What's New in PROC FREQ Procedure, Version 9.

Jyotheeswara Naidu Yellanki, HTC Global Services Inc., Bloomington, IL
Raj Suligavi, HTC Global Services Inc., Bloomington, IL

ABSTRACT:
The FREQ procedure is a descriptive procedure as well as a statistical procedure. It produces one-way counts when you have one variable and 2-way counts when you have two variables and so on. You can use the FREQ procedure to create crosstabulation tables that summarize data for two or more categorical variables by showing the number of observations for each combination of variable values. Frequency distributions work best with variables whose values can be described as categorical or discrete, and whose values are best summarized by counts rather than by averages. There are several options in PROC FREQ to control the type of output produced. SAS version 9 has come up with nice useful options. This paper will investigate and describe with examples how and where to use these options and what are their pros and cons. New options that are associated with PROC FREQ are discussed below.

- NLEVELS options in PROC FREQ statement.
- ZEROS option in WEIGHT statement.
- CONTENTS, CROSSLIST, NOSPARSE and NOWARN options in TABLES statement.

INTRODUCTION:
The FREQ procedure can include many statements and options for controlling frequency output. This paper will start with basic syntax, using the new options in one-dimensional table and two-dimensional table. For simplicity this presentation will limit just to one way and two-way tables.

OVERVIEW:
The general syntax of FREQ procedures is:
PROC FREQ <options>;
    BY <descending> variable-1 <..<descending> variable-n><NOTSORTED>;
    EXACT statistic-options </computation-option>;
    OUTPUT <OUT=SAS-data-set>options;
    TABLES requests </options>;
    TEST options;
    WEIGHT variables </option>;

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

    Proc freq;
    Run;

NEW OPTIONS:

NLEVELS:
This option in PROC FREQ statement is used to display the "Number of Variable Levels" . By using this options we can get the distinct count for each variable listed in the TABLES statement. If TABLES statement is omitted, by default it will display for all variables in that dataset.

For example, if you run the following code:

```
PROC FREQ DATA=SASHELP.CLASS NLEVELS;
TABLES SEX;
RUN;
```

Then the resulting output would look like this:

```
NUMBER OF VARIABLE LEVELS

VARIABLE        LEVELS          ---- this is added output because
-------------------                  of NLEVELS option
SEX              2


                                CUMULATIVE    CUMULATIVE
SEX     FREQUENCY     PERCENT    FREQUENCY      PERCENT
------------------------------------------------------
F            9        47.37          9           47.37
M           10        52.63         19          100.00
```

This will produce one-way frequency table for the variable SEX and also "Number of Variable Levels" for the same variable SEX.

Suppose if you are interested in only "Number of Variable Levels" for the variable, then you can use NOPRINT option in TABLES statement.

```
PROC FREQ DATA=SASHELP.CLASS NLEVELS;
TABLES SEX / NOPRINT;
RUN;
```

```
NUMBER OF VARIABLE LEVELS

VARIABLE        LEVELS
-------------------
SEX              2
```

Some times you might be interested to know the number of levels for all character variables or all numeric variables in that data set with out printing there frequency table.

```
PROC FREQ DATA=SASHELP.CLASS NLEVELS;
TABLES _CHAR_  / NOPRINT;
RUN;
```

```
NUMBER OF VARIABLE LEVELS

VARIABLE        LEVELS
-------------------
NAME            19
SEX              2
```

Similarly you can use

```
TABLES _ NUMERIC _  / NOPRINT;      or
TABLES VAR1--VAR2  / NOPRINT;
```

It is best to use these options to know the number of levels of variable before printing the frequency table if the data set is too large in size and if you unaware of what could be distinct count for each variable.

ZEROS:
ZEROS: Used in WEIGHT statement to include the observations with zero weight values. By default, PROC FREQ ignores observations with zero weights. The frequency and crosstabulation tables will display any levels that corresponding to observations with zero weights. Without the ZEROS option, PROC FREQ does not process observations with zero weights, and so does not display levels that contain only observations with zero weights.

For example a company has 3 business units and each unit quarterly gain/loss(GA_LO) is depicted in the dataset below.

```
DATA DS;
INPUT  BUSINESS $  QUT $  GA_LO @@;
CARDS;
BUS1 Q1 50  BUS2 Q1 60  BUS3 Q1 0
BUS1 Q2 30  BUS2 Q2 30  BUS3 Q2 0
BUS1 Q3 75  BUS2 Q3 15  BUS3 Q3 0
BUS1 Q4 25  BUS2 Q4 25  BUS3 Q4 0
;
RUN;
```

The WEIGHT statement in PROC FREQ will produce the yearly total gain/loss for each business.

```
PROC FREQ DATA=DS;
WEIGHT GA_LO ;
TABLES BUSINESS;
RUN;
```

BUSINESS	FREQUENCY	PERCENT	CUMULATIVE FREQUENCY	CUMULATIVE PERCENT
BUS1	180	58.06	180	58.06
BUS2	130	41.94	310	100.00

Notice that the absence of the BUS3 with the zero count. If you use ZEROS option in the WEIGHT statement then it will show in the output otherwise not .

```
PROC FREQ DATA=DS;
WEIGHT GA_LO / ZEROS;
TABLES BUSINESS;
RUN;
```

BUSINESS	FREQUENCY	PERCENT	CUMULATIVE FREQUENCY	CUMULATIVE PERCENT
BUS1	180	58.06	180	58.06
BUS2	130	41.94	310	100.00
BUS3	0	0.00	310	100.00

CONTENTS:
CONTENTS=link-text: Used in TABLES statement to specify the meaningful text(on the hyper link) for the HTML contents file that links to crosstabulation tables. The CONTENTS= option affects only the HTML contents file, and not the HTML body file. If you omit this option, by default, the HTML link text for crosstabulation tables is "Cross-Tabular Freq Table". This option is useful to specify the different text for different crosstabulationtable links.

The following example shows the output without CONTENTS options. By default it will show 'CROSS-TABULAR FREQ TABLE' in contents page for both 2 way table SEX * AGE and SEX * WEIGHT.

```
PROC FORMAT;
VALUE PHYSIC
        0 - 85  = 'LOW'
       85 - 110 = 'AVERAGE'
      111 - 180 = 'HIGH';
RUN;
```

```
ODS HTML
   BODY     = 'DATA.HTM'
   CONTENTS = 'TOC.HTM'
   FRAME    = 'FRAME.HTM'
   PATH= 'Path name'
   STYLE=DEFAULT;

 PROC FREQ  DATA=SASHELP.CLASS;
  TABLES SEX * AGE;
  TABLES SEX * WEIGHT;
  FORMAT WEIGHT PHYSIC.;
 RUN;
 ODS HTML CLOSE;
```

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Sex by Age

Sex	Age 11	12	13	14	15	16	Total
F	1 5.26 11.11 50.00	2 10.53 22.22 40.00	2 10.53 22.22 66.67	2 10.53 22.22 50.00	2 10.53 22.22 50.00	0 0.00 0.00 0.00	9 47.37
M	1 5.26 10.00 50.00	3 15.79 30.00 60.00	1 5.26 10.00 33.33	2 10.53 20.00 50.00	2 10.53 20.00 50.00	1 5.26 10.00 100.00	10 52.63
Total	2 10.53	5 26.32	3 15.79	4 21.05	4 21.05	1 5.26	19 100.00

Frequency Percent Row Pct Col Pct

Table of Sex by Weight

Sex	Weight Low	Average	High	Total
F	4 21.05 44.44 57.14	3 15.79 33.33 60.00	2 10.53 22.22 28.57	9 47.37
M	3 15.79 30.00 42.86	2 10.53 20.00 40.00	5 26.32 50.00 71.43	10 52.63
Total	7 36.84	5 26.32	7 36.84	19 100.00

By using the option CONTENTS in TABLES statement we can produce the desired or meaningful link-text in the contents page.

```
ODS HTML
   BODY     = 'DATA.HTM'
   CONTENTS = 'TOC.HTM'
   FRAME    = 'FRAME.HTM'
   PATH= 'Path name'
   STYLE=DEFAULT;

 PROC FREQ  DATA=SASHELP.CLASS;
  TABLES SEX * AGE    / CONTENTS= 'First freq table for Sex vs Age';
  TABLES SEX * WEIGHT / CONTENTS = 'Second freq table for Sex vs WEIGHT';
  FORMAT WEIGHT PHYSIC.;
 RUN;
 ODS HTML CLOSE;
```

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Sex by Age						
		Age					
Sex	11	12	13	14	15	16	Total
F	1 5.26 11.11 50.00	2 10.53 22.22 40.00	2 10.53 22.22 66.67	2 10.53 22.22 50.00	2 10.53 22.22 50.00	0 0.00 0.00 0.00	9 47.37
M	1 5.26 10.00 50.00	3 15.79 30.00 60.00	1 5.26 10.00 33.33	2 10.53 20.00 50.00	2 10.53 20.00 50.00	1 5.26 10.00 100.00	10 52.63
Total	2 10.53	5 26.32	3 15.79	4 21.05	4 21.05	1 5.26	19 100.00

Frequency Percent Row Pct Col Pct	Table of Sex by Weight			
		Weight		
Sex	Low	Average	High	Total
F	4 21.05 44.44 57.14	3 15.79 33.33 60.00	2 10.53 22.22 28.57	9 47.37
M	3 15.79 30.00 42.86	2 10.53 20.00 40.00	5 26.32 50.00 71.43	10 52.63
Total	7 36.84	5 26.32	7 36.84	19 100.00

Links to all crosstabulation tables produced by a single/multiple TABLES statement use the same text. To specify different text for different crosstabulation table links request the tables in separate TABLES statements and use the CONTENTS= option in each TABLES statement.

NOTE: The CONTENTS= option applies only to cross-tabulation tables. *It will not work for one-way table*.

CROSSLIST:
CROSSLIST: Used in TABLES statement to display the crosstabulation tables in column format instead of the default crosstabulation cell format. *You cannot specify both the LIST option and the CROSSLIST option in the same TABLES statement.*

The following example with output shows effect of CROSSLIST option in the TABLES statement.

```
PROC FREQ  DATA=SASHELP.CLASS;
TABLES SEX * AGE / CROSSLIST;
RUN;
```

```
                    TABLE OF SEX BY AGE

                                       ROW       COLUMN
SEX        AGE    FREQUENCY    PERCENT  PERCENT   PERCENT
------------------------------------------------------------
F          11         1          5.26    11.11     50.00
           12         2         10.53    22.22     40.00
           13         2         10.53    22.22     66.67
           14         2         10.53    22.22     50.00
           15         2         10.53    22.22     50.00
           16         0          0.00     0.00      0.00

           TOTAL      9         47.37   100.00
------------------------------------------------------------
M          11         1          5.26    10.00     50.00
           12         3         15.79    30.00     60.00
           13         1          5.26    10.00     33.33
           14         2         10.53    20.00     50.00
           15         2         10.53    20.00     50.00
           16         1          5.26    10.00    100.00

           TOTAL     10         52.63   100.00
------------------------------------------------------------
TOTAL      11         2         10.53             100.00
           12         5         26.32             100.00
           13         3         15.79             100.00
           14         4         21.05             100.00
           15         4         21.05             100.00
           16         1          5.26             100.00

           TOTAL     19        100.00
------------------------------------------------------------
```

The CROSSLIST option looks the same as LIST option, but there are few differences as shown below.

	CROSSLIST	LIST
Statistics	FREQUENCY PERCENT ROW PERCENT COLUMN PERCENT	FREQUENCY PERCENT CUMULATIVE FREQUENCY CUMULATIVE PERCENT
Totals	Produces sub totals and final totals.	No totals
Rows with zero freq	Display the variable levels with zero frequencies	Suppress the variable levels with zero frequencies

NOSPARSE:

By default for CROSSLIST tables, as shown in above example, PROC FREQ displays all levels of the column variable within each level of the row variable, including any column variable levels with zero frequency for that row. For CROSSLIST tables, the NOSPARSE option suppresses display of variable levels with zero frequency.

```
PROC FREQ  DATA=SASHELP.CLASS;
TABLES SEX * AGE / CROSSLIST NOSPARSE;
RUN;
```

```
                     TABLE OF SEX BY AGE

                                        ROW       COLUMN
SEX        AGE    FREQUENCY    PERCENT   PERCENT   PERCENT
-------------------------------------------------------------
F          11         1          5.26     11.11     50.00
           12         2         10.53     22.22     40.00
           13         2         10.53     22.22     66.67
           14         2         10.53     22.22     50.00
           15         2         10.53     22.22     50.00

           TOTAL      9         47.37    100.00
-------------------------------------------------------------
M          11         1          5.26     10.00     50.00
           12         3         15.79     30.00     60.00
           13         1          5.26     10.00     33.33
           14         2         10.53     20.00     50.00
           15         2         10.53     20.00     50.00
           16         1          5.26     10.00    100.00

           TOTAL     10         52.63    100.00
-------------------------------------------------------------
TOTAL      11         2         10.53              100.00
           12         5         26.32              100.00
           13         3         15.79              100.00
           14         4         21.05              100.00
           15         4         21.05              100.00
           16         1          5.26              100.00

           TOTAL     19        100.00
-------------------------------------------------------------
```

NOWARN:

Used in TABLES statement to suppress the log warning message that the asymptotic chi-square test may not be valid. By default , PROC FREQ displays this log message when more than 20 percent of the table cells have expected frequencies less than five.

```
PROC FREQ DATA=DS_NAME;
TABLES  VAR1 * VAR2 / CHISQ NOWARN;
RUN;
```

Will suppress the log warning message that the asymptotic chi-square test may not be valid, when more than 20 percent of the table cells have expected frequencies less than five.

CONCLUSION:

PROC FREQ is the most powerful procedure for statistical analysis. The new options have given an added advantage to the PROC FREQ. Other options like BDT used in the TABLES statement is out the scope of this paper. If details are needed, please refer SAS Documentation.

REFERENCES:

SAS Institute Inc., SAS OnlineDoc ® 9, http://v9doc.sas.com/sasdoc/
SAS Institute Inc., Base SAS 9.1 ® Procedures Guide.
SAS Online Proceedings ® http://support.sas.com/usergroups/sugi/proceedings/index.html

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

Jyotheeswara Naidu Yellanki
Ph: 309-662-5163
Email: yjnaidu@hotmail.com

Raj Suligavi
Ph: 309-663-0858
Email: bsuliga@yahoo.com