Chi-Square statistic only. Output 18 shows a printout of the information written to the ***stats*** data set.

Output 18. Contents of data set ***STATS*** resulting from the `OUTPUT OUT=stats PCHI;` statement

Pearson Chi-Square	DF	P Value
17.8543	1	.0001

Thus it is possible to create a printout which includes only the desired information.

The Odds Ratio and Relative Risk

The odds ratio is simply a ratio of odds. Recall that the odds of an event occurring is the ratio of p/q where p is the probability of the event occurring and q is the probability of the event not occurring. In the RESPIRE data (as shown in Output 15), the odds of improvement in the 'test' treatment is 66/24 while the odds of improvement in the placebo group is 38/52 which yields an odds ratio of

$$\frac{66/24}{38/52} = \frac{2.75}{0.731} = 3.76$$

So the odds of improvement are almost 4 times greater in the test group than in the placebo group.

The odds ratio is obtained by including the `CMH` option on the `TABLES` statement. Consider the 2x2 table layout in Figure 2. PROC FREE computes the odds ratio as

$$\frac{a/c}{b/d} = \frac{a*d}{b*c}$$

regardless of what the column and row variables represent. Therefore, it is important that the table be specified properly when requesting the odds ratio so that the odds ratio of interest is produced.

Figure 2. Table of Outcome By Exposure or Treatment

Exposure or Treatment		Outcome 1	Outcome 2	
	1	a	b	r1
	2	c	d	r2
		c1	c2	n

Output 19 shows the output produced by the following statements:

```
proc freq data=respire order=data;
   weight count;
   tables treatmnt*response / CMH;
run;
```

Under the `Estimates of Common Relative Risk (Row1/Row2)` section of the output, the odds ratio is 3.763, and the 95% confidence interval is also given.

The values corresponding to `Cohort (Col1 Risk)` and `Cohort (Col2 Risk)` are Relative Risk estimates.

Relative Risk is the ratio of the incidence of an outcome given one treatment or exposure level to the risk of the outcome in the other level of treatment or exposure. Recall an incidence rate is the proportion of new cases (outcomes) occurring over a period of time whereas prevalence is the proportion of cases existing at any one time. Therefore the risk of an outcome makes sense in the context of prospective cohort studies where the outcome has not occurred in any case at the start of the study. In such a context, referring to, figure 2, the risk of outcome 1 is R1=**a/r1** for exposure level 1 and R2=**c/r2** for exposure level 2. The relative risk of outcome 1 is R1/R2 for subjects with exposure 1 compared to those with exposure 2.

Output 19. Statistics produced by `CMH` for Respire data

```
TABLE OF TREATMNT BY RESPONSE

TREATMNT     RESPONSE

Frequency|
Percent  |
Row Pct  |
Col Pct  |y       |n       |  Total
---------+--------+--------+
test     |     66 |     24 |     90
         |  36.67 |  13.33 |  50.00
         |  73.33 |  26.67 |
         |  63.46 |  31.58 |
---------+--------+--------+
placebo  |     38 |     52 |     90
         |  21.11 |  28.89 |  50.00
         |  42.22 |  57.78 |
         |  36.54 |  68.42 |
---------+--------+--------+
Total         104       76      180
            57.78    42.22   100.00

SUMMARY STATISTICS FOR TREATMNT BY RESPONSE


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

Statistic   Alternative Hypothesis    DF       Value    Prob
---------------------------------------------------------------
   1        Nonzero Correlation        1      17.755   0.001
   2        Row Mean Scores Differ     1      17.755   0.001
   3        General Association        1      17.755   0.001



     Estimates of the Common Relative Risk (Row1/Row2)
                                                        95%
Type of Study    Method              Value   Confidence Bounds
-----------------------------------------------------------------
Case-Control     Mantel-Haenszel     3.763      2.032    6.971
  (Odds Ratio)   Logit               3.763      2.010    7.045

Cohort           Mantel-Haenszel     1.737      1.343    2.245
  (Col1 Risk)    Logit               1.737      1.323    2.280

Cohort           Mantel-Haenszel     0.462      0.322    0.661
  (Col2 Risk)    Logit               0.462      0.314    0.679

The confidence bounds for the M-H estimates are test-based.

Total Sample Size = 180
```

In Output 19, the relative risk of improvement is 1.737 for subjects in the test group relative to the placebo group indicating that subjects in the test treatment are at higher "risk of improvement". Similarly, subjects in the test group are at lower risk of not improving compared to those in the placebo group (`Col2 risk=0.462`).

While the Relative Risk is a measure which is appropriate for prospective cohort studies, the Odds Ratio can be used for cross-sectional case-control studies as well as prospective studies. The estimates in the output are even labeled under '`Type of Study`'. In both cases, a value of 1 indicates no difference between groups.

Finally, the reader should verify that interchanging the row and column variables or modifying the table order by using the `ORDER=` option will result in different values of odds ratio and relative risks. The interpretations should however remain consistent.

Matched Pairs Data

So far, the data has consisted of independent observations, i.e. each subject is measured once for each variable in the data set. Consider the hypothetical example, presented by Cody and Smith, in which 100 people are asked if they have a positive or negative attitude toward smoking. The people are then shown an anti-cigarette advertisement and asked about their attitude again. In this case the data do not consist of independent observations, but rather matched pairs. The frequencies in the 2x2 table in Output 20 represent pairs of observations rather than single counts, i.e. there are 45 people whose attitude was positive before but negative after seeing the ad. ***McNemar's test*** is the appropriate test the hypothesis that the before and after answers are in agreement. It is a chi-square test and the value of the test statistic is 22.273 with a p-value of 0.001. The 45 people whose attitude became negative after seeing the advertisement represent a statistically significant departure from the hypothesis of agreement.

Output 20. McNemar's Test for Matched Pairs Data

```
TABLE OF BEFORE BY AFTER

BEFORE     AFTER

Frequency|Negative|Positive|  Total
---------+--------+--------+
Negative |     30 |     10 |     40
---------+--------+--------+
Positive |     45 |     15 |     60
---------+--------+--------+
Total          75        25      100

STATISTICS FOR TABLE OF BEFORE BY AFTER

McNemar's Test
--------------
Statistic = 22.273    DF = 1    Prob = 0.001


Simple Kappa Coefficient
------------------------
                                95%
                          Confidence Bounds
Kappa = -0.000  ASE = 0.077   -0.151    0.151


Sample Size = 100
```

McNemar's test is requested by including the AGREE option on the TABLES statement as follows:

```
tables before*after / AGREE ;
```

With respect to the notation in figure 1, if $b<4$ or $c<4$ then two conditions must be met: $b^2 \geq c(4-b)$ and $c^2 \geq b(4-c)$ (Walker).

The simple ***kappa coefficient*** is a measure of agreement (sometimes called inter-rater agreement) which is 0 when the agreement is due to chance alone and 1 when there is perfect agreement. The value under the heading ASE is the standard error used to calculate the 95% confidence bounds.

A 2x2x2 Table

Recall that there was a third variable CENTER in the Respire data set. This variable represents the hospital where the patients were treated and is a stratification variable. The association between the treatment and response variables can be evaluated while controlling for CENTER. This is a common strategy, and

Output 21. Respire Data: Controlling for CENTER

```
TABLE 1 OF TREATMNT BY RESPONSE
CONTROLLING FOR CENTER=1

TREATMNT     RESPONSE

Frequency|
Row Pct  |
Col Pct  |y       |n       |  Total
---------+--------+--------+
test     |     29 |     16 |     45
         |  64.44 |  35.56 |
         |  67.44 |  34.04 |
---------+--------+--------+
placebo  |     14 |     31 |     45
         |  31.11 |  68.89 |
         |  32.56 |  65.96 |
---------+--------+--------+
Total          43       47       90


TABLE 2 OF TREATMNT BY RESPONSE
CONTROLLING FOR CENTER=2

TREATMNT     RESPONSE

Frequency|
Row Pct  |
Col Pct  |y       |n       |  Total
---------+--------+--------+
test     |     37 |      8 |     45
         |  82.22 |  17.78 |
         |  60.66 |  27.59 |
---------+--------+--------+
placebo  |     24 |     21 |     45
         |  53.33 |  46.67 |
         |  39.34 |  72.41 |
---------+--------+--------+
Total          61       29       90

SUMMARY STATISTICS FOR TREATMNT BY RESPONSE
CONTROLLING FOR CENTER

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

Statistic   Alternative Hypothesis    DF       Value     Prob
-------------------------------------------------------------
   1        Nonzero Correlation        1      18.411    0.001
   2        Row Mean Scores Differ     1      18.411    0.001
   3        General Association        1      18.411    0.001


     Estimates of the Common Relative Risk (Row1/Row2)
                                                        95%
Type of Study    Method             Value   Confidence Bounds
-------------------------------------------------------------
Case-Control     Mantel-Haenszel    4.029      2.132    7.614
  (Odds Ratio)   Logit              4.029      2.106    7.707

Cohort           Mantel-Haenszel    1.737      1.350    2.235
  (Col1 Risk)    Logit              1.676      1.294    2.170

Cohort           Mantel-Haenszel    0.462      0.324    0.657
  (Col2 Risk)    Logit              0.474      0.326    0.688

The confidence bounds for the M-H estimates are test-based.


Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square =   0.000           DF =   1          Prob = 0.990

Total Sample Size = 180
```

the stratification variable may also be an explanatory variable associated with either of the treatment or response variables. We again request the **Cochran-Mantel-Haenszel test** (`CMH`), while specifying a three-way table (`center*treatmnt*response`). Output 21 shows the results obtained from the following PROC FREE statement:

```
proc freq data=respire order=data;
  weight count;
  tables center*treatmnt*response
       / nopercent cmh;
run;
```

Notice that are two 2x2 tables, one for each level of CENTER. In this example there are only two levels for CENTER, but the stratification variable may have many levels.

First consider the ***Breslow-Day Test for Homogeneity of the Odds Ratios*** appearing at the end of the output. This is a test to determine if the odds ratios at each level of the control variable (CENTER) are homogeneous. If the odds ratios are in opposite directions for example, then it would not be appropriate to compute an overall odds ratio. The null hypothesis here is that the odds ratios are homogeneous. Since we can conclude that this is the case we turn our attention to the CMH statistics.

There are 3 different alternative hypothesis for the Cochran-Mantel-Haenszel statistics: *Nonzero Correlation, Row Mean Scores Differ and General Association*. "These pertain to the situation where you have sets of tables with two ro more rows or columns" and are discussed in detail by Stokes, Davis and Koch. In the case of 2x2 tables, they are the same, and for the Respire data, the alternate hypothesis is that there is an overall association between the treatment and response, adjusting for CENTER. The p-value (Prob) of 0.000 should be interpreted as <0.001 which of course is statistically significant at the 0.05 level.

The overall odds ratio and relative risks are printed with their corresponding 95% confidence intervals.

Had the test for homogeneity of the odds ratios been statistically significant, a closer examination of each 2x2 table at each strata of the stratification variable would be required before making any further interpretations or conclusions.

K x R x C Tables

The analysis covered thus far can be extended to the analysis of more general KxRxC tables. Stokes, Davis and Koch discuss in detail the analyses of 2xC, Kx2xC, Rx2, KxRx2 and KxRxC tables. While it is not possible to cover all of these types of tables in sufficient detail in this paper, the key to understanding the higher dimension tables is to first understand the two way table and sets of 2 way tables.

Cohcran-Mantel-Haenszel: CMH, CMH1 and CMH2

As shown earlier, the CMH option results in 3 statistics with 3 different alternative hypothesis being printed. While all 3 are the same in the 2x2 case, generally, the *Nonzero Correlation* and *Row Mean Scores Differ* are the appropriate alternate hypotheses for a 2xC table, while the Nonzero Correlation alternate hypothesis is appropriate for the Rx2 table (Stokes, Davis and Koch). All three are produced when the `CMH` option is used. The `CMH1` option will produce only the Nonzero Correlation statistic and `CMH2` will produce the *Row Mean Scores Differ* statistics in addition to the *Nonzero Correlation*.

The `SCORES=` option specifies the type of column and row scores to be used in computing the Mantel-Haenszel Chi-Square, Cohchran-Mantel-Haenszel statistics and the Pearson Chi-Square. Each score type assumes certain things about the distribution of the data, and hence this option should be used carefully. See Stokes, Davis and Koch for a discussion of this option.

`MEASURES` and Other Options

The `MEASURES` option will produce measures of association which may be helpful in assessing the strength of the association. Most readers will be familiar with Pearson's correlation coefficient as a measure of association. Pearson's correlation is not appropriate if the data is not on an interval scale. Spearman's correlation is based on ranks and only requires that the data be ordinal. There are other ordinal measures of association. There are also nominal measures of association. In version 8, including the option `CL` will produce confidence intervals for the measures. Not all these measures can be interpreted the same way, and so the choice of measure of association depends on the specific situation at hand. The reader should refer to a statistics book to determine which measure is appropriate to use.

The `EXACT` option on the tables statement will produce Fisher's Exact test for tables that are larger than 2x2. In version 8, this option has been renamed: `FISHER`.

In version 8, there is now an `EXACT` statement which with which exact tests and confidence bounds can be requested.

The `TEST` statement in version 8 is used to request asymptotic tests for measures of association and agreement.

CONCLUSION

PROC FREE is more than just a counting program. While this paper has concentrated on demonstrating many of the options using one-way and two-way tables, the same options can be used in the analysis of multidimensional contingency tables. The text by Stokes, Davis and Koch is a particularly good reference for the reader who wishes to study PROC FREE in more detail.

REFERENCES

Cody, Ronald P, and Smith Jefferey K, *Applied Statistics and the SAS® Programming Language, Fourth Edition,* Prentice-Hall Inc, 1997.

Delwiche, Lora D and Slaughter, Susan J, *The Little SAS® Book, Second Edition*, Cary, NC: SAS Institute Inc., 1998.

Fleiss, Joseph L., *Statistical Methods for Rates and Proportions, Second Edition*, John Wiley and Sons, 1981.

SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SAS Institute Inc., 1989.

SAS Institute Inc., *SAS/STAT® Software Changes and Enhancements Through Release 6.11*, Cary, NC: SAS Institute Inc., 1996.

Stokes, Maura E., Davis, Charles S and Koch, Gary G., *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc., 1995.

Walker, Glenn, *Common Statistical Methods for Clinical Research with SAS® Examples*, Cary, NC: SAS Institute Inc., 1997.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Richard Severino
The Queen's Medical Center
1301 Punchbowl Street
Honolulu, HI 96813
Work Phone: (808) 547-4427
Fax: (808) 537-7897
Email: severino@hawaii.edu, rseverino@queens.org
Web: www.pharmasug.org