Data Analysis and Basic Statistics

Module Content

- PROC MEANS
- PROC SUMMARY
- PROC UNIVARIATE
- PROC CORR

Proc Means

- ✓ What does the MEANS Procedure do ?
  - ✓ Provides summary statistics (descriptive) for variables across observations and within groups

Means Procedure -- Syntax

- ✓ PROC MEANS *<other option(s)> <statistic-keywords>*;

  BY <DESCENDING> variable 1 ……… <DESCENDING> variable n*;*

  VAR variable(s);

  CLASS variable(s);

  OUTPUT <OUT = SAS-dataset> *<output-statistic-specification>*;

Example

Find average credit limit and average risk score in performance data

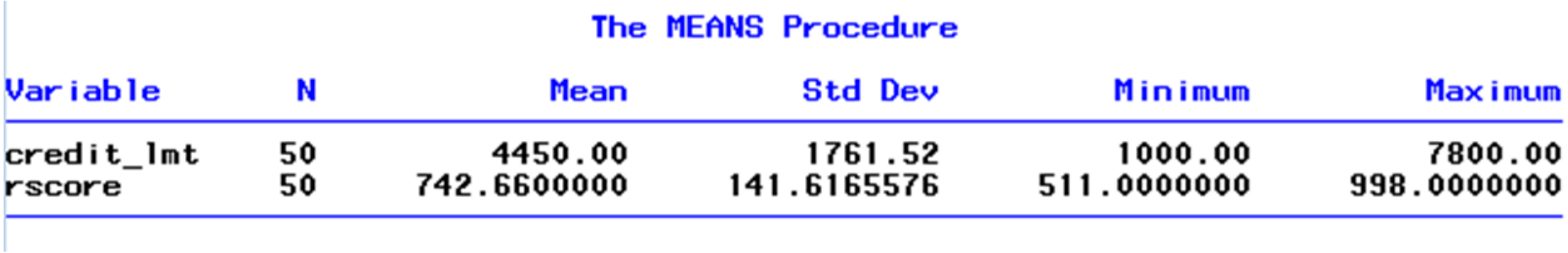
```
proc means data=perf;
var credit_lmt rscore;
run;
```

Find sum of credit limit for each segment

```
proc means data=perf sum;
class segment ;
var credit_lmt rscore;
run;
```

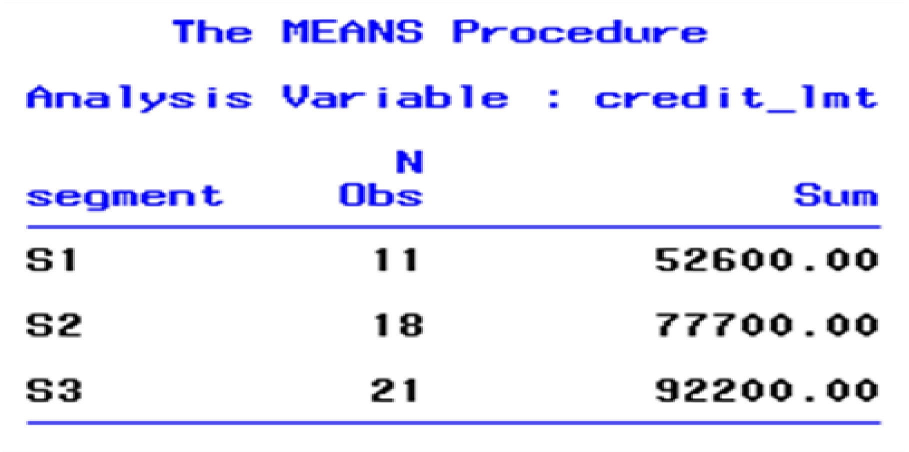
Proc Means – SAS Output

Means on Credit Limit and Risk Score



```
                          The MEANS Procedure

Variable      N           Mean        Std Dev        Minimum        Maximum
---------------------------------------------------------------------------
credit_lmt   50        4450.00        1761.52        1000.00        7800.00
rscore       50    742.6600000    141.6165576    511.0000000    998.0000000
---------------------------------------------------------------------------
```

Total Credit Limit by Each Segment



```
      The MEANS Procedure

Analysis Variable : credit_lmt

                 N
segment        Obs              Sum
-----------------------------------
S1              11         52600.00

S2              18         77700.00

S3              21         92200.00
-----------------------------------
```

Proc Summary

- Similar to Proc Means
- No output is given unless ‘print’ or ‘output’ option is specified
- VAR statement is necessary. In Proc Means if you omit var statement, it gives statistics for all numeric variable in the data

Example

Analyze average credit limit, spend and utilization by combination of Risk Levels and customer segment

Proc Summary – SAS Code

```
data new;
set perf;
if rscore > 800 then risk_level='L';
else if rscore > 600 then risk_level='M';
else risk_level='H';
run;

proc summary data=new nway missing;
class segment risk_level;
var credit_lmt rscore spend;
output out=summ sum=;
run;

proc contents data=summ;
run;

proc print data=summ;
run;
```

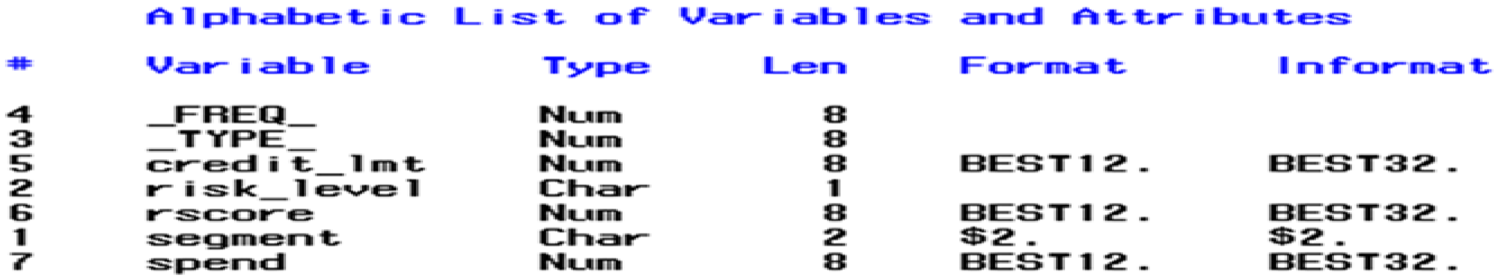
Example

- Omit ‘Nway’ in the proc summary option and notice the change in output
- Missing- Would treat ‘missing’ values in class variables as a separate category
- Sum= specifies that ‘sum’ needs to be output for variables

Proc Summary – Output

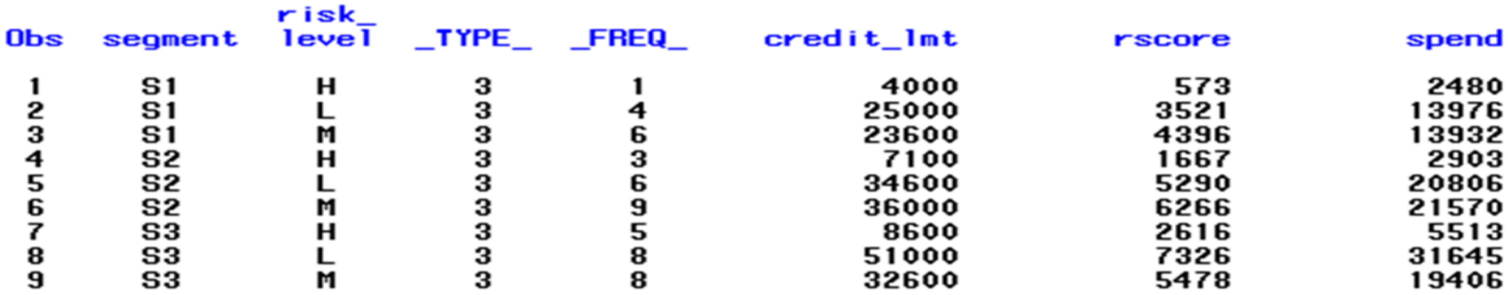
greatlearning
Learning for Life

Contents of Summ Data Set



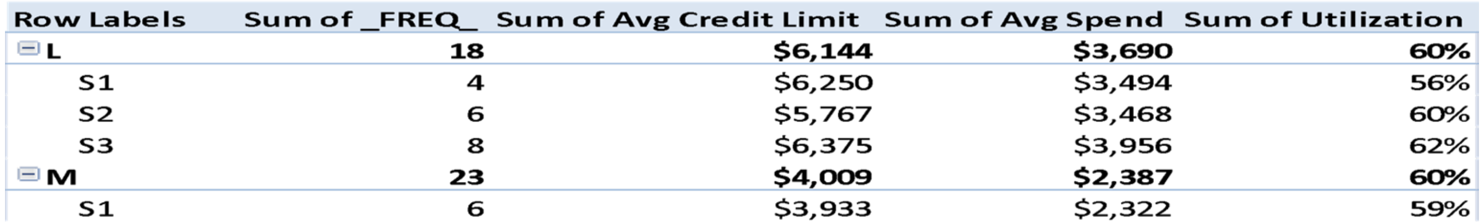
```
          Alphabetic List of Variables and Attributes

#     Variable          Type     Len    Format       Informat

4     _FREQ_            Num        8
3     _TYPE_            Num        8
5     credit_lmt        Num        8    BEST12.      BEST32.
2     risk_level        Char       1
6     rscore            Num        8    BEST12.      BEST32.
1     segment           Char       2    $2.          $2.
7     spend             Num        8    BEST12.      BEST32.
```

Print - Summ Data Set



```
                  risk_
Obs    segment    level    _TYPE_    _FREQ_    credit_lmt      rscore      spend

 1       S1         H         3         1           4000          573       2480
 2       S1         L         3         4          25000         3521      13976
 3       S1         M         3         6          23600         4396      13932
 4       S2         H         3         3           7100         1667       2903
 5       S2         L         3         6          34600         5290      20806
 6       S2         M         3         9          36000         6266      21570
 7       S3         H         3         5           8600         2616       5513
 8       S3         L         3         8          51000         7326      31645
 9       S3         M         3         8          32600         5478      19406
```

Excel Computation



Row Labels	Sum of _FREQ_	Sum of Avg Credit Limit	Sum of Avg Spend	Sum of Utilization
⊟ L	18	$6,144	$3,690	60%
S1	4	$6,250	$3,494	56%
S2	6	$5,767	$3,468	60%
S3	8	$6,375	$3,956	62%
⊟ M	23	$4,009	$2,387	60%
S1	6	$3,933	$2,322	59%

Proc Univariate

- Produces statistics describing distribution of a variable
- Statistics include:
  - Moments (mean, standard deviation, skewness, etc..)
  - Basic Statistical measures (mean , median , mode, range etc)
  - Quantiles (Q1, Q3, Med etc ..)
  - Extreme values
- Syntax

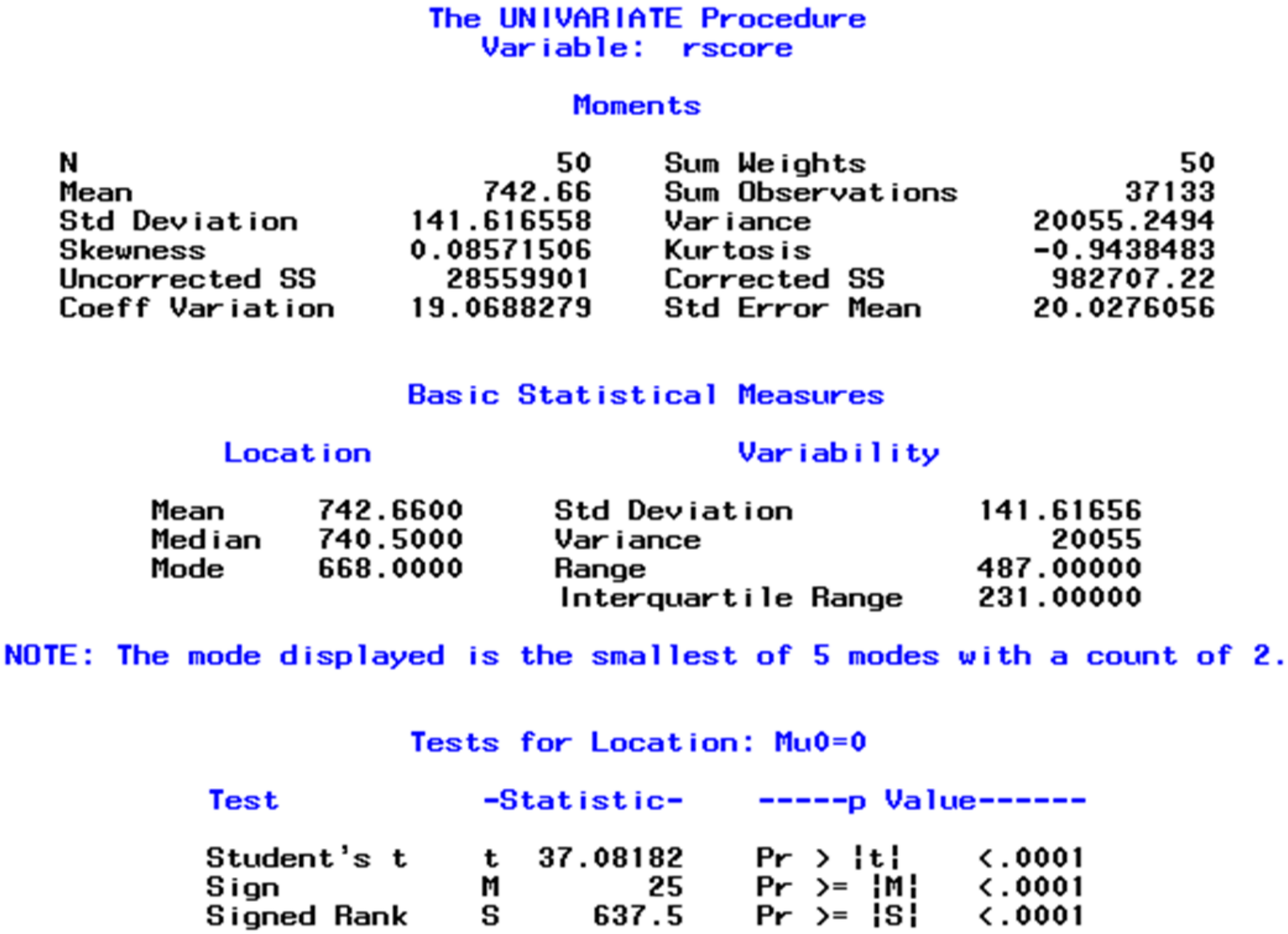
```
proc univariate data=<dataset>;
class <class variables>;
var variable list;
run;
```

Example

Look at the distribution of risk score across performance data

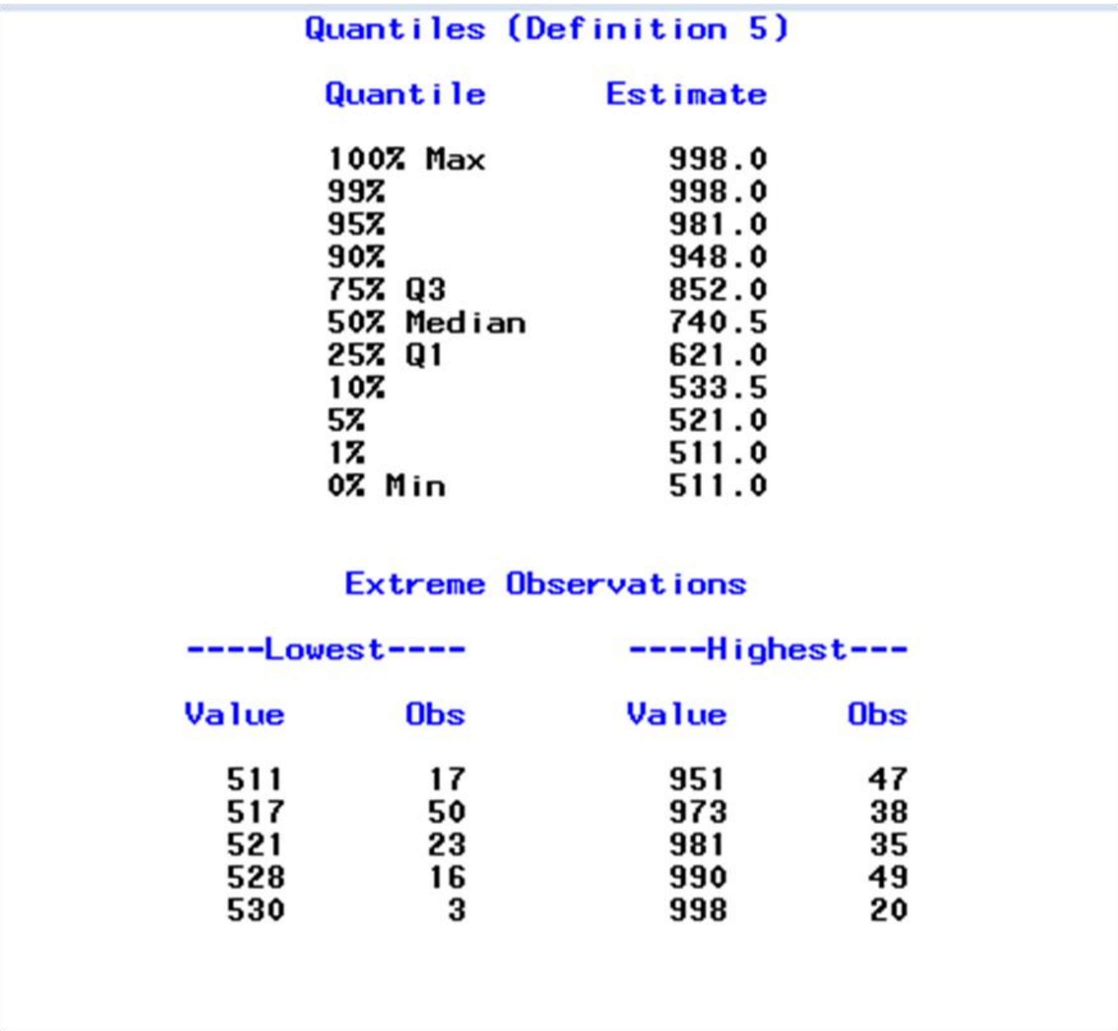
```
proc univariate data=perf;
var rscore;
run;
```

Proc Univariate – SAS Output



```
                      The UNIVARIATE Procedure
                        Variable:  rscore

                              Moments

N                            50    Sum Weights                 50
Mean                     742.66    Sum Observations         37133
Std Deviation        141.616558    Variance            20055.2494
Skewness             0.08571506    Kurtosis            -0.9438483
Uncorrected SS         28559901    Corrected SS         982707.22
Coeff Variation      19.0688279    Std Error Mean      20.0276056

                     Basic Statistical Measures

           Location                    Variability

       Mean     742.6600     Std Deviation           141.61656
       Median   740.5000     Variance                    20055
       Mode     668.0000     Range                   487.00000
                             Interquartile Range     231.00000

NOTE: The mode displayed is the smallest of 5 modes with a count of 2.

                     Tests for Location: Mu0=0

           Test           -Statistic-    -----p Value------

           Student's t    t  37.08182    Pr > |t|    <.0001
           Sign           M        25    Pr >= |M|   <.0001
           Signed Rank    S     637.5    Pr >= |S|   <.0001
```

Proc Univariate – SAS Output

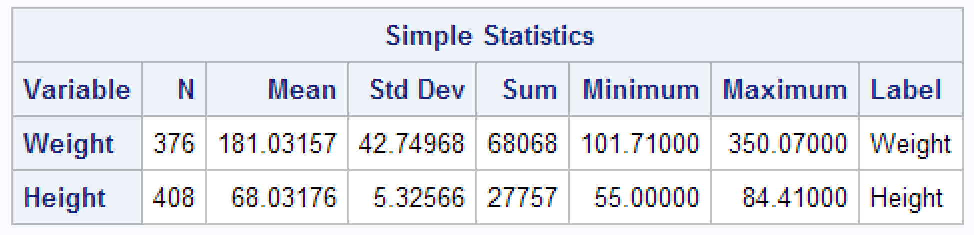


```
               Quantiles (Definition 5)

                Quantile          Estimate

                100% Max             998.0
                99%                  998.0
                95%                  981.0
                90%                  948.0
                75% Q3               852.0
                50% Median           740.5
                25% Q1               621.0
                10%                  533.5
                5%                   521.0
                1%                   511.0
                0% Min               511.0


                 Extreme Observations

        ----Lowest----        ----Highest---

        Value      Obs        Value      Obs

          511       17          951       47
          517       50          973       38
          521       23          981       35
          528       16          990       49
          530        3          998       20
```

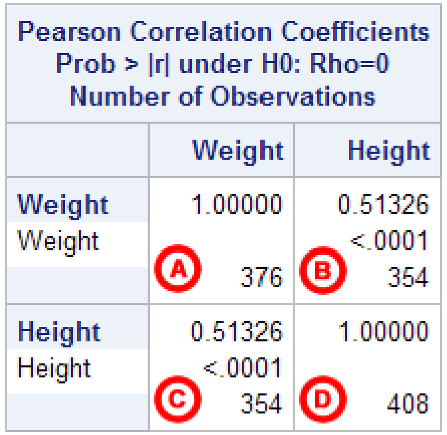
Proc CORR - Pearson Correlation

The correlation or strength of a linear relationship between two continuous numeric variables can be assessed using PROC CORR. This is also known as Pearson Correlation.

```
PROC CORR
DATA=sashelp.class PLOTS=SCATTER(NVAR=all);
VAR height weight;

RUN;
```

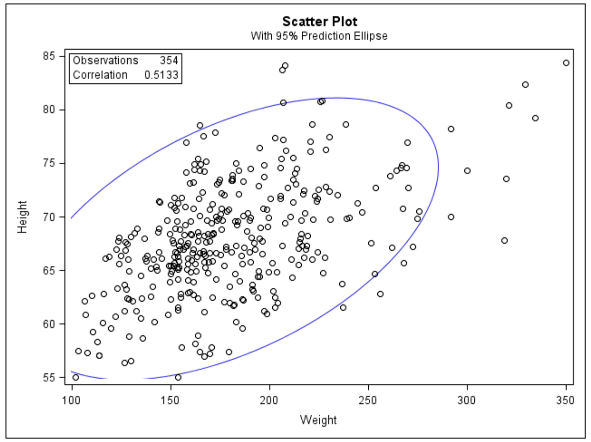


Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Weight	376	181.03157	42.74968	68068	101.71000	350.07000	Weight
Height	408	68.03176	5.32566	27757	55.00000	84.41000	Height



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Weight	Height
Weight Weight	1.00000 A 376	0.51326 <.0001 B 354
Height Height	0.51326 <.0001 C 354	1.00000 D 408



THANK YOU