NOTE #7: DESCRIPTIVE AND UNIVARIATE STATISTICS II

PROC UNIVARIATE; The procedure provides more extensive list of statistics and is one of the most useful procedures.

```
PROC UNIVARIATE <DATA=mydata> < / options>;
  VAR variable1 variable2, …;
  < statements >
  OUTPUT OUT=outdataname
         Statistics = names;
RUN;
```

The default output of this procedure is very comprehensive. We will discuss a sample output in class. Following is list of statements that can be used with the procedure. We will discuss the statements via following examples.

List of statements:

```
BY variables ;
CLASS variables ;
FREQ variables ;
HISTOGRAM < variables > < / options > ;
ID variables ;
PROBPLOT < variables > < / options > ;
QQPLOT < variables > < / options > ;
VAR variables ;
WEIGHT variable ;
```

```
/* Example 7-1 */
OPTIONS LINESIZE=75 PAGESIZE=54 NODATE PAGENO=1;
DM "output;clear;log;clear";

GOPTIONS reset=all goutmode=replace
                          Vsize=6 Hsize=6 Horigin=1.2
                          htitle=1.0 ftitle=simplex
                          htext=1.0 ftext=simplex;

DATA GNP; SET SASHELP.GNP;
Year =year(date);
quarter = qtr(date);
RUN;

PROC SORT; by quarter; RUN;
PROC UNIVARIATE DATA = GNP NORMAL;
var GNP;
by quarter;
QQPLOT GNP;
HISTOGRAM GNP /Midpoints = 600 to 5400 by 600  NORMAL CFILL = ltgray;
INSET MEAN = 'Mean' (8.1)  STD = 'StdDev' (8.1) / POSITION = NE;
RUN;
```

All statistics provided in PROC MEANS also available with PROC UNIVARIATE. Beside those statistics, PROC UNIVARIATE computes much more comprehensive list of statistics (see sample output). In OUTPUT statement we can make use of the statistics used with PROC MEANS. Using PROC UNIVARIATE we can output more percentiles than those automatically calculated with the procedure. See the example below.

```
/*Example 7-2 */
DATA Rand_Norm;
  Count=100; N=30; MU=5; STD=2; seed=0;
  DO I=1 TO Count;
   DO K=1 TO N;
    X=MU+STD*RANNOR(seed); OUTPUT;
   END;
  END;
RUN;

Proc Univariate data = Rand_Norm;
 var X;
 by I;
 output out=norm_out  pctlpre=P_  pctlpts=2.5 to 10 by 2.5 95 to 100 by 2.5;
run;
proc print; run;
```

Making output data set from portion of SAS output

We have noticed that the SAS output from PROC UNIVARIATE is very comprehensive and you may not need all of them. Using ODS (Output Delivery System) you can choose parts of SAS output to be printed or to be saved as an output data that can be used for further analysis. To have parts to be selected we first need to know the parts names. For this, you can first run the following, for example,

```
        /*Example 7-3 */
        ODS TRACE ON;
        PROC UNIVARIATE DATA = SASHELP.GNP NORMAL;
        var GNP INVEST;
        RUN;
        ODS TRACE OFF;
```

In the LOG window you will see the following list. Note that the following is a part of the list.

```
Output Added:
-------------
Name:       Moments
Label:      Moments
Template:   base.univariate.Moments
Path:       Univariate.GNP.Moments
-------------

Output Added:
-------------
Name:       BasicMeasures
Label:      Basic Measures of Location and Variability
Template:   base.univariate.Measures
Path:       Univariate.GNP.BasicMeasures
-------------
```

…

Now, you can print the moment part of the SAS output using ODS LISTING statement in the procedure.

```
/*Example 7-4 */
ODS LISTING EXCLUDE ALL; *this will turn off all output;

PROC UNIVARIATE DATA = SASHELP.GNP NORMAL;
 var GNP INVEST;
 ODS LISTING SELECT Moments;
RUN;

ODS LISTING; *this will turn back on all output;
```

Note that you will see two moment statistics tables; one for GNP and another for INVEST. This is because both peaces have the same output name, “Moments”. To have the output for GNP only you can use the path name. For example, you can replace the ODS LISING statements as

```
ODS LISTING SELECT Univariate.GNP.Moments;
```

Next step is to create a SAS data from parts of SAS output. For this we can use ODS OUPUT statement.

```
/*Example 7-5 */
PROC UNIVARIATE DATA = SASHELP.GNP NORMAL;
var GNP INVEST;
ODS OUTPUT Univariate.GNP.Quantiles = GNP_Q;
ODS OUTPUT Univariate.INVEST.Quantiles = INV_Q;
RUN;

DATA QUANT; SET GNP_Q INV_Q;
PROC PRINT DATA=QUANT NOOBS;
 WHERE Quantile in ('5%','95%'); RUN;
```

Though mostly useful with a comprehensive procedure like UNIVARIATE, ODS can be used with any statistical procedures. For example,

```
/*Example 7-6 */
DATA GNP; SET SASHELP.GNP;
Year =year(date);
quarter = qtr(date); RUN;

PROC TTEST DATA = GNP; where quarter in (1,2);
class quarter;
var  GNP ;
ODS OUTPUT ttests=GNP_t; RUN;

DATA _NULL_; SET GNP_t;
IF variances = 'Equal';
FILE PRINT;
PUT @10 "Two sample T-Test " / @10  "Comparing mean GNP for first and
second quarter" /
@10 60*'-' / ;
IF Probt < .05 Then
   PUT @10 "Since P-value = " Probt " less than .05, we reject Ho: equal
mean at .05 level";
ELSE IF Probt >= .05 Then
   PUT @10 "Since P-value = " Probt " greater than .05, " / @10  "we DO
NOT reject Ho: equal mean at .05 level"; RUN;
```

/* INCLASS PRACTICE 1*/

1. Simulate a random sample of size 1000 from Uniform(0,1). Write a SAS code to provide the test for normality. Also provide histogram and QQ plot. Based on the result of the normality test provide either 95% t-confidence interval or 95% percentile interval. You should use the ODS and your output should look something like:

```
Variable : XXXX
Sampling Distribution: Uniform (0,1)
Normality Test (Shapiro-Wilk): Rejected/accepted with P-value xxx

Sample mean: xxx
Sample std: xxx
95% confidence interval: (xxx,xxx)

NOTE: Due to the lack of normality the CI is based on 2.5 and
97.5 percentiles. (this note is not necessary for normal sample)
```

2. Repeat with a random sample of size 1000 from Poisson distribution with mean of your choice.

/* INCLASS PRACTICE 2*/

We study the approximate normality of the sum of continuous uniform random variables.

(1) Simulate a random sample, Y_1, of size 1000 from Unif(0,1)
(2) Simulate another r.s., U, from Unif(0,1) and let Y_2=Y_1+U
(3) Simulate another r.s., U, from Unif(0,1) and let Y_3=Y_2+U
(4) Continue until you have Y_1,Y_2,…,Y_{10}. Your data should consist of these 10 variables.
(5) Using the Univariate procedure, provide histogram for each variable. Also provide test results for normality for each variable. You output should look something like:

Variable	P-value(S-W)	Normality
Y1	xxx	YES/NO
Y2	xxx	YES/NO
…	…	…

(6) Depending on the result from normality test, provide either 95% t-confidence interval or 95% percentile interval for mean for each variable. Add the information to the result in (5), for example,

Variable	P-value(S-W)	Normality	Sample mean	CI
Y1	xxx	YES/NO	xxx	(xxx,xxx)
Y2	xxx	YES/NO	xxx	(xxx,xxx)
…	…	…	…	…

NOTE that due to the lack of normality for variable xx,xx,xx,and xx, the CI’s are based on 2.5 and 97.5 percentile.