SAS Tutorial: PROC UNIVARIATE

Data Set: dog_mandible_measurements

Tutorial Instructions:

For this tutorial we demonstrate how to use SAS procedure PROC UNIVARIATE to examine distributions of data and produce several graphs that are useful in exploratory data analysis (EDA).

(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 and a sample of the observations here.

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

 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
```

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

(1) We will use PROC UNIVARIATE to produce several statistics and graphs that describe the distribution of a single variable. The statistics include the mean, median, mode, standard deviation, skewness, and kurtosis.

```
* examine descriptive statistics and distributions of data set
variables;
Title "PROC UNIVARIATE EDA - Examine Variable Descriptive Statistics";
proc univariate data=dogjaw;
var X1 X2 X3 X4 X5 X6 X7 X8 X9 SEX;
run;
```

The output for the PROC CORR procedure is given in Table 1, Table 2, Table 3, Table 4, and Table 5. The results are for the variable X1. Table 1 shows the descriptive statistics of the data set. It includes the number of observations, the distribution mean, variance, and standard deviation among others.

Moments			
N	77	**Sum Weights**	77
Mean	128.974026	**Sum Observations**	9931
Std Deviation	17.5018601	**Variance**	306.315106
Skewness	0.85736651	**Kurtosis**	0.09631491
Uncorrected SS	1304121	**Corrected SS**	23279.9481
Coeff Variation	13.5700657	**Std Error Mean**	1.99452206

Table 1: Proc Univariate – Simple Summary Statistics

Table 2 shows basic statistical measures that describe the distribution along with the main three measures of central tendency.

Basic Statistical Measures			
Location		**Variability**	
Mean	128.9740	**Std Deviation**	17.50186
Median	125.0000	**Variance**	306.31511
Mode	111.0000	**Range**	72.00000
		Interquartile Range	23.00000

Table 2: Proc Univariate – Basic Statistical Measures

Table 3 is a test that the mean of the population is zero. For variable X1 from the dog mandible measurements, we can reject the null hypothesis and conclude that the population mean is not equal to zero with statistical significance less than 0.0001.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	64.66413	**Pr > |t|**	<.0001
Sign	**M**	38.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	1501.5	**Pr >= |S|**	<.0001

Table 3: Test Hypothesis that the Population Mean is 0

Table 4 shows the distribution quantiles and Table 5 identifies the five lowest and highest values.

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	177
99%	177
95%	165
90%	163
75% Q3	137
50% Median	125
25% Q1	114
10%	110
5%	107
1%	105
0% Min	105

Table 4: Distribution Quantiles

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
105	32	164	64
106	34	165	61
107	33	166	59
107	29	167	54
107	18	177	58

Table 5: Distribution Extreme Observations

Within data set, GROUP_NAME and coded version GROUP_NBR segment the data. During EDA, it may be worthwhile to examine distributions within the variable by these segments. Below is a sample script to examine the variable X1 by GROUP_NAME class.

```
* examine descriptive statistics and distributions of data set
variables using CLASS statement;
Title "PROC UNIVARIATE w/ Class - Examine Variable Descriptive
Statistics";
proc univariate data=dogjaw;
class GROUP_NAME;
var X1; *X2 X3 X4 X5 X6 X7 X8 X9 SEX;
run;
```

Note that the other variables are commented out. In example, I elected to examine only one variable. I can include statement of all variables. This may be become processing intensive. An alternative is I can write a macro to repeat the actions for each variable. Below is the script for a macro `%myUNIVARIATEbyCLASS()` that will examine the distribution statistics for each variable by the class GROUP_NAME.

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

%macro myUNIVARIATEbyCLASS(x);    *macro name & start;
TITLE "PROC UNIVARIATE w/ Class - Examine Variable Descriptive
Statistics Colony by &x";
proc univariate data=dogjaw;
class GROUP_NAME;
var &x;
run;
%mend myUNIVARIATEbyCLASS;        *macro end;

* calls to macro for each variable;
%myUNIVARIATEbyCLASS(x=X1);
%myUNIVARIATEbyCLASS(x=X2);
%myUNIVARIATEbyCLASS(x=X3);
```

In the example script, I conclude with macro for variable X3. I can add additional statements to review the other variables.

(2) Within PROC UNIVARIATE we can produce several graphs that are useful for EDA. The format is similar to the statements used to examine the distributions in table format. The plot options are cdfplot, for the cumulative distribution plot; histogram, to view distribution of variable; pplot, probplot, and qqplot, these probabilities plots compare data to theoretical distributions.

```
* create statistical graphics with PROC UNIVARIATE;
Title "PROC UNIVARIATE Statistical Plots";
proc univariate data=dogjaw;
var X1; *X2 X3 X4 X5 X6 X7 X8 X9 SEX;
cdfplot X1;              * cumulative distribution plot;
histogram X1 / normal;   * histogram plot;
ppplot X1;               * probability-probability plot;
probplot X1;             * probability plot;
qqplot X1;               * quantile-quantile plot;
run;
```

The table results are the similar to those shown in the basic statement. The request for the histogram with normal option includes tests for the normal distributions goodness-of-fit. The requested graphs are shown in Figure 1.

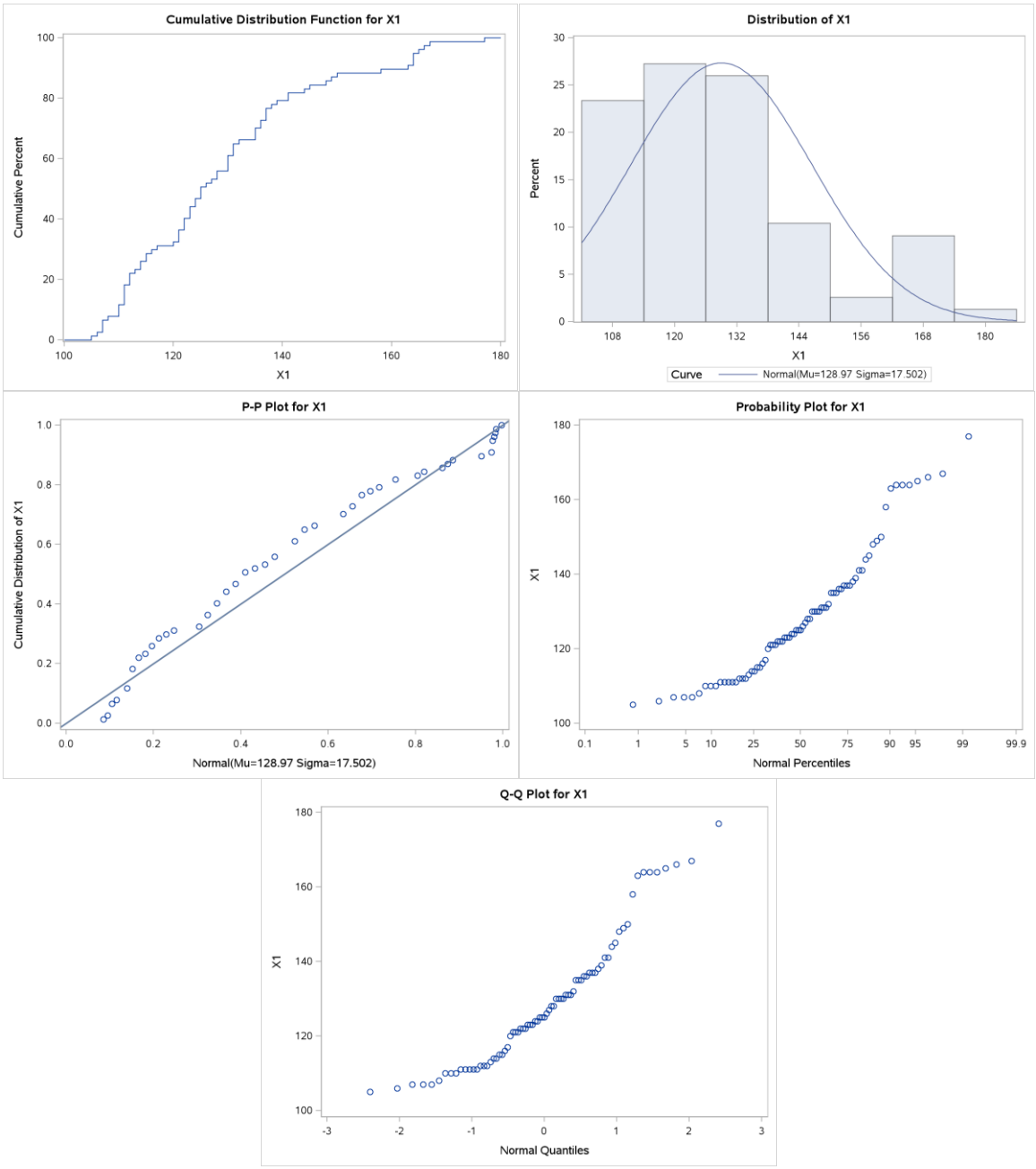


Figure 1: PROC UNIVARIATE Plots a) CDF, b) Histogram, c) P-P Plot, d) Probability Plot, e) Q-Q Plot

As was shown with the macro to examine the distributions by class, another macro can be written to examine the graphical details for each variable. The macro is done for X1, X2, and X3. It could be expanded to include the other variables.

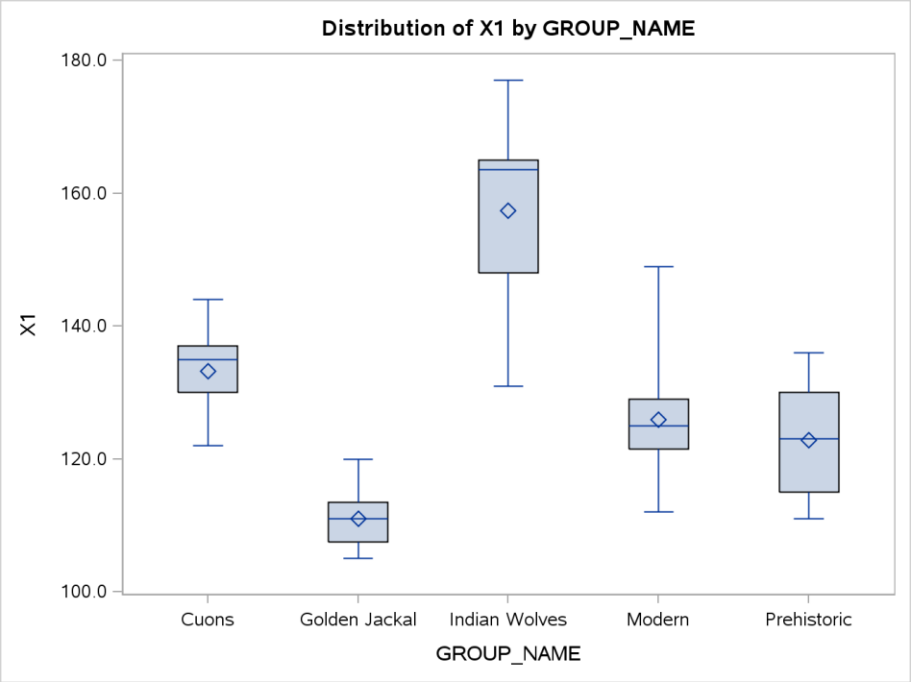
```
* Note that I have created a "macro function" named
%myUNIVARIATEplots()
which has a "macro variable" x as an argument.;

%macro myUNIVARIATEplots(x);   *macro name & start;
TITLE "PROC UNIVARIATE Statistical Plots by &x";
proc univariate data=dogjaw;
var &x;
cdfplot &x;              * cumulative distribution plot;
histogram &x / normal;   * histogram plot;
ppplot &x;               * probability-probability plot;
probplot &x;             * probability plot;
qqplot &x;               * quantile-quantile plot;
run;
%mend myUNIVARIATEplots;       *macro end;

* calls to macro for each variable;
%myUNIVARIATEplots(x=X1);
%myUNIVARIATEplots(x=X2);
%myUNIVARIATEplots(x=X3);
```

(3) Boxplots are a good visual technique to observe a possible difference between two groups. This is not a plot option within PROC UNIVARIATE; however, it may be useful in the EDA done during use of PROC UNIVARIATE.

```
* boxplot is a good visual approach to observe differences in
distributions;
* recommended practice is to sort the data by the group variable;
proc sort data=dogjaw;
by GROUP_NAME;
run;
Title "Variable EDA - Boxplot of Variables";
proc boxplot data=dogjaw;
plot X1 * GROUP_NAME;
run;
```



As described earlier, you may elect to write a macro to repeat the plot for each variable. We can observe a possible difference in the mean values and might consider a t-test to compare the means of two groups. The Logistic Regression tutorial discusses options for performing a t-test to compare group means.