Paper 056-2012

Increase your OUTPUT with PROC MEANS and PROC FREQ

Julie Kezik MS, Melissa Hill, MPH, Yale University, New Haven, CT, USA

Abstract

How many people have 'x and y'? How many people are in category 'z'? The questions are endless, but the solutions are not! Using an output statement in PROC MEANS or PROC FREQ can easily answer those inquiries.

The purpose of this paper is to utilize the OUTPUT statement in a PROC MEANS or PROC FREQ to get simple statistics for a selected variable or group of variables. The resulting SAS® dataset can be used for future analyses and/or be easily exported to create graphical or tabular displays.

Problem

With large datasets it is often hard to tease out necessary simple statistics that analysts are required to produce on a regular basis. Using PROC MEANS and PROC FREQ can be solutions to frequently asked questions and reporting queries.

The Question

We are asked to find simple statistics on the age distribution of children with allergies, stratified by gender.
Let's first take a look at the variables, we can utilize the means and frequency procedures to do the investigating.

PROC MEANS analyzes the dataset according to the variable listed in the VAR statement and is a simple way to see standard statistics of continuous variables. The syntax below requests statistics for child age (ch_age).

```
proc means data =tmp1.tablevars;
var ch_age ;run;
```

Analysis Variable : CH_AGE				
N	Mean	Std Dev	Minimum	Maximum
1233	7.390916	1.697873	4.0000000	11.000000

Output 1. Proc Means

Output 1 includes the N, Mean, Standard Deviation, Maximum and Minimum of child's age for the dataset "tablevars".

PROC FREQ analyzes the dataset according to variables listed in the TABLES statement; this is an efficient way to see a snapshot of categorical variables. The following frequency procedure syntax includes a TABLES statement for two variables, gender and allergic status.

```
proc freq data = tmp1.tablevars;
tables gender allergy; run;
```

child's sex				
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 =Male	726	58.88	726	58.88
2 =Female	507	41.12	1233	100.00

ALLERGY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0 =No	423	34.31	423	34.31
1 = Yes	810	65.69	1233	100.00

Output 2. Proc Freq

Output 2 displays the frequency and percent values, as well as indicating to us that there are 810 children with Allergy = 1 ('Yes'). This is a value we will need to carry along to keep children with allergies in our analyses.

It looks like these variables are practical to work with, let's create an output dataset to answer the final question.

```
proc sort data =tmp1.tablevars;
by gender; run;
proc means data= tmp1.tablevars noprint;
var ch_age; by gender ;
where allergy = 1;
output out = Boss1; run;
```

Don't forget when using a BY variable in the PROC MEANS statement, make sure to use a PROC SORT first!

By adding a WHERE statement to PROC MEANS we are able to limit our dataset to only children with allergies. The BY statement allows another column to be added in our analysis dataset so not only are we looking at summary statistics for age, but stratifying those statistics by gender. The output statement sends the created variables into a new SAS dataset named Boss1.

Obs	GENDER	_TYPE_	_FREQ_	_STAT_	CH_AGE
1	1	0	493	N	493.000
2	1	0	493	MIN	4.000
3	1	0	493	MAX	11.000
4	1	0	493	MEAN	7.505
5	1	0	493	STD	1.713
6	2	0	317	N	317.000
7	2	0	317	MIN	4.000
8	2	0	317	MAX	11.000
9	2	0	317	MEAN	7.599
10	2	0	317	STD	1.669

Output 3. Proc Means dataset "Boss1"

Output 3, PROC MEANS includes a "_STAT_" column, containing standard PROC MEANS statistics. We now can see that the mean age of a male with allergies is 7.5 and the mean age of a female with allergies is 7.6.

```
proc freq data = tmp2.tablevars noprint;
tables ch_age*gender/norow nocol out=Boss2;
where allergy = 1; run;
```

The above PROC FREQ syntax requests a table of child age by gender, within the same statement we have included 'out =' which will output the requested statistics in to the new dataset Boss2. Again, adding a WHERE statement allows us to only include children with allergies.

Obs	CH_AGE	GENDER	COUNT	PERCENT
1	4	1	1	0.1235
2	5	1	85	10.4938
3	6	1	77	9.5062
4	7	1	71	8.7654
5	8	1	100	12.3457
6	9	1	79	9.7531
7	10	1	79	9.7531
8	11	1	1	0.1235
9	4	2	1	0.1235
10	5	2	40	4.9383
11	6	2	55	6.7901
12	7	2	56	6.9136
13	8	2	59	7.2840
14	9	2	50	6.1728
15	10	2	55	6.7901
16	11	2	1	0.1235

Output 4. Proc Freq dataset "Boss2"

Output 4, PROC FREQ, displays a distribution of the data with summary statistics. It contains the column “percent”, which refers to the percentage of each age, stratified by gender that is included in the allergy analysis. This output, although informative, is not optimal for reporting. Displaying the data in a graph will make viewing the dataset BOSS2 much less complicated.

Results

Now that we have created a dataset with summary variables there are several ways to display the dataset BOSS1. Using ODS in combination with a simple PROC PRINT is an expedient way to get the output dataset from a PROC MEANS into a tabular display.

```
ods rtf file = 'tab_disp.rtf';
proc print data = boss1;
where _STAT_ = "MEAN";
run;
```

“Mean Age of Children with Allergies, Stratified by Gender”

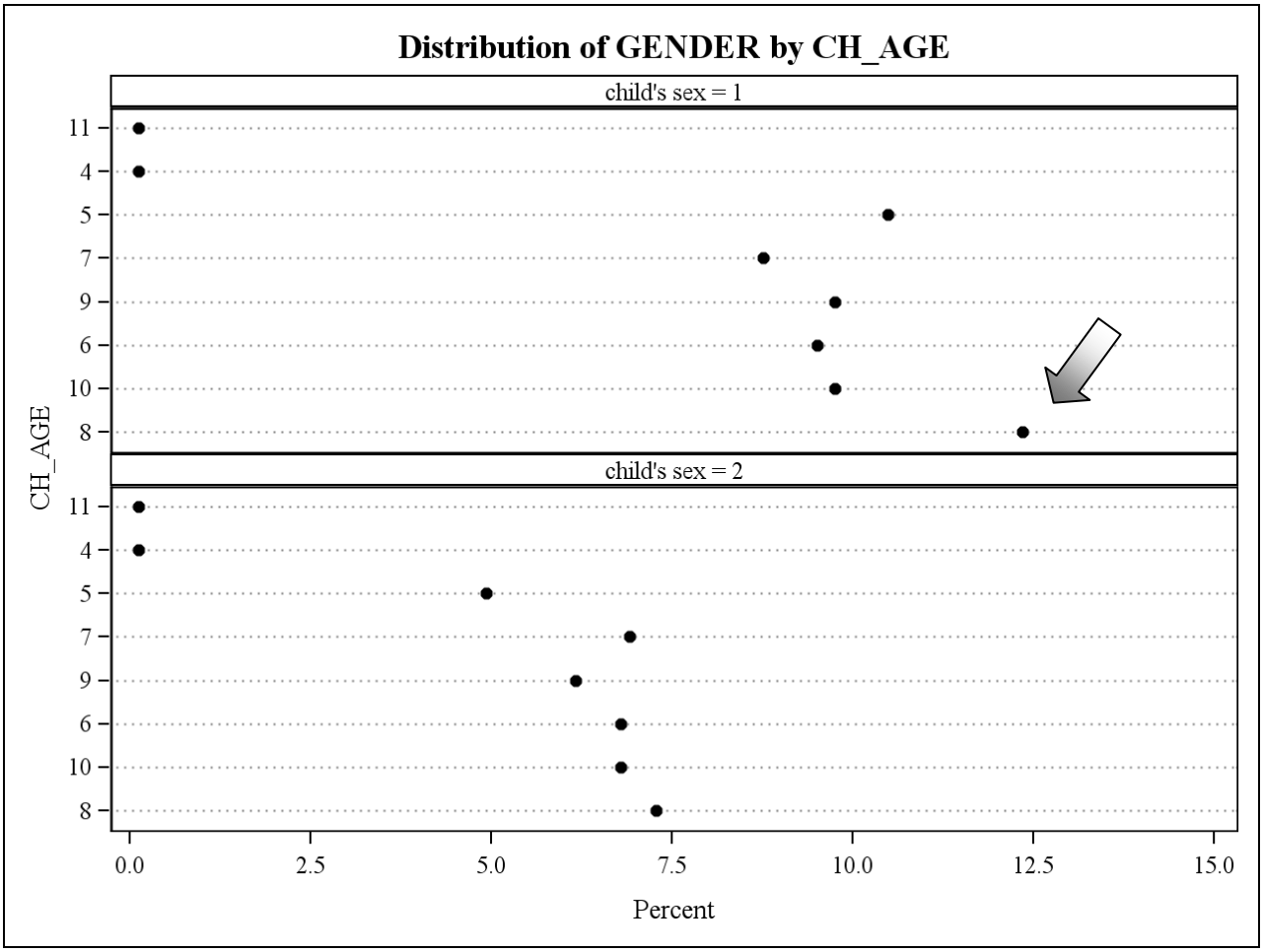
Gender	Frequency	Mean Age
Male	493	7.50507
Female	317	7.59937

Output 5. Using ODS with proc print

Output 5 is an example of edited ODS output, the table was exported into a word document (tab_disp.rtf). In this format the document can be easily edited and ready for distribution in minutes.

There are also several options to graphically display the data, one way, is to utilize PROC FREQ in combination with ODS to create a frequency dot plot which clearly shows the distribution of age by gender for the allergic children.

```
ods rtf file =graph_disp.rtf';
ods graphics on;
proc freq data = boss2 order=freq;
tables gender*ch_age/plots=freqplot
(type=dot scale =percent);
weight count;
run;
```



Graph 1. Frequency Plot – Distribution of Allergic Children by Gender and Age

Graph 1 indicates by gender, the percent of each age with allergies. Clearly illustrating the highest overall percentage of children with allergies is in the category of eight year old males.

Conclusion

PROC MEANS and PROC FREQ offer an OUTPUT statement which can be utilized to create datasets with simple statistics and produce graphical or tabular displays. Using the OUTPUT statement can be an ally in everyday programming.

CONTACT INFORMATION

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

Name: Julie Kezik, MS
Enterprise: Yale University Center for Perinatal, Pediatric and Environmental Epidemiology
Address: One Church Street, 6th Floor
City, State ZIP: New Haven, CT 06510
Work Phone: 203-764-9375
Fax: 203-764-9378
E-mail: julie.kezik@yale.edu
Web: