Guido's Guide to PROC FREQ – A Tutorial for Beginners Using the SAS® System

Joseph J. Guido, University of Rochester Medical Center, Rochester, NY

ABSTRACT

PROC FREQ is an essential procedure within BASE SAS® used primarily for counting, displaying and analyzing categorical type data. It is such a powerful procedure that you will find it documented not only in BASE SAS but also in SAS®/STAT documentation. This Beginning Tutorial will touch upon both the uses of PROC FREQ in BASE SAS and SAS/STAT. Don't worry, I promise that you do not need a statistical background to understand this procedure. This tutorial will teach you the basics of PROC FREQ and give you a framework to build upon and extend your knowledge of the SAS System.

INTRODUCTION

According to the paper by Carrie Mariner entitled "Answering the Right Question with the Right PROC", PROC FREQ answers the question *How many?*, PROC MEANS answers the question *How much?* and PROC REPORT will answer *Can you produce a report that looks like this?* We are going to answer the question "*How many?"* as we work through some basic PROC FREQ examples in this paper.

The Version 9 SAS® Procedure Manual states, "The FREQ procedure produces one-way to *n*-way frequency and cross tabulation (contingency) tables. For two-way tables, PROC FREQ computes tests and measures of association. For *n*-way tables, PROC FREQ does stratified analysis, computing statistics within, as well as across, strata. Frequencies and statistics can also be output to SAS data sets."

We will begin with the very basics and consider the one-way frequency tables. These are used by SAS Programmers and Analysts all the time without giving the matter a second thought. What I mean is that if one is counting, doing error checking of the data or categorizing data, then PROC FREQ is the usual choice. The variables in the dataset can be either character or numeric.

The following statements are used in PROC FREQ according to the SAS® Procedure Manual:

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

I have underlined the 4 statements in PROC FREQ which I will be discussing in this paper. The PROC FREQ statement is the only required statement for the FREQ procedure. If you specify the following statements, PROC FREQ produces a one-way frequency table for each variable in the most recently created data set.

```
PROC FREQ;
RUN;
```

DISCUSSION

Using the dataset Trial.sas7bdat from the Glenn Walker book "Common Statistical Methods for Clinical Research with SAS® Examples" used in the SAS courses that I teach will illustrate an example. In this fictitious dataset there are 100 patients, and we want to know how many are males and how many are females. We also want to know how many males are over age 55.

Notice that the two questions ask how many and so we know that PROC FREQ is the procedure of choice. We begin by asking SAS for frequencies (one-way) on each of the variables of interest SEX and AGE. On the TABLES statement we list both variables separated by a space.

Example 1

```
PROC FREQ Data=Trial;
     TABLES Sex Age;
RUN;
```

[See Example 1 Output in Appendix]

While these results are informative they do not give us the desired end which is "How many males are over the age of 55". We have the number of males and females and we have the age distribution but we don't have an answer to our question.

We know there are 44 males and we know that there are 18 patients over the age of 55.
So we decide to use a WHERE statement to select only those patients who are male.

Example 2

```
PROC FREQ Data=Trial;
     TABLES Age;
     WHERE Sex='M';
RUN;
```

[See Example 2 Output in Appendix]

So while we can answer the questions about how many males are over the age of 55 (there are 8 males over the age of 55), we may want to reshape our data into meaningful groupings in case there are other similar questions that arise at a later time (e.g. How many females are aged 46-55?). So we will create groupings of the ages using a PROC FORMAT statement.

Example 3

```
PROC FORMAT;
VALUE Age_Fmt
          Low-15='Less than 16 years'
          16-25='16 – 25 years'
          26-35='26 – 35 years'
          36-45='36 – 45 years'
          46-55='46 – 55 years'
          56-High='Over 55 years';
 RUN;
```

Note that each original group of values used to create a format (called a range) is found on the left of the equal sign; and on the right of the equal sign we assign text that describes the group. The '–' and 'Low' and 'High' symbols in the original values need some explanation. To use these, note that any value next to a '–' symbol (called a range indicator) is included in the range. The keyword 'Low' means the lowest numeric value in a given dataset (which may be less than zero or missing). The keyword 'High' means the highest numeric value in a given dataset.

Once we have our format created, we need to apply the format we created with a FORMAT statement in PROC FREQ. We also add a LABEL statement to further describe the variables in our report. Finally, rather than limiting the report to only Males, we create a two-dimensional table using the asterisk between the Age and Sex variables. Because PROC FREQ will by default add several unwanted statistics to the table when we define two dimensions, we also add the NOCOL, NOROW, and PERCENT options to remove those statistics.

Example 4

```
PROC FREQ Data=Trial;
     TABLES Age*Sex / nocol norow nopercent;
     FORMAT Age Age_Fmt.;
     LABEL Age='Age of Patient'
           Sex='Sex of Patient';
RUN;
```

[See Example 4 Output in Appendix]

Well believe it or not we actually got through one-way and two-way tables with PROC FREQ based on Example 4 above. To take things one step further (stratification) we introduce the three-way table. In this case we may want to know the output of Example 4 above by each Center in our study. There are actually several different ways to accomplish this task and I am going to demonstrate the two most common.

In the first case we can simple add another "dimension" to our TABLES statement to accomplish this task. With a two-way table the first dimension is the ROW and the second dimension is the COLUMN. If we have a three-way crosstab then the first dimension is the STRATA, the second dimension is the ROW and the third dimension is the COLUMN.

Example 5

```
PROC FREQ Data=Trial;
     TABLES Center*Age*Sex / nocol norow nopercent;
     FORMAT Age Age_Fmt.;
     LABEL Age='Age of Patient'
           Sex='Sex of Patient'
           Center='Study Center';
RUN;
```

[See Example 5 Output in Appendix]

Another way to accomplish the same thing as Example 5 is to use a BY statement. However when we use a BY statement we must ALWAYS sort the SAS dataset by the Key Variable or Variables. In this case we would sort the dataset "Trial" by the variable "Center".

Example 6

```
PROC SORT Data=Trial Out=TrialSorted;
     BY Center;
RUN;
```

Now that we have sorted the dataset Trial by the variable Center and created a new dataset called TrialSorted using the Out= option on PROC SORT we are ready to proceed.

Example 7

```
PROC FREQ Data=TrialSorted;
     TABLES Age*Sex / nocol norow nopercent;
     FORMAT Age Age_Fmt.;
     LABEL Age='Age of Patient'
           Sex='Sex of Patient'
           Center='Study Center';
     BY Center;
RUN;
```

[See Example 7 Output in Appendix]

You will note that while Example 5 and Example 7 give the same numeric results, the appearance is slightly different. Example 7 puts each table on a separate page and begins with a dashed line followed by Study Center = N where N is 1, 2 or 3 and then the dashed line continues.

Now we are ready for some simple statistical computations using PROC FREQ. We will look at the calculation of the Chi-square statistic and McNemar's statistic. The first statistic is used on independent groups in the data (Males and Females). We compare these two groups by a variable called RESP which indicates response (0=No, 1=Yes).
Without even knowing much about statistics we conclude that there is a statistically significant difference between the response variable for men and women. This is determined because the Chi-square statistic of 5.18 has an associated probability (p-value) of 0.0228. This means that there is a less than 1/20 chance that this finding is due to chance alone.

Example 8

```
PROC FREQ Data=Trial;
     TABLES Sex*Resp / CHISQ;
     LABEL Sex='Sex of Patient'
           Resp='Response of Patient';
RUN;
```

[See Example 8 Output in Appendix]

Now let's say that we have some additional data for this fictitious group of 100 patients. They are asked their opinion about a certain experimental procedure. Then we introduce an intervention (an educational program). Finally we survey the SAME 100 patients after the intervention (say in 6 weeks). We want to know if there has been a significant change or shift in the patient's original opinion (yes or no).

We cannot use the CHISQ option with PROC FREQ because these are not independent groups. They are in fact the same 100 patients who we surveyed initially and then at 6 weeks. The SAS System has an option in PROC FREQ to handle this type of analysis. It is the AGREE option. There is one more twist here because we collected these new data after the original Trial dataset was created we need to either add the individual data points to Trial or we can create a dataset using the aggregate number for our two new pieces of data and then use the WEIGHT statement in PROC FREQ to analyze.

So let's create a new dataset called Trial_aggr based on the counts of the data. We will be creating a 2 x 2 table as follows; 30 patients answered Yes initially and at follow-up,
10 patients answered Yes initially and then No at follow-up. Forty (40) patients answered No initially and then Yes at follow-up and 20 patients answer No initially and then No at follow-up.

Example 9

```
DATA Trial_aggr;
INPUT Pre $ Post $ Count;
DATALINES;
Yes Yes 30
Yes No  10
No  Yes 40
No  No  20
;
RUN;
```

Now we are ready to run PROC FREQ on our newly constructed dataset Trial_aggr. You will notice that the code or syntax for Example 8 and Example 10 look similar except that we have replaced the CHISQ option with the AGREE option since we are working with data that is not independent (actually called paired data). Also the WEIGHT statement is used since without it SAS would report that we only had 4 observations (instead of 100 observations).

Example 10

```
PROC FREQ Data=Trial_aggr;
     TABLES Pre*Post / AGREE;
     LABEL Pre='Initial Response of Patient'
           Post='Response of Patient after Intervention';
WEIGHT Count;
RUN;
```

[See Example 10 Output in Appendix]

We have completed our Tutorial and now the rest is up to you. The best ways to improve your SAS skills are to practice, practice, and practice. The SAS Online Help facility and SAS manuals are excellent ways to do this. Both are available to you under the Help dropdown (Learning SAS Programming and SAS Help and Documentation).

CONCLUSION

PROC FREQ is a very powerful but simple and necessary procedure in SAS. This Beginning Tutorial has just scratched the surface of the functionality of PROC FREQ. The author's hope is that these several basic examples will serve as a guide for the user to extend their knowledge of PROC FREQ and experiment with other uses for their specific data needs.

REFERENCES

SAS Institute, Inc. (2002). *Base SAS® 9 Procedures Guide*. Cary, NC: SAS Institute, Inc.

Walker, Glenn A. (2002). "Common Statistical Methods for Clinical Research with SAS® Examples", 2nd Edition, SAS Institute: Cary, NC.

Mariner, Carrie (2006). "Answering the Right Question with the Right PROC", SouthEast SAS Users Group 2006 Proceedings. The Scholars Digital Library of Analytics (Beta) http://www8.sas.com/scholars/Proceedings/2006/HOW/HW01_06.PDF

ACKNOWLEDGEMENTS

CONTACT INFORMATION

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

Joseph J Guido, MS
University of Rochester Medical Center
Department of Community and Preventive Medicine
Division of Social and Behavioral Medicine
120 Corporate Woods, Suite 350
Rochester, New York 14623
Phone: (585) 758-7818
Fax: (585) 424-1469
Email: Joseph_Guido@urmc.rochester.edu

APPENDIX

Example 1

SEX				
SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	56	56.00	56	56.00
M	44	44.00	100	100.00

AGE				
AGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
19	2	2.00	2	2.00
21	1	1.00	3	3.00
24	2	2.00	5	5.00
26	2	2.00	7	7.00
27	4	4.00	11	11.00
28	3	3.00	14	14.00
29	1	1.00	15	15.00
30	1	1.00	16	16.00
31	3	3.00	19	19.00
32	5	5.00	24	24.00
33	2	2.00	26	26.00
34	2	2.00	28	28.00
35	2	2.00	30	30.00
36	4	4.00	34	34.00
37	3	3.00	37	37.00
38	3	3.00	40	40.00
39	3	3.00	43	43.00
40	2	2.00	45	45.00
41	4	4.00	49	49.00
42	6	6.00	55	55.00
43	1	1.00	56	56.00
44	3	3.00	59	59.00
45	4	4.00	63	63.00
46	1	1.00	64	64.00
47	2	2.00	66	66.00
48	3	3.00	69	69.00
49	1	1.00	70	70.00
50	2	2.00	72	72.00
51	5	5.00	77	77.00
52	2	2.00	79	79.00
53	1	1.00	80	80.00
54	1	1.00	81	81.00
55	1	1.00	82	82.00

AGE				
AGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
56	2	2.00	84	84.00
57	3	3.00	87	87.00
58	1	1.00	88	88.00
59	2	2.00	90	90.00
61	2	2.00	92	92.00
62	2	2.00	94	94.00
63	1	1.00	95	95.00
64	2	2.00	97	97.00
65	1	1.00	98	98.00
69	1	1.00	99	99.00
70	1	1.00	100	100.00

Example 2

AGE				
AGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
19	1	2.27	1	2.27
24	1	2.27	2	4.55
26	1	2.27	3	6.82
27	2	4.55	5	11.36
29	1	2.27	6	13.64
31	2	4.55	8	18.18
32	2	4.55	10	22.73
36	2	4.55	12	27.27
37	3	6.82	15	34.09
38	1	2.27	16	36.36
39	2	4.55	18	40.91
40	2	4.55	20	45.45
41	1	2.27	21	47.73
42	3	6.82	24	54.55
44	1	2.27	25	56.82
45	2	4.55	27	61.36
46	1	2.27	28	63.64
47	1	2.27	29	65.91
48	1	2.27	30	68.18
49	1	2.27	31	70.45
51	2	4.55	33	75.00
52	2	4.55	35	79.55
55	1	2.27	36	81.82
56	1	2.27	37	84.09
58	1	2.27	38	86.36
59	1	2.27	39	88.64
61	2	4.55	41	93.18
62	1	2.27	42	95.45
65	1	2.27	43	97.73
70	1	2.27	44	100.00

Example 4

Table of AGE by SEX			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	3	2	5
26 - 35 years	17	8	25
36 - 45 years	16	17	33
46 - 55 years	10	9	19
Over 55 years	10	8	18
Total	56	44	100

Example 5

Table 1 of AGE by SEX			
Controlling for CENTER=1			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	2	1	3
26 - 35 years	6	5	11
36 - 45 years	7	4	11
46 - 55 years	6	4	10
Over 55 years	3	2	5
Total	24	16	40

Table 2 of AGE by SEX			
Controlling for CENTER=2			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	1	1	2
26 - 35 years	7	2	9
36 - 45 years	7	8	15
46 - 55 years	3	2	5
Over 55 years	2	2	4
Total	20	15	35

Table 3 of AGE by SEX			
Controlling for CENTER=3			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	0	0	0
26 - 35 years	4	1	5
36 - 45 years	2	5	7
46 - 55 years	1	3	4
Over 55 years	5	4	9
Total	12	13	25

Example 7

Study Center = 1

Table of AGE by SEX			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	2	1	3
26 - 35 years	6	5	11
36 - 45 years	7	4	11
46 - 55 years	6	4	10
Over 55 years	3	2	5
Total	24	16	40

Study Center = 2

Table of AGE by SEX			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
16 - 25 years	1	1	2
26 - 35 years	7	2	9
36 - 45 years	7	8	15
46 - 55 years	3	2	5
Over 55 years	2	2	4
Total	20	15	35

Study Center = 3

Table of AGE by SEX			
AGE(Age of Patient)	SEX(Sex of Patient)		
Frequency	F	M	Total
26 - 35 years	4	1	5
36 - 45 years	2	5	7
46 - 55 years	1	3	4
Over 55 years	5	4	9
Total	12	13	25

Example 8

Table of SEX by RESP			
SEX(Sex of Patient)	RESP(Response of Patient)		
Frequency Percent Row Pct Col Pct	0	1	Total
F	17 17.00 30.36 77.27	39 39.00 69.64 50.00	56 56.00
M	5 5.00 11.36 22.73	39 39.00 88.64 50.00	44 44.00
Total	22 22.00	78 78.00	100 100.0

Statistic	DF	Value	Prob
Chi-Square	1	5.180	.0228
Likelihood Ratio Chi-Square	1	5.472	.0193
Continuity Adj. Chi-Square	1	4.132	.0421
Mantel-Haenszel Chi-Square	1	5.128	.0235
Phi Coefficient		0.227	
Contingency Coefficient		0.221	
Cramer's V		0.227	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	17
Left-sided Pr <= F	0.995
Right-sided Pr >= F	0.019
Table Probability (P)	0.014
Two-sided Pr <= P	0.028

Sample Size = 100

Example 10

Table of Pre by Post			
Pre(Initial Response of Patient)	**Post(Response of Patient after Intervention)**		
Frequency Percent Row Pct Col Pct	**Yes**	**No**	**Total**
Yes	30 30.00 75.00 42.86	10 10.00 25.00 33.33	40 40.00
No	40 40.00 66.67 57.14	20 20.00 33.33 66.67	60 60.00
Total	70 70.00	30 30.00	100 100.0

Statistics for Table of Pre by Post

McNemar's Test	
Statistic (S)	18.0000
DF	1
Pr > S	<.0001

Simple Kappa Coefficient	
Kappa	0.074
ASE	0.081
95% Lower Conf Limit	-0.086
95% Upper Conf Limit	0.234

Sample Size = 100