FREQ-OUT:
An Applied Presentation of the Options and Output of the FREQ Procedure

Pamela Landsman MPH, Merck & Co., Inc, West Point, PA

Abstract:

Have you ever been told "compare the rate of death by gender", or "give me the p-value for the difference between these 2 percentages", or "is there a trend in this association", or even "check the p-value that's reported in this paper for this 2x2 table"? You know you need PROC FREQ, but how do you get the data in? Which statistic should you report; SAS® presents 4 p-values just by running the procedure with the CHISQ option for a simple 2x2 table. This presentation should put you on the right path.

Guidelines for when to use the different tests of significance and interpretation of the output from PROC FREQ will be presented. Designed for the non-statistician, this presentation will teach you about different data structures (e.g., binary versus ordinal data); how to determine what the investigator really wants (e.g., when to run a stratified analysis and when to use the BY statement); and which p-value to report. Additional topics covered will be tables greater than 2x2, options such as ORDER=, MISSING, MISSPRINT, SCORES=, EXACT, CMH, and the WEIGHT statement.

Introduction:

PROC FREQ can be used to analyze nominal, ordinal, and even continuous response data. Nominal data (e.g. race, gender, model of car owned) are categorical responses with no inherent ordering. Ordinal data (e.g. shirt size of S, M, L, XL, or clinical efficacy such as Cure, Improve, Failure) are categorical responses that are ordered however, the number of values is limited. In order for continuous response data (e.g., age weight, exam score) to be analyzed utilizing PROC FREQ, the data has to be grouped into discrete categories such as 10 year groups for age or $5000 increments for income. This can be accomplished in a data step by creating a new variable or by formatting the values. Although PROC FREQ has only 4 statements, there are nearly 30 options that can provide important information for data management purposes as well as statistical analysis of these types of data.

Example:

Suppose your company, "Programmers Programmers" places SAS® programmers in corporations of all shapes and sizes. You are asked to provide an analysis of which applicants were hired in the past year, focusing on gender and previous experience information. Excerpts from the PROC CONTENTS of your dataset TEST.HIRED follow:

```
----Alphabetic List of Variables and Attributes-----
#    Variable    Type    Len    Pos    Label
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
1    GENDER      Char     1      0     Gender (M,F)
3    HIRED       Char     1      9     Hired(0=No,1=Yes)
2    LEVEL       Num      8      1     Yrs. Experience
                                         (0,1,2,3,4,5+)
```

The questions you are asked to answer are:

1. What is the gender and experience distribution of our applicants?
2. Are those hired more experienced than those not hired?
3. Do we need to account for any differences in gender when comparing experience and hiring status?

Each of these questions will be answered in order. They will demonstrate the increasing sophistication of code and the nuances of analyzing this type of data.

Question 1: What is the gender distribution of our clients?

Ideally you should look at the distribution of each of your variables. This will aid you in determining exactly what type of statistical tests you might need to conduct. The code below utilizes the PROC and TABLES statements and the DATA= option.

```
PROC FREQ DATA=TEST.HIRED ;
   TABLES GENDER ;
RUN ;
```

```
                Gender (M,F)
                                   Cumulative  Cumulative
GENDER    Frequency     Percent     Frequency     Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
F                81       49.7            81       49.7
M                82       50.3           163      100.0
               Frequency Missing = 4
```

Variables for which you want to know the distribution can be added to the list as part of the TABLES statement. You have answered your first question, women comprise 49.7% of your population, or have you? Note there are 163 clients with known genders and 4 clients with missing values for GENDER.

You may be asking yourself, do these clients have any other missing information, are there clients with known gender but other information missing? What proportion of the data do those with missing gender represent? What proportion of the 167 clients have complete data? These questions can be answered with a slight modification to the list of variables in the TABLES statement and by adding the LIST and MISSING options to the above code. The LIST opion presents the combination of multiple variables in list format rather than the default of tabular format. An additional change you would like to make is a cosmetic one. If we include the HIRED variable in the TABLES statement, you will see that it has the values of 1 and 0. Instead you would like to see the words 'Hired" and 'Not Hired' appear; this is accomplished with a combinaton of PROC FORMAT and the FORMAT statement.

```
PROC FORMAT ;
   VALUE $HIRE '0'='Not Hired' '1'='Hired' ;
RUN ;
PROC FREQ DATA=TEST.HIRED ;
   FORMAT HIRED HIRE. ;
   TABLES GENDER*LEVEL*HIRED/LIST MISSING ;
RUN ;
```

Sample output:

GENDER	LEVEL	HIRED	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	.	Hired	4	2.4	4	2.4
F	.	Not Hired	1	0.6	5	3.0
F	0	Not Hired	9	5.4	14	8.4
F	0	Hired	8	4.8	22	13.2
F	1	Not Hired	7	4.2	29	17.4
F	1	Hired	4	2.4	33	19.8
F	2	Not Hired	22	13.2	55	32.9
F	2	Hired	1	0.6	56	33.5
F	3	Not Hired	12	7.2	68	40.7
F	3	Hired	2	1.2	70	41.9
F	4	Not Hired	10	6.0	80	47.9
F	4	Hired	1	0.6	81	48.5
F	5	Not Hired	4	2.4	85	50.9
M	.	Not Hired	1	0.6	86	51.5
M	.	Hired	1	0.6	87	52.1
M	0	Not Hired	33	19.8	120	71.9
M	0	Hired	1	0.6	121	72.5
M	1	Not Hired	15	9.0	136	81.4
M	1	Hired	3	1.8	139	83.2
M	2	Not Hired	4	2.4	143	85.6
M	3	Not Hired	7	4.2	150	89.8
M	3	Hired	3	1.8	153	91.6
M	4	Hired	8	4.8	161	96.4
M	5	Not Hired	5	3.0	166	99.4
M	5	Hired	1	0.6	167	100.0

Notice that the observations with missing values are included in the denominator of the percent calculations. In the previous example (without the MISSING option) they were not. If you would like the missing values as part of your table but not included in the denominator, you can use the MISSPRINT option instead.

As you can observe, 4 of those hired have both a missing gender and experience level (2.4% of the total), 1(0.6%) female client has a missing experience level, and 1(0.6%) male who was hired and 1(0.6%) not hired also have missing experience levels. This totals 7(4.2%) of the 167 clients have missing data. You might now decide to drop these cases from further analysis, which will be done for the remainder of the examples. The total number of observations in the dataset will be 160. You might also rerun the analysis used to answer Question 1.

Question 2: Are those hired more experienced than those not hired?

This question asks you to compare the distribution of experience for those hired with the distribution of experience for those not hired. Experince level is a numeric value that takes on the values 0-5. From the label of the variable LEVEL you notice the value "5" does not represent "5 years", but "5 or more years" of experience. You therefore rule out tests of significance such as the t-test or Wilcoxon Rank Sums. Based on your knowledge of analyzing this type of data the decision made is to evaluate a chi-square test of significance. You also want to make some cosmetic changes as well.

From the distribution of "HIRED" you notice that "Not Hired" is appearing first in the list. SAS® prints the formatted values in the order of the underlying values, 0 (not hired) comes before 1 (hired). This is the default. How can you get SAS® to print the order differently? This time utilize the ORDER= option with the PROC statement.

```
PROC FREQ DATA=TEST.HIRED2
          ORDER=FORMATTED ;
   FORMAT HIRED $HIRE. ;
   TABLES HIRED*LEVEL/CHISQ NOCOL
                              NOPCT;
RUN ;
```

```
HIRED(Hired(0=No,1=Yes))
        LEVEL(Yrs. Experience (0,1,2,3,4,5+))

           ,    0,    1,    2,    3,    4,    5,  Total
  ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆ
  Not Hired ,  42 ,  22 ,  26 ,  19 ,  10 ,   9 ,    128
           ,32.81,17.19,20.31,14.84, 7.81, 7.03,  80.00
  ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆ
  Hired     ,   9 ,   7 ,   1 ,   5 ,   9 ,   1 ,     32
           ,28.13,21.88, 3.13,15.63,28.13, 3.13,  20.00
  ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆ
  Total        51    29    27    24    19    10     160
           31.88 18.13 16.88 15.00 11.88 6.25   100.00
```

Statistic	DF	Value	Prob
Chi-Square	5	14.498	0.013
Likelihood Ratio Chi-Square	5	14.636	0.012
Mantel-Haenszel Chi-Square	1	1.163	0.281

Notice that the chi-square test shows a significant relationship (p=0.013) but the Mantel Haenszel chi-square shows a non-significant relationship (p=0.281). Without going into the statistical theory, just be assured that when you have a table where at least one of the variables is ordered (row or column) and that variable has more than 2 levels, the Mantel-Haenszel (M-H) statistic is the statistic to use. Thus, there appears to be no statistically significant relationship between hiring status and years of experience (p=0.281). *Note: There are extensions of the M-H chi-square that you can take advantage of using the SCORES= option on the TABLES statement. However, this topic will not be discussed here. The references at the end all include discussions of this very useful option.*

Take another look at the table above. Is the presentation of the data optimal? How do you relay this information to your company?. As previously noted, 5 years of experience is really 5+ years. Also, is management able to communicate to future employers the difference between 3 years of experience and 4 years? Maybe it would be better to group the years of experience into more generic categories. Entry, Intermediate, and Advanced programmer are much more preferred as categories when trying to communicate experience level to potential employers. You write the code to format the data such that 0-1 years=Entry, 2-4 years=Intermediate, and 5+ years=Advanced, and rerun the analysis.

```
PROC FORMAT ;
```

```
        VALUE EXP 0-1='Entry' 2-4='Intermediate'
5='Advanced' ;
        RUN ;

PROC FREQ DATA=TEST.HIRED2
                ORDER=FORMATTED ;
        FORMAT HIRED $HIRE. LEVEL EXP. ;
        TABLES HIRED*LEVEL/CHISQ NOCOL NOPCT;
        RUN ;
```

First take a look at the table that is generated.

	,Advanced	,Entry	,Intermed, iate	Total
Hired	1	16	15	32
	3.13	50.00	46.88	20.00
Not Hired	9	64	55	128
	7.03	50.00	42.97	80.00
Total	10	80	70	160
	6.25	50.00	43.75	100.00

Notice anything unusual? The values of LEVEL seem to be mixed up. Remember, ORDER=FORMATTED puts the data in alphabetical order of the formats. In this case you may want the data to be in the order of the underlying values instead. The default is ORDER=DATA. The analysis is rerun and the following table and statistics are presented.

	,Entry	,Intermed iate	,Advanced	Total
Not Hired	64	55	9	128
	50.00	42.97	7.03	80.00
Hired	16	15	1	32
	50.00	46.88	3.13	20.00
Total	80	70	10	160
	50.00	43.75	6.25	100.00

Statistic	DF	Value	Prob
Chi-Square	2	0.714	0.700
Likelihood Ratio Chi-Square	2	0.822	0.663
Mantel-Haenszel Chi-Square	1	0.184	0.668

The relationship remains non-significant. However, after discussing this result with your colleagues you are asked to rerun the analysis. They believe that since there are so few advanced candidates (n=10) compared to total number of applicants (n=160; 6.25%) and only 1 of those were hired they would like to see LEVEL reformatted so that the comparison groups will be “Entry” and “Non-Entry”.

	,Entry	,Non- Entry,	Total
Not Hired	64	64	128
	50.00	50.00	80.00
Hired	16	16	32
	50.00	50.00	20.00
Total	80	80	160
	50.00	50.00	100.00

Statistic	DF	Value	Prob
Chi-Square	1	0.000	1.000
Likelihood Ratio Chi-Square	1	0.000	1.000
Continuity Adj. Chi-Square	1	0.000	1.000
Mantel-Haenszel Chi-Square	1	0.000	1.000

Your conclusion remains the same; question 2 has been answered.

Question 3: Do we need to account for any differences in gender when comparing experience and hiring status?

One of the first decisions you make in order to answer this last question is that you will again format LEVEL into 2 responses. Next you may want to see how men and women differ for the individual factors of hiring status and experience. For this example, significance testing will not be conducted.

```
PROC FREQ DATA=TEST.HIRING
                ORDER=FORMATTED ;
    FORMAT HIRED $HIRE. LEVEL EXP. ;
    TABLES GENDER*(HIRED LEVEL) ;
RUN ;
```

Frequency, Percent , Row Pct , Col Pct	,Hired	,Not Hire,d	Total
F	16	64	80
	10.00	40.00	50.00
	20.00	80.00	
	50.00	50.00	
M	16	64	80
	10.00	40.00	50.00
	20.00	80.00	
	50.00	50.00	
Total	32	128	160
	20.00	80.00	100.00

Frequency, Percent , Row Pct , Col Pct	,Entry	,Non- Entry,	Total
F	28	52	80
	17.50	32.50	50.00
	35.00	65.00	
	35.00	65.00	
M	52	28	80
	32.50	17.50	50.00
	65.00	35.00	
	65.00	35.00	
Total	80	80	160
	50.00	50.00	100.00

The above tables show that 20% of the women and 20% of the men applicants are hired. However while 35% of the women applicants would be considered at the entry level, 65% of the men are at the same level. Now you need to combine all of this information to answer question 3.

How can you "adjust" for possible gender differences. Should you decide to use the BY statement to look at hiring status and experience by gender, you would be on the right track. However, this will produce 2 separate tables, one for experience and hiring status for women and one for men. This would result in the individual statistics that test for the relationship of interest but not "account for possible differences" between men and women. Other terms you may see, which all mean the same thing are "are men different then women" , "can we ignore gender when we are looking at hiring status and experience?", "stratify by gender" or "control for gender". In essence what is required is a way to statistically combine the two tables that would be generated utilizing the BY statement. You want a multi-way table. This can be accomplished with the CMH option in the TABLES statement and adding an additional level of the table to the TABLES statement. CMH stands for Cochran-Mantel-Haenszel. There are additional uses for the CMH option that can be found in the references cited.

```
PROC FREQ DATA=TEST.HIRED2
                  ORDER=FORMATTED ;
   FORMAT HIRED HIRE. LEVEL EXP. ;
   TABLES GENDER*HIRED*LEVEL/CHISQ CMH ;
RUN ;
```

First take a look at the 2 tables that are generated by this code and the individual tests of significance for the individual associations.

CONTROLLING FOR GENDER=F

HIRED(Hired(0=No,1=Yes)) by LEVEL(Yrs. Experience (0,1,2,3,4,5+))

Frequency , Percent , Row Pct , Col Pct	Entry	Non-Entry	Total
Hired	12	4	16
	75.00	25.00	20.00
Not Hired	16	48	64
	25.00	75.00	80.00
Total	28	52	80
	35.00	65.00	100.00

Statistic	DF	Value	Prob
Chi-Square	1	14.066	0.001
Likelihood Ratio Chi-Square	1	13.618	0.001
Continuity Adj. Chi-Square	1	11.954	0.001
Mantel-Haenszel Chi-Square	1	13.890	0.001
Fisher's Exact Test (Left)			3.38E-04
(Right)			1.000
(2-Tail)			3.38E-04

CONTROLLING FOR GENDER=M

HIRED(Hired(0=No,1=Yes)) by LEVEL(Yrs. Experience (0,1,2,3,4,5+))

Frequency , Percent , Row Pct , Col Pct	Entry	Non-Entry	Total
Hired	4	12	16
	25.00	75.00	20.00
Not Hired	48	16	64
	75.00	25.00	80.00
Total	52	28	80
	65.00	35.00	100.00

Statistic	DF	Value	Prob
Chi-Square	1	14.066	0.001
Likelihood Ratio Chi-Square	1	13.618	0.001
Continuity Adj. Chi-Square	1	11.954	0.001
Mantel-Haenszel Chi-Square	1	13.890	0.001
Fisher's Exact Test (Left)			1.000
(Right)			3.38E-04
(2-Tail)			3.38E-04

You can see that the tests of significance for both males and females is significant, p<0.001 for each. Next you will want to evaluate the overall test of significance "controlling" for gender.

SUMMARY STATISTICS FOR HIRED BY LEVEL
CONTROLLING FOR GENDER

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.000	1.000
2	Row Mean Scores Differ	1	0.000	1.000
3	General Association	1	0.000	1.000

For this example all the statistics are the same and it appears that there is no relationship of hiring status and experience once you account for differences of gender. Wait a minute, the individual tests were both significant? Individually there is a very strong relationship, why did it disappear? But this should make sense. In question 2 the test ignoring or "pooling" the two groups showed no relationship between hiring status and experience (p=0.281). Obviously there must be some difference between men and women to give opposite results when you stratify (highly significant relationships) than when you do not (strongly non-significant).

Take a look back at the two tables. Notice that for the women, 75% of those that are hired are at the entry level. For the men, 25% of those hired are also at the entry level. The tables go in the opposite direction! This phenomenon is referred to as Simpson's Paradox. Had the tables comparing gender with hiring status also shown a difference as the table of gender and experience level the phenomenon above would be referred to as confounding. What is happening is the difference seen

for the men is being canceled by the difference seen for the women. To answer question 3 you would say yes, gender matters, and present management with the results of the individual analyses. You would also have to tell your colleques that they must report these results and not the results of question 2. As just demonstrated, the analysis of question 2 is true, but misleading.

Neat Trick:

Now that you have conducted your analysis and presented the results you go back to your desk. The phone rings. It is one of the managers who attended the meeting and she has an additional question. She tells you that while your analysis showed equal placement rates of 20% for men and women, for the years of 1990-1996 similar numbers were found; 25% of female clients and 20% of male clients were placed. She is writing a report and needs the p-value showing that the rates for the 6 years in total are not different. How can you run this analysis using PROC FREQ? The raw data used to generate these numbers is not available.

Here is a neat trick that you can use to produce a 2x2 table (or larger). First you need a little more information, you need to know the numerator and denominator associated with each percentage. You are told that 25% represents 145/578 women and 20% represents 97/496 men. This will be demonstrated below.
Start by drawing a picture of what your contingency table should look like based on the known information.

	Hiring Status		Total
Gender	No	Yes	
Female		145	578
Male		97	495
Total		242	1073

The non-hires can be determined by subtracting the hires from the row totals.

	Hiring Status		Total
Gender	No	Yes	
Female	433	145	578
Male	398	97	495
Total	831	242	1073

You are now ready to produce this table and the corresponding statistical tests using PROC FREQ. A dataset to be utilized by the procedure is needed. The raw data is not available, however all the information needed is provided in the table above. Create a dataset that contains one observation for each cell, the variables will be GENDER, HIRED, and COUNT.

```
DATA HIRE9096 ;
INPUT @1 GENDER $6.
      @8 HIRED $3.
      @12 COUNT 3.
      @ ;
CARDS ;
Female No  433
Female Yes 145
Male   No  398
Male   Yes  97
;
RUN ;
```

Note: Similar code can be used to analyze count data represented in any data table as long as you have the totals for each final cell. If you have additional variables you want to use as stratification variables each unique combination of responses (cell) will be represented by a record in the dataset.

Next you can run PROC FREQ with DATA-HIRE96 utilizing the WEIGHT statement.

```
PROC FREQ DATA=HIRE9096 ;
WEIGHT COUNT ;
TABLES GENDER*HIRED/CHISQ ;
RUN ;
```

GENDER HIRED

Frequency, Percent, Row Pct, Col Pct	No	Yes	Total
Female	433	145	578
	40.35	13.51	53.87
	74.91	25.09	
	52.11	59.92	
Male	398	97	495
	37.09	9.04	46.13
	80.40	19.60	
	47.89	40.08	
Total	831	242	1073
	77.45	22.55	100.00

STATISTICS FOR TABLE OF GENDER BY HIRED

Statistic	DF	Value	Prob
Chi-Square	1	4.602	0.032
Likelihood Ratio Chi-Square	1	4.632	0.031
Continuity Adj. Chi-Square	1	4.293	0.038
Mantel-Haenszel Chi-Square	1	4.598	0.032
Fisher's Exact Test (Left)			0.019
(Right)			0.987
(2-Tail)			0.034

There is a difference in placement rates for the previous 6 years, p=0.032. Bring the output with you when the results are reported. The manager who requested the information assumed the rates were essentially equal.

Conclusions:

This paper walked you through the steps of a what, at first glance, was believed to be a simple analysis. Examples were used to illustrate the use of possible combinations of statements and options available. These can be used to alter the cosmetics of your output, making it not only easier to read but also interpret. These combinations can also produce various statistical tests

for the many types of categorical response data and data structures you may encounter. In addition, the examples were used to attempt to demonstrate the many intricacies of analyzing and interpreting categorical response data.

Hopefully this has encouraged you to find learn more about the many different pairings of data and statistical tests available within PROC FREQ.

Acknowledgments:

Thanks go out to my colleagues in Outcomes Research and Management at Merck: Margaret Coughlin, David Manfredonia, James Murray, and Linda Nelsen. In addition, thanks go out to Janet Stuelpner of ASG, Inc.

References:

London, W. (199), "How to Gain a Working Knowledge of the FREQ Procedure Without Freaking Out," *Proceedings of the Seventeenth Annual SAS Users Group International Conference*, 17, 251-258.

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

SAS Institute Inc. (1990), *SAS/STAT User's Guide Volume 1, Version 6, Fourth Edition*, Cary, NC: SAS Institute Inc.

Author:

Pamela Landsman, MPH
Manager, Outcomes Research & Management
Merck & Co., Inc.
P.O. Box 4, WP39-158
West Point, PA 19486
phone: 215-652-7492
fax: 215-652-0860
email: pamela_landsman@merck.com