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

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

ABSTRACT

PROC MEANS is a basic procedure within BASE SAS® used primarily for answering questions about quantities (How much?, What is the average?, What is the total?, etc.) It is the procedure that I use second only to PROC FREQ in both data management and basic data analysis. PROC MEANS can also be used to conduct some basic statistical analysis. This beginning tutorial will touch upon many of the practical uses of PROC MEANS and some helpful tips to expand one's knowledge of numeric type data and give a framework to build upon and extend your knowledge of the SAS System.

INTRODUCTION

The first in this series, "Guido's Guide to PROC FREQ – A Tutorial for Beginners Using the SAS® System", dealt with answering the Question of "How Many?". This second guide concentrates on answering the question "How much?".

The Version 9 SAS® Procedure Manual states, "The MEANS procedure provides data summarization tools to computer descriptive statistics across all observations and within groups of observations. For example, PROC MEANS calculates descriptive statistics based on moments, estimates quantiles, which includes the median, calculates confidence limits for the mean, identifies extreme values and performs a t-test".

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

```
PROC MEANS <option(s)> ;
      BY variable(s);
      CLASS variable(s) </ option(s)>;
      VAR variable(s) </ WEIGHT=weight-variable>;
      WEIGHT variable;
      FREQ variable;
      ID variable(s);
      TYPES request(s);
      WAYS list;
      OUTPUT <OUT=SAS-data-set> ... </ options>;
RUN;
```

I have underlined the 4 statements in PROC MEANS which I will be discussing in this paper. The PROC MEANS statement is the only required statement for the MEANS procedure. If you specify the following statements, PROC MEANS produces five basic stats (N, Min, Max, Mean, SD) for each numeric variable in the last created dataset.

```
PROC MEANS;
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 the average age (mean) of the 100 patients as well as the average age (mean) of the males and females

Notice that the these questions ask how much and so we know that PROC MEANS is the procedure of choice. We begin by asking SAS for a simple table on the variable AGE using the VAR statement. Then we add a CLASS statement to find out the answer for the men and the women.

Example 1

```
PROC MEANS DATA=Trial;
     VAR Age;
RUN;
```

The SAS System

The MEANS Procedure

Analysis Variable : AGE				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
100	42.5800000	12.0169745	19.0000000	70.0000000

The output above gives us 5 simple statistics. The number of subjects is represented by **N** (N=100). The Minimum Age of the Subjects is represented by **Minimum** (Min=19) and the Maximum Age of the Subjects is represented by **Maximum** (Max=70). The Mean Age of the Subjects is represented by **Mean** (Mean=42.58) and the Standard Deviation of the Mean (**Std Dev** = 12.0169745). So the answer to our first question about what is the average age of the 100 subjects is 42.58 years.

Now we want to know what is the mean age of the men and the mean age of the women and so we can add a CLASS statement to our program to answer this question.

Example 2

```
PROC FREQ DATA=Trial;
     VAR Age;
     CLASS Sex;
RUN;
```

The SAS System

The MEANS Procedure

Analysis Variable : AGE						
SEX	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
F	56	56	42.0892857	12.1464949	19.0000000	69.0000000
M	44	44	43.2045455	11.9603600	19.0000000	70.0000000

The great thing about the SAS System is there is almost always two or more ways to do the same thing and so another way to calculate the mean age of men and the mean age of women is to us a BY statement instead of a CLASS statement. The only caveat is that whenever you use a BY statement, the SAS dataset must be sorted. Let's take a look at the syntax and output.

Example 3

```
PROC SORT DATA=Trial OUT=TrialSorted;
     BY Sex;
RUN;
PROC MEANS DATA=TrialSorted;
     BY Sex;
     VAR Age;
RUN;
```

The SAS System

The MEANS Procedure

SEX=F

Analysis Variable : AGE				
N	Mean	Std Dev	Minimum	Maximum
56	42.0892857	12.1464949	19.0000000	69.0000000

SEX=M

Analysis Variable : AGE				
N	Mean	Std Dev	Minimum	Maximum
44	43.2045455	11.9603600	19.0000000	70.0000000

Now you may be asking yourself, why not just use the CLASS statement and then you won't have to sort the data. While that is correct, there may be times when you want to use both a CLASS statement and a BY statement depending on the problem. In the next example we will use both. In this example we will use Center as our CLASS variable and use Sex as our BY variable. Then we will repeat the analysis using only the CLASS statement.

Example 4

```
PROC MEANS DATA=TrialSorted;
     BY Sex;
     CLASS Center;
     VAR Age;
RUN;
```

The SAS System

The MEANS Procedure

SEX=F

Analysis Variable : AGE						
CENTER	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	24	24	41.3750000	12.3914855	19.0000000	69.0000000
2	20	20	39.6500000	10.7472151	24.0000000	63.0000000
3	12	12	47.5833333	13.1249820	30.0000000	64.0000000

SEX=M

Analysis Variable : AGE						
CENTER	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	16	16	41.5000000	13.4956783	19.0000000	70.0000000
2	15	15	41.0666667	10.0247313	24.0000000	58.0000000
3	13	13	47.7692308	11.6415481	27.0000000	65.0000000

<u>Example 5</u>

```
PROC MEANS DATA=TrialSorted;
     CLASS Center Sex;
     VAR Age;
RUN;
```

The SAS System

The MEANS Procedure

Analysis Variable : AGE							
CENTER	**SEX**	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
1	F	24	24	41.3750000	12.3914855	19.0000000	69.0000000
	M	16	16	41.5000000	13.4956783	19.0000000	70.0000000
2	F	20	20	39.6500000	10.7472151	24.0000000	63.0000000
	M	15	15	41.0666667	10.0247313	24.0000000	58.0000000
3	F	12	12	47.5833333	13.1249820	30.0000000	64.0000000
	M	13	13	47.7692308	11.6415481	27.0000000	65.0000000

While we have concisely produced the above table without sorting the data and using the CLASS statement we could still do more to make it aesthetically pleasing to the eye. So let's decrease the decimal places to two and format the Center and Sex variables.

Example 6

```
PROC FORMAT;
      VALUE Centerf 1='1:Austin'
                    2='2:Dallas'
                    3='3:Conroe';
      VALUE $Sexf 'F'='F:Female'
                  'M'='M:Male';
RUN;

PROC MEANS DATA=TrialSorted MAXDEC=2;
TITLE 'Guido''s Guide to PROC MEANS';
TITLE2 'Example 6 – CLASS, FORMAT and MAXDEC';
     CLASS Center Sex;
     VAR Age;
     FORMAT Center Centerf. Sex Sexf.;
RUN;
```

Guido's Guide to PROC MEANS
Example 6 – CLASS, FORMAT and MAXDEC

The MEANS Procedure

Analysis Variable : AGE

CENTER	SEX	N Obs	N	Mean	Std Dev	Minimum	Maximum
1:Austin	Female	24	24	41.38	12.39	19.00	69.00
	Male	16	16	41.50	13.50	19.00	70.00
2:Dallas	Female	20	20	39.65	10.75	24.00	63.00
	Male	15	15	41.07	10.02	24.00	58.00
3:Conroe	Female	12	12	47.58	13.12	30.00	64.00
	Male	13	13	47.77	11.64	27.00	65.00

Up to this point we have been letting PROC MEANS produce the "default" statistics of N, MIN, MAX, MEAN and STD DEV. (See Appendix A for available statistics from PROC MEANS)

Suppose that we want to see the MEAN, MEDIAN and the 95% Confidence Limits of the Mean. Whenever we want anything other than the default statistics we have to explicitly ask for them.

Example 7

```
PROC MEANS DATA=TrialSorted LCLM MEAN UCLM MEDIAN MAXDEC=2;
TITLE 'Guido''s Guide to PROC MEANS';
TITLE2 'Example 7 – Selected Statistics for Age';
     CLASS Center Sex;
     VAR Age;
     FORMAT Center Centerf. Sex Sexf.;
RUN;
```

Guido's Guide to PROC MEANS
Example 7 – Selected Statistics for Age

The MEANS Procedure

Analysis Variable : AGE							
CENTER	**SEX**	**N Obs**	**Lower 95% CL for Mean**	**Mean**	**Upper 95% CL for Mean**	**Std Dev**	**Median**
1:Austin	Female	24	36.14	41.38	46.61	12.39	41.50
	Male	16	34.31	41.50	48.69	13.50	41.00
2:Dallas	Female	20	34.62	39.65	44.68	10.75	39.50
	Male	15	35.52	41.07	46.62	10.02	42.00
3:Conroe	Female	12	39.24	47.58	55.92	13.12	48.00
	Male	13	40.73	47.77	54.80	11.64	46.00

We now have a report that transmits the data very succinctly and clearly. Let's try to do some basic statistical analyses using PROC MEANS.

If we look at the output in Example 7, then we can see that for each center there appears to be no statistically significant difference between the mean ages of the men and women. For example, in the Austin center the mean age for women is 41.38 with LCLM equal to 36.14 and UCLM equal to 46.61. The mean age for men is 41.50 with LCLM equal to 34.31 and UCLM equal to 48.69. Generally speaking, if the mean for one group is contained with the LCLM and UCLM for the other group, there is no statically significant difference in the two groups. Repeating this observation for the Dallas center, we find that there is no statistically significant difference in the mean ages of women versus men. Finally, there is also no statistically significant difference in the mean ages of the women versus the men in the Conroe center.

Now let's try a slightly different statistical analysis and let SAS do the testing. We can consider an example from Glenn Walker's book in Chapter 4. Here is a synopsis of the problem:

Mylitech is developing a new appetite suppressing compound for use in weight reduction. A preliminary study of 35 obese patients provided data before and after 10 weeks of treatment with the new compound. Does the new treatment look at all promising? Let's take a look at the VIEWTABLE version of the SAS Dataset – Work.Obese

VIEWTABLE: Work.Obese

	subj	wtpre	wtpost	wtloss
1	1	165	160	-5
2	2	202	200	-2
3	3	256	259	3
4	4	155	156	1
5	5	135	134	-1
6	6	175	162	-13
7	7	180	187	7
8	8	174	172	-2
9	9	136	138	2
10	10	168	162	-6
11	11	207	197	-10
12	12	155	155	0
13	13	220	205	-15
14	14	163	153	-10
15	15	159	150	-9
16	16	253	255	2
17	17	138	128	-10
18	18	287	280	-7
19	19	177	171	-6
20	20	181	170	-11
21	21	148	154	6
22	22	167	170	3
23	23	190	180	-10
24	24	165	154	-11
25	25	155	150	-5
26	26	153	145	-8
27	27	205	206	1
28	28	186	184	-2
29	29	178	166	-12
30	30	129	132	3
31	31	125	127	2
32	32	165	169	4
33	33	156	158	2
34	34	170	161	-9
35	35	145	152	7

Notice that some subjects have a negative wtloss (this means they lost weight after the 10 weeks of treatment with the new compound). Some subjects have a positive wtloss (this means they gained weight after the 10 weeks of treatment with the new compound). If the average wtloss is not different from 0, then we conclude that there is no statistically significant difference between the beginning weight (wtpre) and the ending weight (wtloss) which is represented by the variable wtloss. PROC MEANS will test this hypothesis (referred to as "The Null Hypothesis").

Example 8

```
PROC MEANS DATA=Obese N MEAN STD T PRT MAXDEC=2;
TITLE 'Guido''s Guide to PROC MEANS';
TITLE2 'Example 8 – Paired t-Test for Weight Loss';
     VAR wtloss;
RUN;
```

Guido's Guide to PROC MEANS
Example 8 – Paired t-Test for Weight Loss

The MEANS Procedure

Analysis Variable : wtloss				
N	**Mean**	**Std Dev**	**t Value**	**Pr > |t|**
35	-3.46	6.34	-3.23	0.0028

If we examine the output from Example 8 then for the 35 subjects we find that the mean difference in weight loss is -3.46 pounds, the standard deviation is 6.34, the t-value is -3.23 and the p-value is 0.0028. If the p-value is less than 0.05 then we may reject 'The Null Hypothesis". The p-value is 0.0028 and so we can reject "The Null Hypothesis" and conclude that there is a statistically significant difference in weight loss of the 35 subjects between pre and post treatment weights.

There are other procedures in the SAS System that can answer this question. You could use PROC UNIVARIATE which give a plethora of output, PROC SUMMARY which gives no output (by default) and since the emergence of version 7 of the SAS System you can use PROC TTEST to do the paired t-Test analysis.

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).

APPENDIX A – STATISTIC KEYWORDS FOR PROC MEANS STATEMENT

DESCRIPTIVE STATISTIC KEYWORDS

CLM – Two sided Confidence Limit of the Mean

CSS – Corrected Sum of Squares

CV – Coefficient of Variation

KURTOSIS|KURT – Kurtosis

LCLM – Lower Confidence Limit of Mean

MAX - Maximum

MEAN – Average

MIN – Minimum

N - Number of non-missing values

NMISS – Number of missing values

RANGE – Maximum minus Minimum

SKEWNESS|SKEW - Skewness

STDDEV|STD – Standard Deviation

STDERR – Standard Error of the Mean

SUM – Sum of the

SUMWGT – Sum of the Weights

UCLM – Upper Confidence Limit of Mean

USS – Uncorrected Sum of Squares

VAR - Variance

QUANTILE STATISTIC KEYWORDS

MEDIAN|P50 – Median or 50th Percentile

P1 - 1st Percentile

P5 – 5th Percentile

P10 – 10th Percentile

Q1|P25 – 1st Quartile or 25th Percentile

Q3|P75 – 3rd Quartile or 75th Percentile

P90 - 90th Percentile

P95 – 95th Percentile

P99 – 99th Percentile

QRANGE – Interquartile Range (Q3 – Q1)

HYPOTHESIS STATISTIC KEYWORDS

PROBT – two-tailed p-value for Student's t statistic

T – Student's t statistic

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

Guido, Joseph J. (2007). "Guido's Guide to PROC FREQ – A Tutorial for Beginners Using the SAS® System", Proceedings of the 20th annual North East SAS Users Group Conference, Baltimore, MD, 2007, paper #FF07.

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

ACKNOWLEDGEMENTS
SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

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