SAS Tutorial: PROC MEANS

Data Set: dog_mandible_measurements

Tutorial Instructions:

For this tutorial we demonstrate how to use SAS procedure PROC MEANS to examine distributions and summary statistics of data that are useful in exploratory data analysis (EDA). Many of the descriptive statistics available within PROC UNIVARIATE can be produced using PROC MEANS. PROC MEANS is useful when you want specific descriptive statistics. PROC MEANS does not produce any graphical output.

(0) To begin we Set the Dog Mandible Measurements data set with a shorter name for reference

```
* Set the dog mandible measurements data set to short name;
data dogjaw;
set mydata.dog_mandible_measurements;
run;
```

A description is available in our data dictionary. For the convenience of this tutorial we will post that description here.

```
Multivariate Statistical Methods: A Primer 3rd Edition ISBN 1584884142
pp. 55-57
This SAS Tutorial shows how to produce the following:
 1. PROC MEANS

 Data set contains 77 observations of dog mandible measurements.

 Variables:
 GROUP_NAME: type of dog
 CASE: observation number for the group
 GROUP_NBR: coded value for the GROUP_NAME
 X1: length of mandible
 X2: breadth of mandible below first molar
 X3: breadth of articular condyle
 X4: height of mandible below first molar
 X5: length of first molar
 X6: breadth of first molar
 X7: length of first to third molar
 X8: length from first fourth premolar
 X9: breadth of lower canine
 SEX: 1=male, 2=female, 0=unknown

* print first 6 observations checking values of selected variables;
Title "Observe Data Set";
options obs=6;
proc print data=dogjaw; run;
options obs=max; * reset options to analyze and report on all data;
```

Table 1 is the output from the PROC PRINT script that shows a sample of six observations from the data set. There may be times when you first load the data set, add, or transform variables that you will want to look at the values in the data set. If you do look at a select number of observations, make certain you reset to `options obs = max;` to reset to full data set.

Obs	GROUP_NAME	CASE	GROUP_NBR	X1	X2	X3	X4	X5	X6	X7	X8	X9	SEX
1	Modern	1	1	123.0	10.1	23.0	23.0	19.0	7.8	32.0	33.0	5.6	1
2	Modern	2	1	137.0	9.6	19.0	22.0	19.0	7.8	32.0	40.0	5.8	1
3	Modern	3	1	121.0	10.2	18.0	21.0	21.0	7.9	35.0	38.0	6.2	1
4	Modern	4	1	130.0	10.7	24.0	22.0	20.0	7.9	32.0	37.0	5.9	1
5	Modern	5	1	149.0	12.0	25.0	25.0	21.0	8.4	35.0	43.0	6.6	1
6	Modern	6	1	125.0	9.5	23.0	20.0	20.0	7.8	33.0	37.0	6.3	1

Table 1: Sample of Six Observations from Data Set

(1) We will use PROC MEANS to produce several descriptive statistics that describe the distribution of each identified variable. If no variables are identified, the procedure will provide results on all variables in the data set. I selected only the continuous variables (X1-X9) from the data set.

```
* produce default proc means descriptive statistics;
Title "PROC MEANS EDA - Examine Variable Descriptive Statistics";
proc means data=dogjaw;
var X1 X2 X3 X4 X5 X6 X7 X8 X9;
run;
```

Table 2 shows the default descriptive statistics if no statistic keywords are provided. For each variable, the default statistics are the number of non-missing observations, the variable mean, standard deviation, and the minimum and maximum observation value.

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	77	128.9740260	17.5018601	105.0000000	177.0000000
X2	77	9.9610390	1.4030189	7.2000000	13.4000000
X3	77	21.9480519	3.5832068	17.0000000	32.0000000
X4	77	21.4935065	3.3780381	15.0000000	28.0000000
X5	77	20.4935065	2.4901034	17.0000000	27.0000000
X6	77	8.0000000	1.0236673	6.0000000	10.5000000
X7	77	32.5194805	4.1726250	26.0000000	43.0000000
X8	77	37.4025974	4.4047216	31.0000000	50.0000000
X9	77	6.0753247	1.0196950	4.3000000	8.5000000

Table 2: PROC MEANS – Simple Descriptive Statistics

You can request specific descriptive statistics using keywords. If you request any keyword, this replaces the default results and you must request all desired measures.

```
* produce specific descriptive proc means statistics using selected
options;
Title "PROC MEANS EDA - Examine Variable for Specific Descriptive
Statistics";
proc means data=dogjaw n q1 mean q3 stddev median clm ndec=3;
var X1 X2 X3 X4 X5 X6 X7 X8 X9;
run;
```

The script requests the number of non-missing observations, the 25% quantile value, the mean, the 75% quantile value, standard deviation, median, and upper and lower confidence limits at the default 0.05% and 95% levels with the results shown in Table 3. Additional options can be found in *The Little SAS Book* on page 258.

Variable	N	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 95% CL for Mean	Upper 95% CL for Mean
X1	77	114.000	128.974	137.000	17.502	125.000	125.002	132.946
X2	77	8.700	9.961	10.900	1.403	10.000	9.643	10.279
X3	77	19.000	21.948	25.000	3.583	22.000	21.135	22.761
X4	77	18.000	21.494	24.000	3.378	22.000	20.727	22.260
X5	77	19.000	20.494	22.000	2.490	20.000	19.928	21.059
X6	77	7.100	8.000	8.700	1.024	7.900	7.768	8.232
X7	77	30.000	32.519	33.000	4.173	31.000	31.572	33.467
X8	77	34.000	37.403	39.000	4.405	36.000	36.403	38.402
X9	77	5.300	6.075	6.800	1.020	6.100	5.844	6.307

Table 3: PROC MEANS – Requested Statistical Measures

Inclusion of the ndec=3 sets the number of decimals to 3 places to allow for better table layout, especially when requesting several descriptive statistics.

(2) As with PROC UNIVARIATE, PROC MEANS can be used to examine statistics using the BY and CLASS statements. It is recommended to sort the data by the grouping variable when using the BY statement.

```
* recommended practice is to sort the data by the group variable;
proc sort data=dogjaw;
by SEX;
run;
```

The data set has two categorical variables that could serve to segment the data set. For this tutorial, I chose to examine the data by the sex of the dog. Once the segmentation variable is sorted, include the BY statement in the procedure.

```
* produce descriptive statistics using BY statement;
Title "PROC MEANS w/ BY Statement - Examine Variable for Specific
Descriptive Statistics";
proc means data=dogjaw n q1 mean q3 stddev median clm ndec=3;
var X1 X2; *X3 X4 X5 X6 X7 X8 X9;
by SEX;
run;
```

Table 4 shows the output for variables X1 and X2 by the dog's SEX.

SEX=0

Variable	N	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 95% CL for Mean	Upper 95% CL for Mean
X1	10	115.000	122.800	130.000	8.417	123.000	116.779	128.821
X2	10	9.800	10.340	10.700	0.771	10.050	9.789	10.891

SEX=1

Variable	N	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 95% CL for Mean	Upper 95% CL for Mean
X1	35	117.000	133.686	145.000	19.129	131.000	127.115	140.257
X2	35	8.700	10.274	11.500	1.523	10.400	9.751	10.797

SEX=2

Variable	N	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 95% CL for Mean	Upper 95% CL for Mean
X1	32	111.000	125.750	134.000	16.762	123.500	119.707	131.793
X2	32	8.400	9.500	10.700	1.316	9.600	9.025	9.975

Table 4: PROC MEANS Descriptive Statistics BY Categorical Variable

(3) Another useful view using PROC MEANS is to examine the distribution using CLASS statement with the quantile keywords. The CLASS statement is similar to BY statement and segments the output by the selected variable, typically a categorical variable.

```
* examine means at 5th, 10th, 25th, 50th, 75th, 90th, and 95th
percentile;
proc means data=dogjaw p5 p10 p25 p50 p75 p90 p95 ndec=2;
class SEX;
var X1 X2; *X3 X4 X5 X6 X7 X8 X9;
run;
```

Within this script example, I set the number of decimals to two decimals places as another example of formatting the SAS output. The results shown in Table 5 are helpful in examining for breaks (or cuts) in data that may be useful in creating discretized variables from the continuous variables.

SEX	N Obs	Variable	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl
0	10	X1	111.00	111.50	115.00	123.00	130.00	134.00	136.00
		X2	9.50	9.55	9.80	10.05	10.70	11.55	11.90
1	35	X1	110.00	112.00	117.00	131.00	145.00	165.00	167.00
		X2	8.10	8.20	8.70	10.40	11.50	12.30	12.60
2	32	X1	106.00	107.00	111.00	123.50	134.00	148.00	163.00
		X2	7.30	7.70	8.40	9.60	10.70	10.90	11.30

Table 5: PROC MEANS Results Showing Quantile Results

As shown in other tutorials, a macro is a convenient method to repeat steps. The scripts could be combined in to a single macro using the following script.

```
* Note that I have created a "macro function" named %myMEANS()
which has a "macro variable" x as an argument.;

%macro myMEANS(x, y);    *macro name & start;
TITLE "PROC MEANS - Examine Variable Descriptive Statistics for &x by
&y";
proc means data=dogjaw n nmiss q1 mean q3 stddev median clm alpha=0.10
ndec=3;
var &x;
by &y;
run;

proc means data=dogjaw p5 p10 p25 p50 p75 p90 p95 ndec=2;
class &y;
var &x;
run;
%mend myMEANS;        *macro end;

* calls to macro for each variable;
%myMEANS(x=X1, y=SEX);
%myMEANS(x=X2, y=SEX);
%myMEANS(x=X3, y=SEX);
```

In the macro, you define your input variables using the '&' symbol in front of the declared variable. In the example, '&x' defines the predictor variable and '&y' defines the BY and CLASS variable, which are provided in the statement that calls the macro. Table 6 shows the results for the macro statement using X3. It includes the two separate PROC MEANS descriptive statistics requests. Notice that the output title includes the "X3 by SEX" label. These values were passed into the script through the macro TITLE statement. This is a useful technique to organize your output, especially when reviewing large output files.

PROC MEANS - Examine Variable Descriptive Statistics for X3 by SEX

The MEANS Procedure

SEX=0

Analysis Variable : X3								
N	N Miss	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 90% CL for Mean	Upper 90% CL for Mean
10	0	19.000	20.000	22.000	1.944	19.500	18.873	21.127

SEX=1

Analysis Variable : X3								
N	N Miss	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 90% CL for Mean	Upper 90% CL for Mean
35	0	19.000	22.914	25.000	4.147	23.000	21.729	24.100

SEX=2

Analysis Variable : X3								
N	N Miss	Lower Quartile	Mean	Upper Quartile	Std Dev	Median	Lower 90% CL for Mean	Upper 90% CL for Mean
32	0	19.000	21.500	24.000	3.005	21.500	20.599	22.401

Analysis Variable : X3								
SEX	N Obs	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl
0	10	17.00	17.50	19.00	19.50	22.00	22.50	23.00
1	35	17.00	18.00	19.00	23.00	25.00	29.00	31.00
2	32	17.00	18.00	19.00	21.50	24.00	26.00	27.00

Table 6: PROC MEANS Output for X3 by SEX using Macro script