

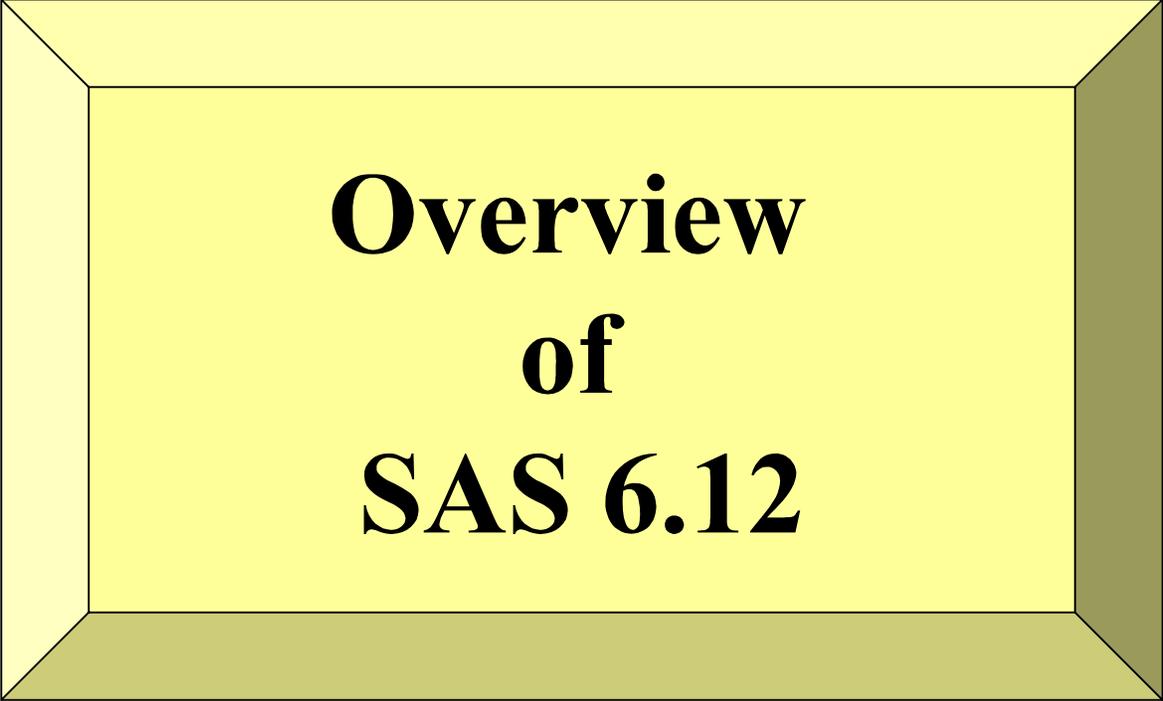
Using SAS for Elementary Statistics

Fairouz Makhlouf

Statistical Support Staff
Center For Information Technology
National Institute of Health

Contents:

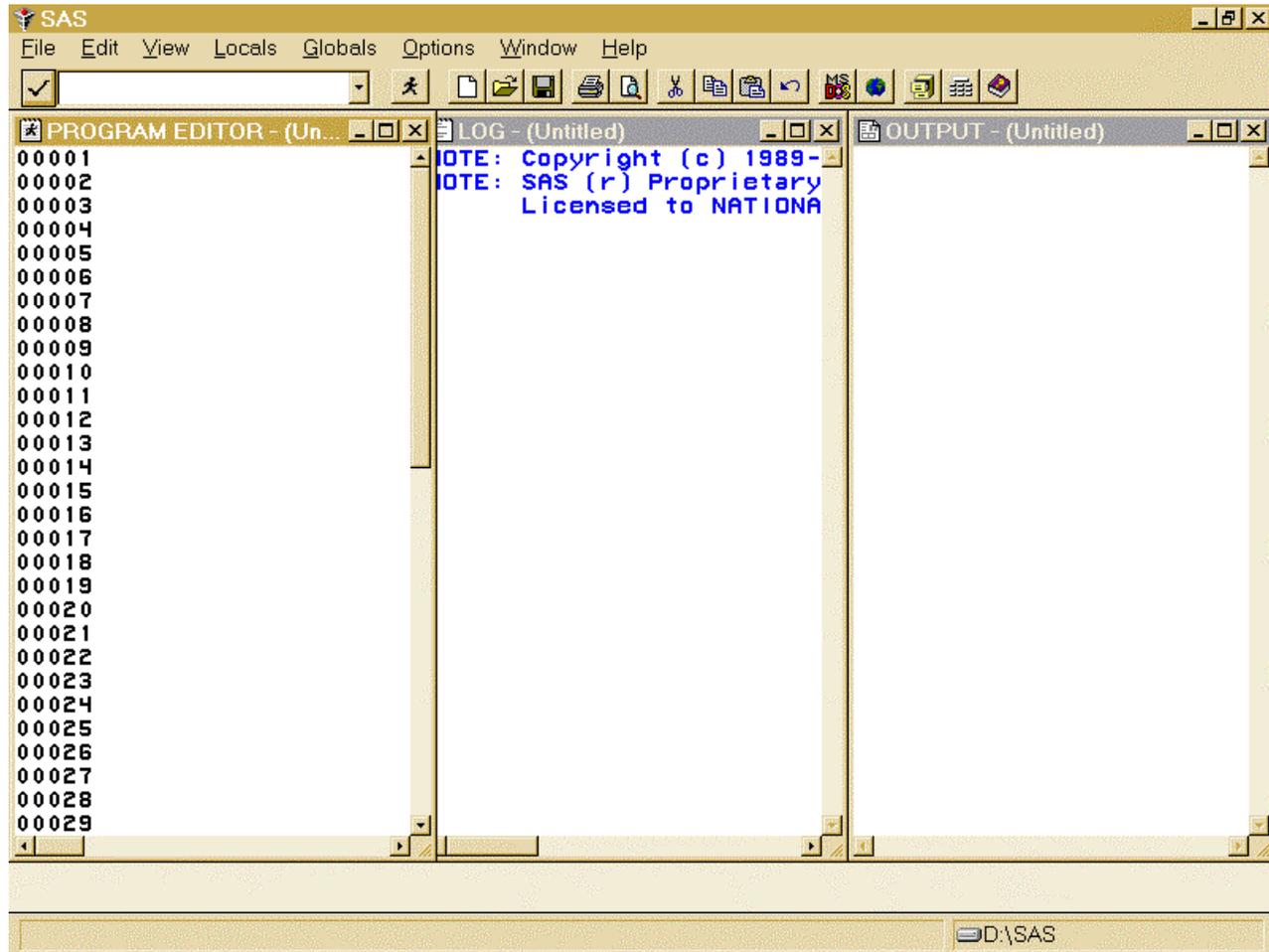
- Overview of SAS 6.12
- Describing DATA
- More Descriptive Statistics
- Descriptive statistics Broken by subgroups
- Frequency Distribution
- Bar Graph
- Plotting Data
- Creating Summary Data sets with **PROC MEANS** ,
PROC UNIVARIATE and **PROC FREQ**
- Outputting statistics other than Means
- Creating a Summary Data Set Containing a Median ²



**Overview
of
SAS 6.12**

Three Important Windows

- **Program Editor**
 - load, create or modify a SAS program
- **LOG**
 - contains the execution log of the session
- **OUTPUT**
 - contains the output of the SAS program

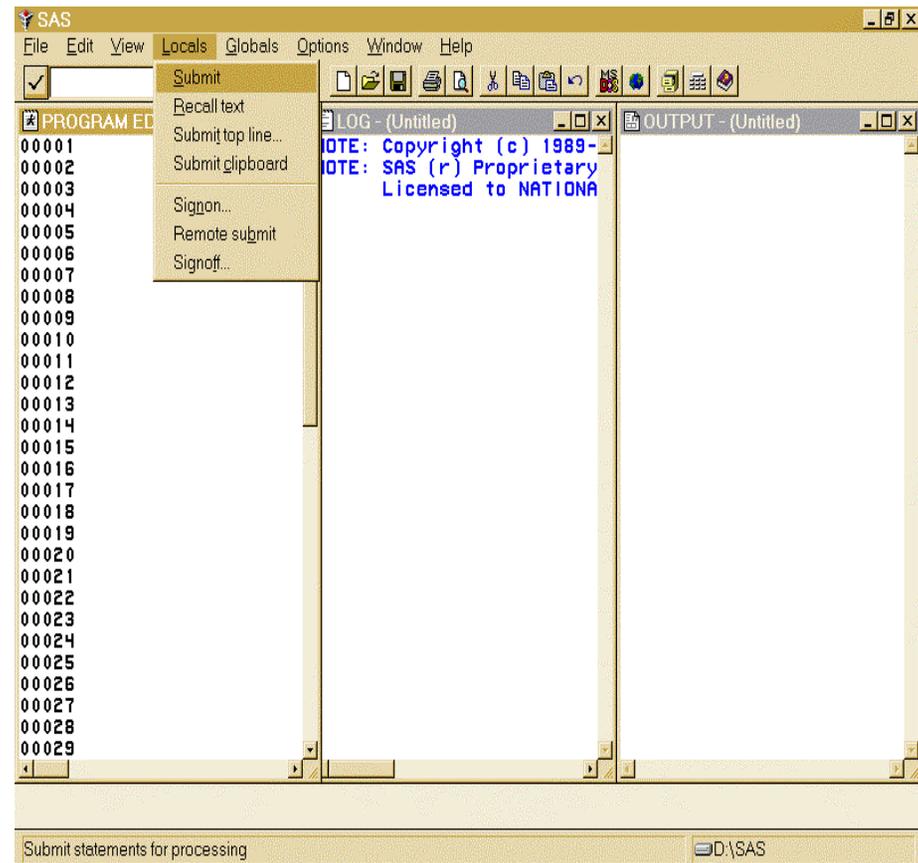


Saving

- The SAS source code should have extension **.sas**
- The log of a session should have **.log**
- The output should have extension **.lst**

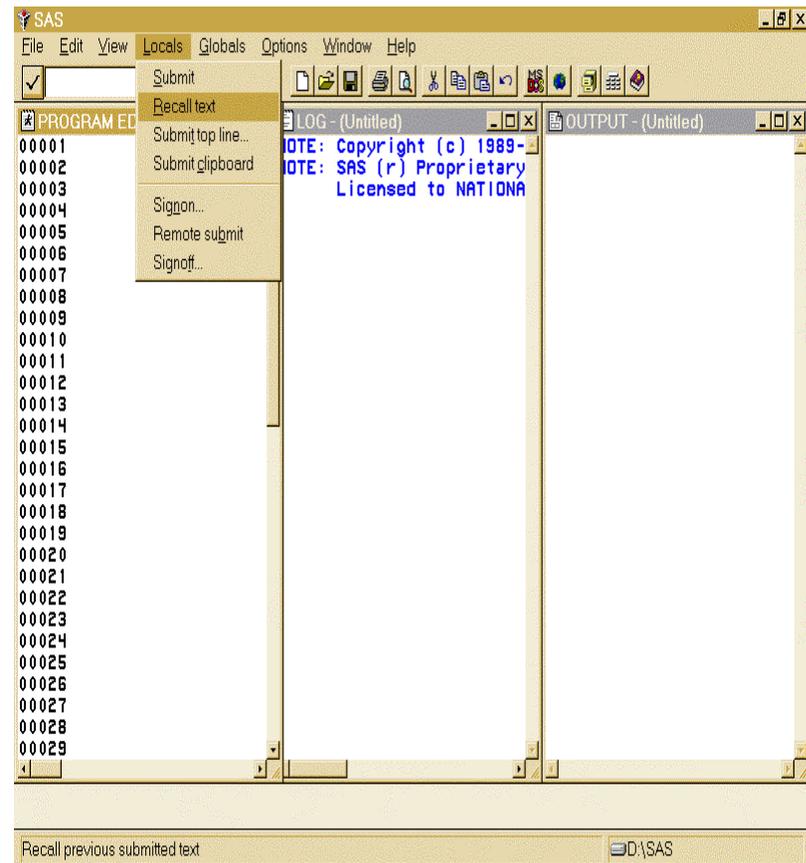
Submit a Job

- Go to **Locals**
 - Choose **Submit**



Recall Text

- Go to **Locals**
 - Choose **Recall text**



Syntax of SAS programs

- Main steps in a SAS program are
- DATA Step
 - includes statements to create SAS datasets
- PROC Step
 - call a procedure from its library

Rules for SAS Statements

- SAS statements should end with a semicolon (;)
- SAS statements are not case sensitive
- SAS programs are freely formatted

Rules for SAS Names

- A SAS name can contain at most 8 characters.
- The first letter must be a character or an underscore (_).
- SAS names are not case sensitive.

Comments in SAS

- `/* This is a comment in SAS. */`
 - Any text between `/*` and `*/` are ignored
- `* This is an comment in SAS. ;`
 - Any text between `*` and `;` are ignored.

DataSets in SAS

- DataSet in SAS is a table-like structure
 - The rows are called observations
 - The columns are called the variables

**SAS
Sample
Program**

```
options nodate nocenter;
```

```
/******  
/* Create a data set called "class" */  
/******
```

```
DATA class;
```

```
input name $ height weight age sex $;
```

```
cards;
```

```
Alfred 69.0 112.5 14 M  
Alice 56.5 84.0 13 F  
Barbara 65.3 98.0 13 F  
Carol 62.8 102.5 14 F  
Henry 63.5 102.5 14 M  
James 57.3 83.0 12 M  
Jane 59.8 84.5 12 F  
Janet 62.5 112.5 15 F  
Jeffrey 62.5 84.0 13 M  
John 59.0 99.5 12 M  
Joyce 51.3 50.5 11 F  
Judy 64.3 90.0 14 F  
Louise 56.3 77.0 12 F  
Mary 66.5 112.0 15 F  
Philip 72.0 150.0 16 M  
Robert 64.8 128.0 12 M  
Ronald 67.0 133.0 15 M  
Thomas 57.5 85.0 11 M  
William 66.5 112.0 15 M
```

```
;
```

```
PROC PRINT DATA=CLASS;
```

```
RUN;
```

NOTE: Copyright (c) 1989-1996 by SAS Institute Inc., Cary, NC, USA.

NOTE: SAS (r) Proprietary Software Release 6.12 TS050
Licensed to NATIONAL INSTITUTES OF HEALTH, Site 0008995010.



```
1  options nodate nocenter;
2
3  /*****
4  /* Create a data set called "class"
5  /*****
6
7  DATA class;
8      input name $ height weight age sex $;
9  cards;
```

NOTE: The data set WORK.CLASS has 19 observations and 5 variables.

NOTE: The DATA statement used 0.29 seconds.

```
29  ;
30
31  PROC PRINT DATA=CLASS;
32  RUN;
```

NOTE: The PROCEDURE PRINT used 0.06 seconds.

**SAS
OUTPUT**

The SAS System

OBS	NAME	HEI GHT	WEI GHT	AGE	SEX
1	Al fred	69.0	112.5	14	M
2	Al i ce	56.5	84.0	13	F
3	Barbara	65.3	98.0	13	F
4	Carol	62.8	102.5	14	F
5	Henry	63.5	102.5	14	M
6	James	57.3	83.0	12	M
7	Jane	59.8	84.5	12	F
8	Janet	62.5	112.5	15	F
9	Jeffrey	62.5	84.0	13	M
10	John	59.0	99.5	12	M
11	Joyce	51.3	50.5	11	F
12	Judy	64.3	90.0	14	F
13	Loui se	56.3	77.0	12	F
14	Mary	66.5	112.0	15	F
15	Phi l i p	72.0	150.0	16	M
16	Robert	64.8	128.0	12	M
17	Ronal d	67.0	133.0	15	M
18	Thomas	57.5	85.0	11	M
19	Wi l l i am	66.5	112.0	15	M

Sources Of Data

- Include raw data in the SAS program
- External file which can be accessed by SAS
- SAS data sets

Temporary SAS data sets

- A temporary SAS data set exists until the end of the current SAS session
- A temporary SAS data set could be created by using 1-level data set name in the DATA statement because the SAS system places the data set in a SAS library referred to as WORK.

Example:

```
DATA class;  
.....  
RUN;
```

This creates a temporary SAS data set, named WORK.class.

Example:

```
DATA temp1;  
  input x y z;  
  Cards;  
1 1 1  
2 4 8  
3 9 27  
4 16 64  
;  
RUN;  
  
PROC PRINT data=temp1;  
  title " DataSet temp1";  
RUN;
```

Temp1 is a temporary SAS data set that was created by including the raw data in the SAS program.

The **INPUT** statement indicates the fields to be read and the SAS variable names to be created.

The **CARDS** statement indicates that the data lines follow. It must appear after the **INPUT** statement and should be followed by a semicolon.

DataSet temp1			
OBS	X	Y	Z
1	1	1	1
2	2	4	8
3	3	9	27
4	4	16	64

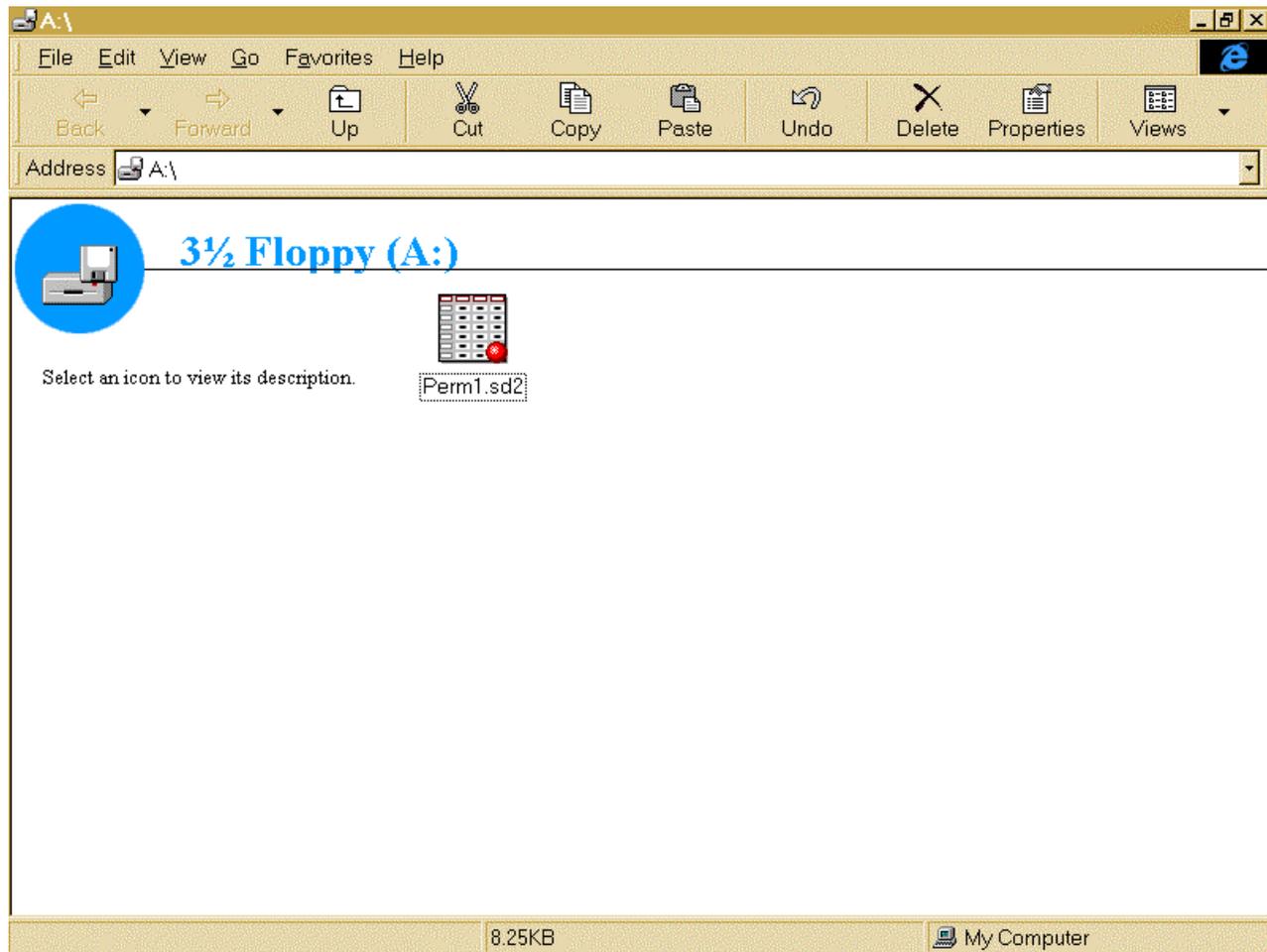
Example:

```
LIBNAME mydir 'a:\';  
DATA mydir.perm1;  
  input x y z;  
  Cards;  
1 1 1  
2 4 8  
3 9 27  
4 16 64  
;  
RUN;  
  
PROC PRINT data=mydir.perm1;  
  title "DataSet perm1";  
RUN;
```

The above DATA step creates a permanent SAS data set called mydir.perm1.

The LIBNAME statement identifies the directory to which the data should be written which is a:\ in this example.

DataSet perm1			
OBS	X	Y	Z
1	1	1	1
2	2	4	8
3	3	9	27
4	4	16	64



Example:

```
DATA temp2;  
  infile 'a:\authors.txt';  
  input F_Name $ L_Name $ Age Income Home $;  
;  
RUN;  
  
PROC PRINT data=temp2;  
  title " DataSet temp2";  
RUN;
```

The above DATA step creates a Temporary SAS data set called WORK.temp2.

This is done by using the **INFILE** statement which must precede the **INPUT** statement.

DataSet temp2					
OBS	F_NAME	L_NAME	AGE	I NCOME	HOME
1	Lorne	Green	82	1200000	LA
2	Loren	Jaye	40	40000	DC
3	Robi n	Green	45	25000	MD
4	Bi l l y	Jaye	40	27500	VA

Creating Variables

- In SAS you can create variables by using assignment statements with the DATA step
- The syntax is
 - new variable name = expression

Example:

```
DATA temp3;  
  infile 'a:\test.txt';  
  input Name $ test1 test2 test3 test4;  
  average = (test1+test2+test3+test4)/4;  
;  
RUN;  
  
PROC PRINT data=temp3;  
  title " DataSet temp3";  
RUN;
```

Average = (test1+test2+test3+test4)/4 is an assignment used to create a new variable “Average”

DataSet temp3

OBS	NAME	TEST1	TEST2	TEST3	TEST4	AVERAGE
1	Sam	11	12	13	14	12.5
2	Hilda	21	22	23	24	22.5
3	Rita	31	32	33	34	32.5
4	Eric	41	42	43	44	42.5
5	Dan	51	52	53	54	52.5
6	Tom	61	62	63	64	62.5
7	Alice	71	72	73	74	72.5
8	Martin	81	82	83	84	82.5
9	Dave	91	92	93	94	92.5



Describing Data

- The most common way of describing data is by reporting
 - the number of observations in the sample (sample size)
 - the mean of the scores
 - the arithmetic mean
 - the standard deviation
 - a measure of how widely spread the scores are
 - if the scores are from a normal distribution then
 - about 68% of the scores fall within 1 standard deviation
 - about 95% of the scores fall within 2 standard deviation
 - about 99.7% of the scores fall within 3 standard deviation

Descriptive Statistics Using SAS

- Using **PROC MEANS**
- Using **PROC UNIVARIATE**

PROC MEANS: *Introduction*

- The MEANS procedure produces simple univariate descriptive statistics for numeric variables.
- PROC MEANS computes statistics for an entire SAS data set or for groups of observations in the data set.
- If you use a **BY** statement, **PROC MEANS** calculates descriptive statistics separately for groups of observations. Each group is composed of observations having the same values of the variables used in the **BY** statement. The groups can be further subdivided by the use of the **CLASS** statement.
- **PROC MEANS** can optionally create one or more SAS data sets containing the statistics calculated.

PROC MEANS: *Syntax*

```
PROC MEANS <option-list> <statistic-keyword-list>;  
  VAR variable-list;  
  BY variable-list;  
  CLASS variable-list;  
  FREQ variable;  
  WEIGHT variable;  
  ID variable-list;  
  OUTPUT <OUT= SAS-data-set> <output-statistic-list>  
    <MINID|MAXID <(var-1<(id-list-1)>  
    <...var-n<(id-list-n)>>>)=name-list>;
```

<option-list>

- ALPHA= value
 - it sets the confidence level used for the confidence limits.
- DATA= SAS-data-set
 - names the SAS data set to be analyzed by PROC MEANS.
 - If DATA= is omitted, the most recently created SAS data set is used
- NOPRINT
 - The NOPRINT option suppresses the printing of the descriptive statistics.
 - Use NOPRINT when the only purpose of using PROC MEANS is to create new SAS data sets.
- MAXDEC= number
 - The MAXDEC option specifies the maximum number of decimal places to use in printing results.
 - Range: 0 - 8
 - Default: 7

- **FW= field-width**
 - The FW= option specifies the field width to use in printing each statistic.
 - Default: 12
- **MISSING**
 - The MISSING option requests that missing values be treated as valid subgroup values for the CLASS variables.
- **NWAY**
 - The NWAY option specifies that statistics for only the observation with the highest `_TYPE_` value (highest level of interaction among CLASS variables) are to be output.
- **IDMIN**
 - The IDMIN option specifies that the value of the ID variable in the output data set should be its minimum rather than its maximum value for the corresponding observations of the input data set.

- **ORDER= INTERNAL | FREQ | DATA | EXTERNAL | FORMATTED**
 - The ORDER= option specifies the order to sort the values of the CLASS variables for display or for the output data set.
 - INTERNAL orders class values by their internal representation.
 - FREQ orders class values by descending frequency count.
 - DATA orders class values as they were ordered in the input data set.
 - EXTERNAL orders class values by their formatted (external)
 - FORMATTED representation.
 - Default: INTERNAL
 - Note: The ORDER= option does not apply to missing values, which are always ordered first.

- **DESCENDING**
 - The DESCENDING option requests that the output data set be ordered by descending `_TYPE_` value.
 - Default: ascending
 - Note: This option has no effect if the `NWAY` option is also specified.
- **VARDEF= DF | WEIGHT | WGT | N | WDF**
 - The VARDEF option specifies the divisor to be used for calculation of the variance.
 - DF uses degrees of freedom (N-1).
 - WEIGHT or WGT uses the sum of the weights.
 - N uses the number of observations (N).
 - WDF uses the sum of the weights minus one.
 - Default: DF

<statistic-keyword-list>

- **N** is the number of observations in the subgroup with non missing values.
- **NMISS** is the number of observations in the subgroup having missing values for the variable.
- **MEAN** is the arithmetic mean, the sum of the values of a variable divided by the number of values.
- **STD** is the standard deviation.
- **MIN** is the minimum value.
- **MAX** is the maximum value.
- **RANGE** is the difference between the maximum and the minimum values.
- **SUM** is the total of all the values of a variable.

- VAR is the variance.
- USS is the uncorrected sum of squares.
- CSS is the corrected sum of squares.
- CV is the coefficient of variation expressed as a percentage.
- STDERR is the standard error of the mean.
- T is Student's t value for testing the hypothesis that the population mean is zero.
- PRT is the probability of a greater absolute value of the Student's t value under the hypothesis that the mean is zero.

- SUMWGT is the sum of the WEIGHT variable values.
- Skewness is a measure of the tendency of the deviations to be larger in one direction than in the other.
 - Measures heaviness of tails--that is, whether some values are very distant from the mean for the population.
- CLM is upper and lower confidence limits.
- LCLM is lower confidence limit.
- UCLM is upper confidence limit.

VAR variable-list

- The VAR statement identifies the numeric variables for which statistics are to be calculated. The results are printed in the order of the variables in the VAR statement.
- If a VAR statement is not used, all numeric variables in the input data set are analyzed, except those listed in BY, CLASS, ID, FREQ, or WEIGHT statements.

BY <DESCENDING> variables...<NOTSORTED>

- A BY statement is used with a procedure to obtain separate analyses on observations in groups defined by the BY variables. The data set being processed need not have been previously sorted by the SORT procedure. However, the data set must be in the same order as though PROC SORT had sorted it unless NOTSORTED is specified.
- If you have used a FORMAT or ATTRIB statement to group a continuous variable into discrete groups, the BY statement creates BY groups based on the formatted values.
- You can also ensure that variables are processed in ascending order by creating an index for one or more variables in the SAS data set. The usage of the BY statement differ in each procedure. Please refer to the Users' Guide for the details.

CLASS variable-list

- The **CLASS** statement assigns the variables used to form subgroups.
- The **CLASS** statement has basically the same effect on the statistics computed as that of the **BY** statement. The differences are in the format of the printed output and in the sorting requirements of the **BY** statement.

FREQ variable

- The **FREQ** statement specifies a numeric variable. Each observation in the input data set is assumed to represent N observations where N is the value of the **FREQ** variable.
- If the value of the **FREQ** variable is less than 1 or is missing, the observation is not used in the calculations. If the value is not an integer, only the integer portion is used.

Example :

For the following data, use SAS to compute some descriptive statistics

Gender	Height	Weight
M	68.5	155
F	61.2	99
F	63.0	115
M	70.0	205
M	68.6	170
F	65.1	125
M	72.4	220

Creating Data Set:

```
DATA HTWT;  
  INPUT SUBJECT GENDER $ HEIGHT WEIGHT;  
CARDS;  
  
1 M 68.5 155  
2 F 61.2 99  
3 F 63.0 115  
4 M 70.0 205  
5 M 68.6 170  
6 F 65.1 125  
7 M 72.4 220  
;
```

Default Options with PROC MEANS

```
PROC MEANS DATA=HTWT;  
  TITLE 'SIMPLE DESCRIPTIVE STATISTICS';  
RUN;
```

SIMPLE DESCRIPTIVE STATISTICS

Variable	N	Mean	Std Dev	Minimum	Maximum
SUBJECT	7	4.0000000	2.1602469	1.0000000	7.0000000
HEIGHT	7	66.9714286	4.0044618	61.2000000	72.4000000
WEIGHT	7	155.5714286	45.7961321	99.0000000	220.0000000

USING THE VAR STATEMENT

```
PROC MEANS DATA=HTWT;  
  TITLE 'SIMPLE DESCRIPTIVE STATISTICS FOR THE WEIGHT';  
  VAR WEIGHT;  
RUN;
```

SIMPLE DESCRIPTIVE STATISTICS FOR THE WEIGHT

Analysis Variable : WEIGHT

N	Mean	Std Dev	Minimum	Maximum
7	155.5714286	45.7961321	99.0000000	220.0000000

USING THE *CLASS* STATEMENT

```
PROC MEANS DATA=HTWT;  
  TITLE 'SIMPLE DESCRIPTIVE STATISTICS USING  
        THE CLASS STATEMENT';  
CLASS GENDER;  
RUN;
```

```
SIMPLE DESCRIPTIVE STATISTICS USING THE BY STATEMENT  
GENDER   N Obs  Variable  N           Mean           Std Dev           Mini mum           Maxi mum  
-----  
F         3  SUBJECT  3           3.6666667        2.0816660        2.0000000        6.0000000  
          3  HEIGHT  3           63.1000000        1.9519221        61.2000000        65.1000000  
          3  WEIGHT  3           113.0000000       13.1148770       99.0000000       125.0000000  
M         4  SUBJECT  4           4.2500000        2.5000000        1.0000000        7.0000000  
          4  HEIGHT  4           69.8750000        1.8172782        68.5000000       72.4000000  
          4  WEIGHT  4           187.5000000       30.1385689       155.0000000      220.0000000  
-----
```

USING THE *BY* STATEMENT

```
PROC SORT DATA=HTWT;  
BY GENDER;  
RUN;
```

```
PROC MEANS DATA=HTWT;  
  TITLE 'SIMPLE DESCRIPTIVE STATISTICS USING  
        THE BY STATEMENT';  
BY GENDER;  
RUN;
```

SIMPLE DESCRIPTIVE STATISTICS USING THE BY STATEMENT

GENDER=F

Variabl e	N	Mean	Std Dev	Mi ni mum	Maxi mum
SUBJECT	3	3.6666667	2.0816660	2.0000000	6.0000000
HEI GHT	3	63.1000000	1.9519221	61.2000000	65.1000000
WEI GHT	3	113.0000000	13.1148770	99.0000000	125.0000000

GENDER=M

Variabl e	N	Mean	Std Dev	Mi ni mum	Maxi mum
SUBJECT	4	4.2500000	2.5000000	1.0000000	7.0000000
HEI GHT	4	69.8750000	1.8172782	68.5000000	72.4000000
WEI GHT	4	187.5000000	30.1385689	155.0000000	220.0000000

Difference between the **CLASS** and the **BY** Statement

- To use the BY statement you need to sort the data first while the CLASS statement does not need sorting. This saves a lot of the processing time. On the other hand the CLASS statement requires more memory.

Some **statistic-keyword-list**

```
PROC MEANS DATA=HTWT N MEAN VAR STD STERR CLM T  
                PRT SKEWNESS KURTOSIS MAXDEC=4  
                USS CSS CV ;  
    TITLE 'SOME STATISTICAL OPTIONS WITH PROC MEANS' ;  
RUN;
```

SOME STATISTICAL OPTIONS WITH PROC MEANS

Variable	N	Mean	Variance	Std Dev	Std Error	Lower 95.0% CLM
SUBJECT	7	4.0000	4.6667	2.1602	0.8165	2.0021
HEIGHT	7	66.9714	16.0357	4.0045	1.5135	63.2679
WEIGHT	7	155.5714	2097.2857	45.7961	17.3093	113.2171

Variable	Upper 95.0% CLM	T	Prob> T	Skewness	Kurtosis	USS
SUBJECT	5.9979	4.8990	0.0027	0.0000	-1.2000	140.0000
HEIGHT	70.6749	44.2481	0.0001	-0.2391	-1.1613	31492.4200
WEIGHT	197.9258	8.9877	0.0001	0.2789	-1.4672	182001.0000

Variable	CSS	CV
SUBJECT	28.0000	54.0062
HEIGHT	96.2143	5.9794
WEIGHT	12583.7143	29.4374

- When the data is a random sample of a population,
 - The **mean** is used as an estimate of the population parameter.
 - The **standard error** of the mean is used to tell us how far this estimate might be from the population mean (The standard deviation divided by the square root of n).
 - The **CLM** option print a 95% confidence interval of the mean. This interval gives us a 95% confident that the true mean is within the limits of the interval.
 - The **T** value is the Student's t-test for testing the null hypothesis that the population mean is 0.
 - The **PRT** is the p value for the preceding t test i.e the probability of obtaining a t-value this large or larger if the population mean were zero.
 - The **CV** is the coefficient of variation and it is a measure of relative dispersion for a data set, found by dividing the standard deviation by the mean and multiplying by 100. It is helpful in comparing the relative dispersion in several variables that have different means and different standard deviation.
 - The **USS** Uncorrected sum of squares (the sum of the scores squared)

- The **CSS** is the corrected sum of squares(sum of squares about the mean)
- **Skewness** refers to the extent to which the sample distribution departs from the normal curve because of a long tail on one side of the distribution.
 - **Positively skewed** when the long tail appears on the right side of the distribution.
 - **Negatively skewed** when the long tail appears on the left side of the distribution.
- **Kurtosis** refers to the extent to which the sample distribution departs from the normal curve because it is either peaked or flat
 - **Positive kurtosis** when the sample distribution is relatively peaked.
 - **Negative kurtosis** when the sample distribution is relatively flat.

PROC UNIVARIATE: *Introduction*

The UNIVARIATE procedure provides detail on the distribution of a variable. Features include:

- detail on the extreme values of a variable
- quantiles, such as the median
- several plots to picture the distribution
- frequency tables
- a test to determine that the data are normally distributed.
- If a BY statement is used, descriptive statistics are calculated separately for groups of observations.

PROC UNIVARIATE: *Syntax*

PROC UNIVARIATE DATA= SASdataset

 NOPRINT

 PLOT

 FREQ

 NORMAL

 PCTLDEF= value

 VARDEF= DF|WEIGHT|WGT|N|WDF

 ROUND= roundoff unit...;

VAR variables;

BY variables;

FREQ variable;

WEIGHT variable;

ID variables;

OUTPUT OUT= SASdataset keyword= names...;

Some OPTIONS

- **PCTLDEF= value**
 - The PCTLDEF= option calculates percentiles. Valid values are 1-5. These values correspond to five computational methods. Refer to the SAS Procedures Guide, Version 6 Edition for an explanation of the computational methods.
 - Default: 5
- **NORMAL**
 - The NORMAL option computes a test statistic for the hypothesis that the input data come from a normal distribution. The probability of a more extreme value of the test statistic is also printed.

- **PLOT**

- The PLOT option causes PROC UNIVARIATE to produce a stem-and-leaf plot (or a horizontal bar chart), a box plot, and a normal probability plot.
- If a BY statement is used, box plots for groups defined by the BY variables also appear side-by-side.

- **ID variables;**

- When an ID statement is used, up to eight characters of the first variable specified are used to identify observations in the printed listing of the five largest and five smallest values.
- The five largest and five smallest values of variables in the VAR statement are printed by PROC UNIVARIATE. If an ID statement is used, the values of the ID variable are used to identify these ten values for each VAR statement variable. If an ID statement is not used, the observation number identifies the values.

Example :

- Produce more descriptive Statistics for the data in example 1 using PROC UNIVARIATE.

```
PROC UNIVARIATE DATA=HTWT NORMAL PLOT;  
    TITLE ' DESCRIPTIVE STATISTICS USING  
          PROC UNIVARIATE' ;  
VAR HEIGHT WEIGHT;  
RUN;
```

DESCRIPTIVE STATISTICS USING PROC UNIVARIATE

Univariate Procedure

Variable=HEIGHT

Moments				Quantiles(Def=5)			
N	7	Sum Wgts	7	100% Max	72.4	99%	72.4
Mean	66.97143	Sum	468.8	75% Q3	70	95%	72.4
Std Dev	4.004462	Variance	16.03571	50% Med	68.5	90%	72.4
Skewness	-0.23905	Kurtosis	-1.16132	25% Q1	63	10%	61.2
USS	31492.42	CSS	96.21429	0% Min	61.2	5%	61.2
CV	5.979358	Std Mean	1.513544			1%	61.2
T: Mean=0	44.24808	Pr> T	0.0001	Range	11.2		
Num ^= 0	7	Num > 0	7	Q3-Q1	7		
M(Sign)	3.5	Pr>= M	0.0156	Mode	61.2		
Sgn Rank	14	Pr>= S	0.0156				
W: Normal	0.954727	Pr<W	0.7849				

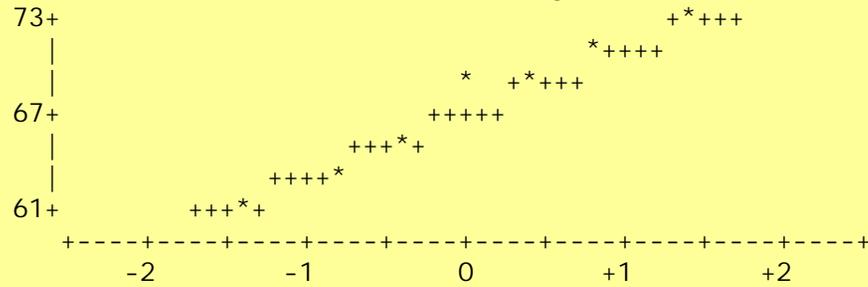
Extremes

Lowest	Obs	Highest	Obs
61.2(2)	65.1(6)
63(3)	68.5(1)
65.1(6)	68.6(5)
68.5(1)	70(4)
68.6(5)	72.4(7)

Stem Leaf	#	Boxplot
72 4	1	
70 0	1	+-----+
68 56	2	*-----*
66		+
64 1	1	
62 0	1	+-----+
60 2	1	

-----+-----+-----+-----+

Normal Probability Plot



DESCRIPTIVE STATISTICS USING PROC UNIVARIATE

Univariate Procedure

Variable=WEIGHT

Moments				Quantiles(Def=5)			
N	7	Sum Wgts	7	100% Max	220	99%	220
Mean	155.5714	Sum	1089	75% Q3	205	95%	220
Std Dev	45.79613	Variance	2097.286	50% Med	155	90%	220
Skewness	0.278915	Kurtosis	-1.46723	25% Q1	115	10%	99
USS	182001	CSS	12583.71	0% Min	99	5%	99
CV	29.43737	Std Mean	17.30931			1%	99
T: Mean=0	8.987731	Pr> T	0.0001	Range	121		
Num ^= 0	7	Num > 0	7	Q3-Q1	90		
M(Sign)	3.5	Pr>= M	0.0156	Mode	99		
Sgn Rank	14	Pr>= S	0.0156				
W: Normal	0.941255	Pr<W	0.6674				

Extremes

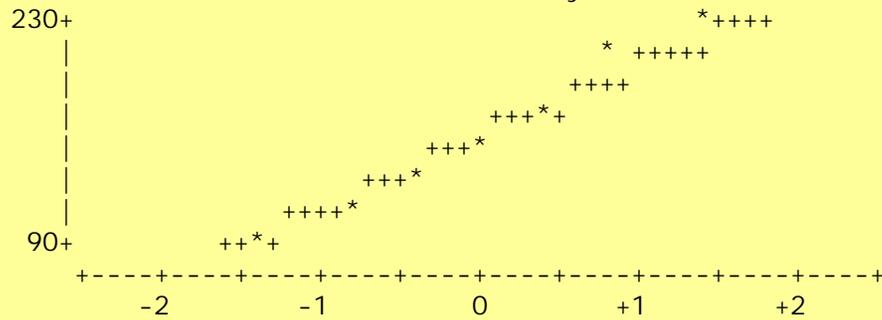
Lowest	Obs	Highest	Obs
99(2)	125(6)
115(3)	155(1)
125(6)	170(5)
155(1)	205(4)
170(5)	220(7)

Stem Leaf	#	Boxplot
22 0	1	
20 5	1	+-----+
18		
16 0	1	
14 5	1	*-----*
12 5	1	
10 5	1	+-----+
8 9	1	

-----+-----+-----+-----+

Multiply Stem. Leaf by 10**+1

Normal Probability Plot



- Sgn Rank is the Wilcoxon signed rank sum (usually used for difference scores)
- Prob >|S| is the p-value for the Sign Rank test
- Num ^=0 is the Number of nonzero observation
- W:Normal Shapiro-Wilk statistic for a test of normality (if the sample size is > 2000 SAS will produce kolomogrove D:Normal test)
- Prob<W is the P-value testing the null hypothesis that the population is normally distributed (when the D:Normal test is done, the statistic is Prob>D)
- A five number summary (Min, Q1 <the first quartile>, Median, Q3 <the third quartile>, Max)

- Some other important percentiles like the 99%, 95% etc...
- Range, Interquartile Range(Q3-Q1), Mode
- Extremes are the five highest and the five lowest values. Next to each extreme value is the corresponding observation number. This can be made more useful if the ID statement is used.
- Stem-and-leaf Plot (explain)
- BoxPlot (explain)
- Normal probability Plot which is used to compare the cumulative frequency distribution to a normal distribution. (explain)

PROC FREQ: *Introduction*

- Frequency tables show the distribution of variable values. Crosstabulation tables show combined frequency distributions for two or more variables.
- For one-way tables, **PROC FREQ** can compute chi-square tests for equal or specified proportions. For two-way tables, **PROC FREQ** computes tests and measures of association. For n-way tables, **PROC FREQ** does stratified analysis, computing statistics within as well as across strata.

PROC FREQ: *Syntax*

PROC FREQ options;

OUTPUT <OUT= SAS-data-set><output-statistic-list>;

TABLES requests / options;

WEIGHT variable;

EXACT statistic-keywords;

BY variable-list;

Some OPTIONS

- The NOCOL option suppresses printing of column percentages in cells of a crosstabulation.
- The NOCUM option suppresses printing of cumulative frequencies and cumulative percentages for one-way frequencies and for frequencies in LIST format.
- The NOPERCENT option suppresses printing of cell percentages for a crosstabulation and also suppresses printing of percentages and cumulative percentages for one-way frequencies and for frequencies in LIST format.

NOCUM *NOPERCENT* OPTIONS

```
PROC FREQ DATA=HTWT ;  
  TITLE ' USING PROC FREQ TO COMPUTE FREQUENCIES' ;  
  TABLES GENDER/ NOCUM NOPERCENT;  
RUN;
```

USING PROC FREQ TO COMPUTE FREQUENCIES

GENDER	Frequency
<i>ffffffffffffffffffff</i>	
F	3
M	4

BAR GRAPH

- To produce **BAR GRAPH** we will use the **CHART** procedure which produces vertical and horizontal bar charts (histograms), block charts, pie charts, and star charts. These charts are useful for showing pictorially a variable's values or the relationships between two or more variables.

PROC CHART: *Introduction*

The appearance of a chart produced by PROC CHART is determined by three factors:

1. the method chosen to present the chart
2. the summary measures that are shown for the variable whose values are charted (the chart variable)
3. features of the procedure that specify how the values are grouped.

PROC CHART produces charts for both numeric and character variables. Character variables and formats cannot exceed a length of sixteen.

Multiple charts may be requested in a single CHART procedure.

PROC CHART: *Syntax*

PROC CHART DATA= SAS-data-set

LPI= p

FORMCHAR (index list) = 'formchar-string';

BY <variable-list>;

VBAR variable-list </ <option-list>>;

HBAR variable-list </ <option-list>>;

BLOCK variable-list </ <option-list>>;

PIE variable-list </ <option-list>>;

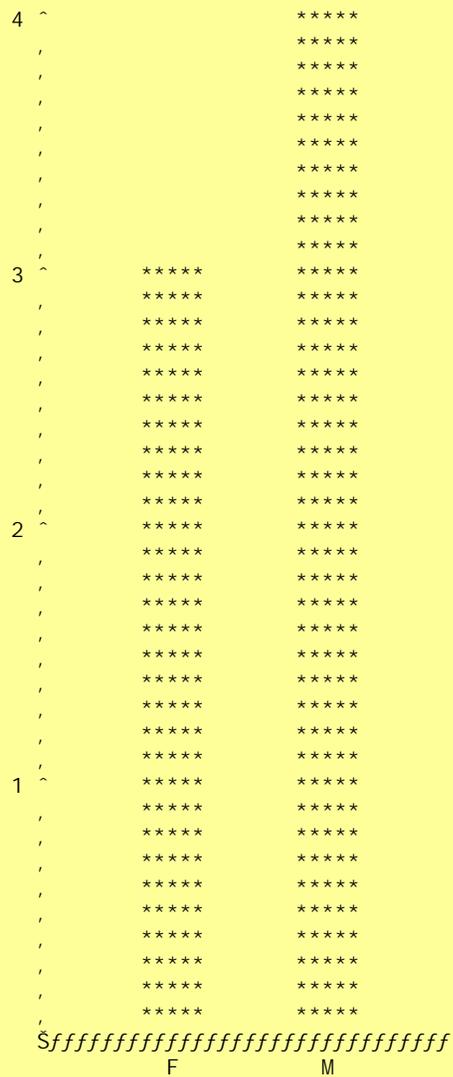
STAR variable-list </ <option-list>>;

VBAR/HBAR OPTION

```
PROC CHART DATA=HTWT;  
  TITLE 'VBAR OPTIONS' ;  
  VBAR GENDER;  
RUN;
```

```
PROC CHART DATA=HTWT;  
  TITLE 'HBAR OPTIONS' ;  
  HBAR GENDER;  
RUN;
```

VBAR OPTI ON
Frequency



GENDER

HBAR OPTI ON
GENDER

		Freq	Cum. Freq	Percent	Cum. Percent
F	*****	3	3	42.86	42.86
M	*****	4	7	57.14	100.00

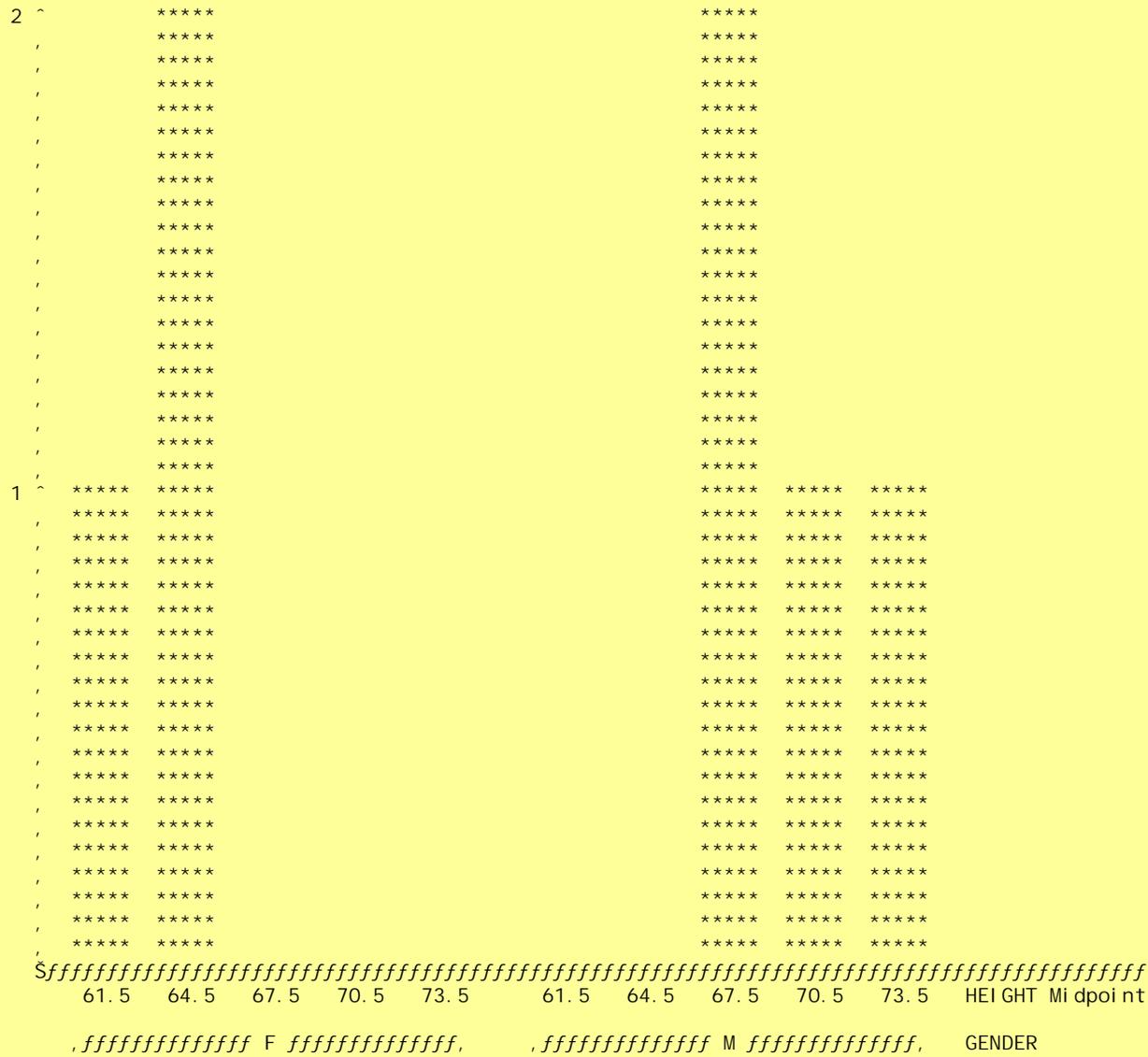
~\$ffffffffff~ffffffffff~ffffffffff~ffffffffff~
 1 2 3 4
 Frequency

LEVELS OPTION

```
PROC CHART DATA=HTWT;  
  TITLE 'DISTRIBUTION OF HEIGHTS' ;  
  VBAR HEIGHT / LEVELS=6;  
RUN;
```


DISTRIBUTION OF HEIGHTS BY GENDER

Frequency



PROC PLOT: *Introduction*

- The PLOT procedure graphs one variable against another, producing a printer plot. The coordinates of each point on the plot correspond to the two variables' values in one or more observations of the input data set.

PROC PLOT: *Syntax*

```
PROC PLOT <option-list>;  
    PLOT request-list </ options >;  
    BY variable-list;  
RUN;
```

```
PROC PLOT DATA=data_set_name;  
    PLOT Y variable * X variable;  
        (Vertical)      (Horizontal)  
    BY variable-list;  
RUN;
```

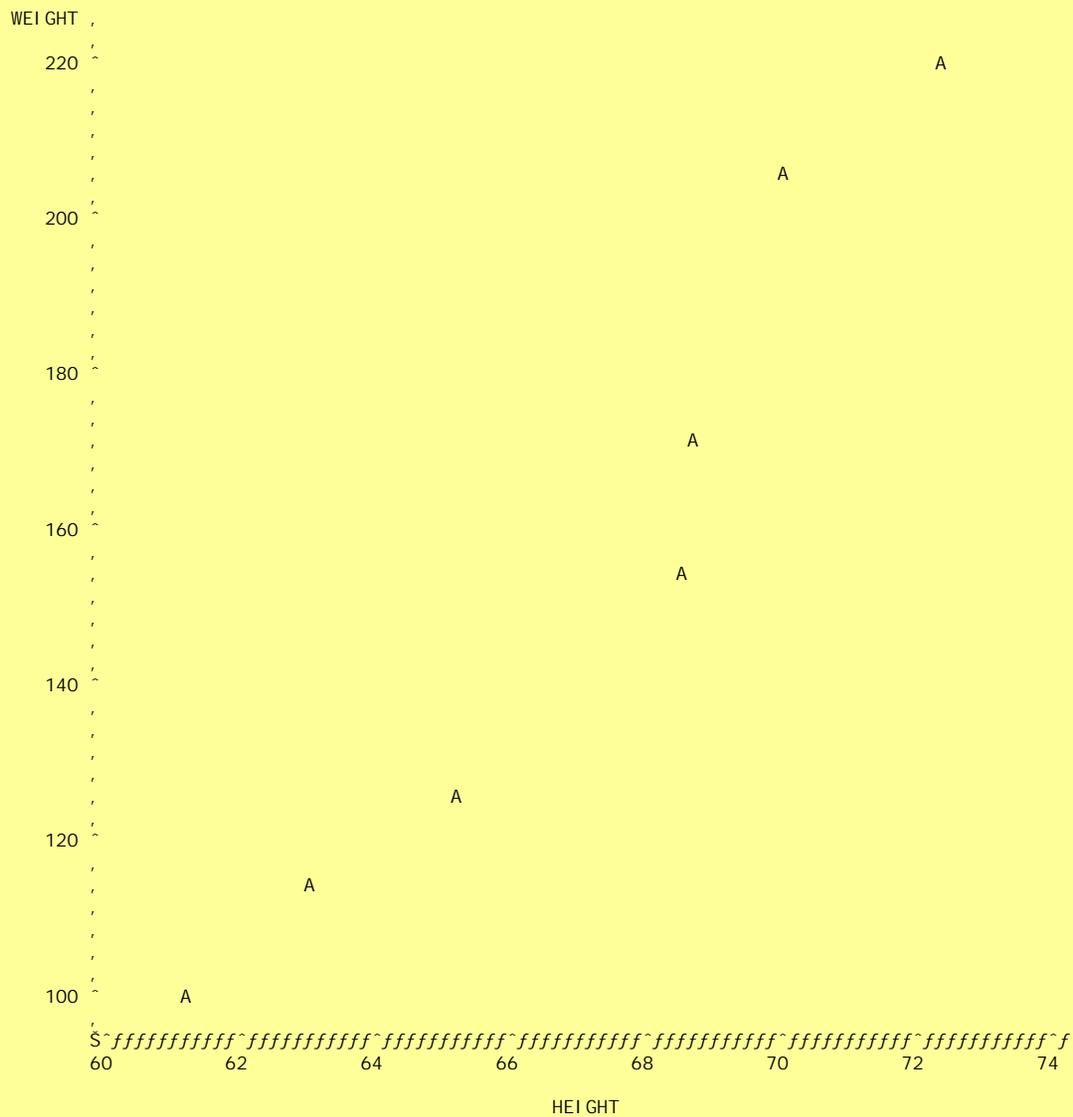
Example:

- Generate a Graph of HEIGHT versus WEIGHT using the data in example 1.

```
PROC PLOT DATA=HTWT;  
    TITLE ' PLOT OF HEIGHT VERSUS WEIGHT' ;  
    PLOT WEIGHT*HEIGHT;  
RUN;
```

PLOT OF HEIGHT VERSUS WEIGHT

Plot of WEIGHT*HEIGHT. Legend: A = 1 obs, B = 2 obs, etc.



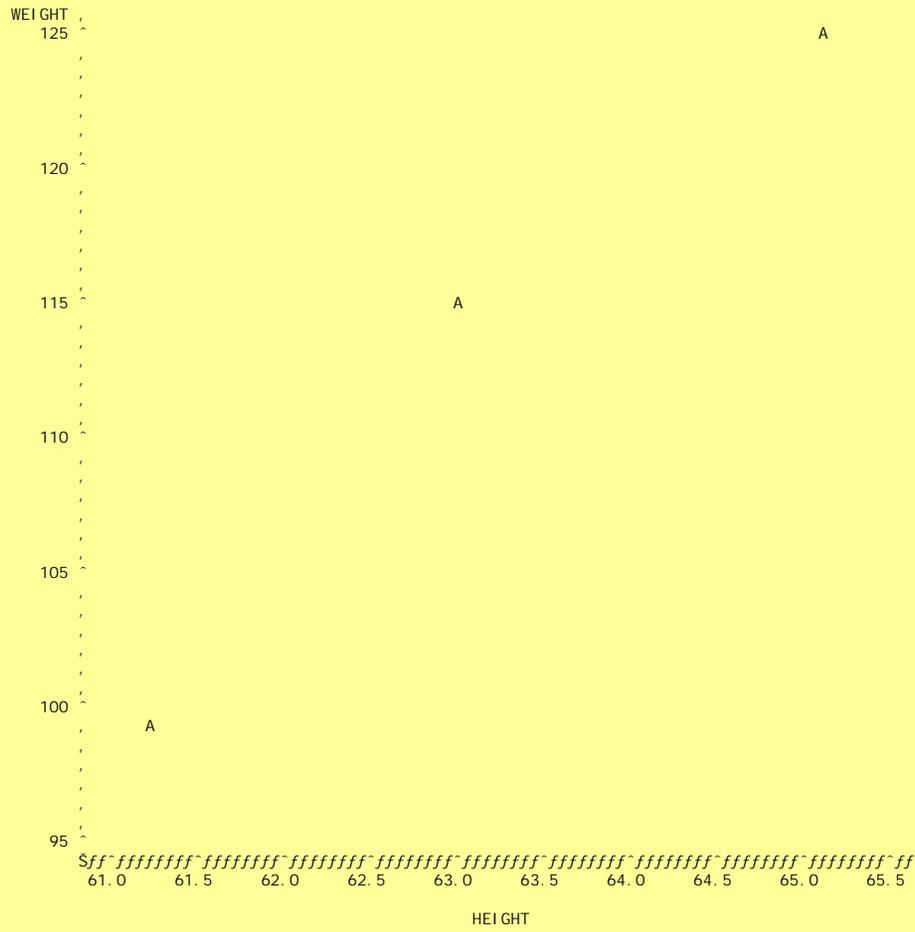
- Generate a separate Graph for males and another for female of HEIGHT versus WEIGHT using the data in example 1 .

```
PROC SORT DATA=HTWT;  
BY GENDER;  
RUN;  
  
PROC PLOT DATA=HTWT;  
  TITLE ' PLOT OF HEIGHT VERSUS WEIGHT BY GENDER' ;  
  PLOT WEIGHT*HEIGHT;  
  BY GENDER;  
RUN;
```

PLOT OF HEIGHT VERSUS WEIGHT BY GENDER

GENDER=F

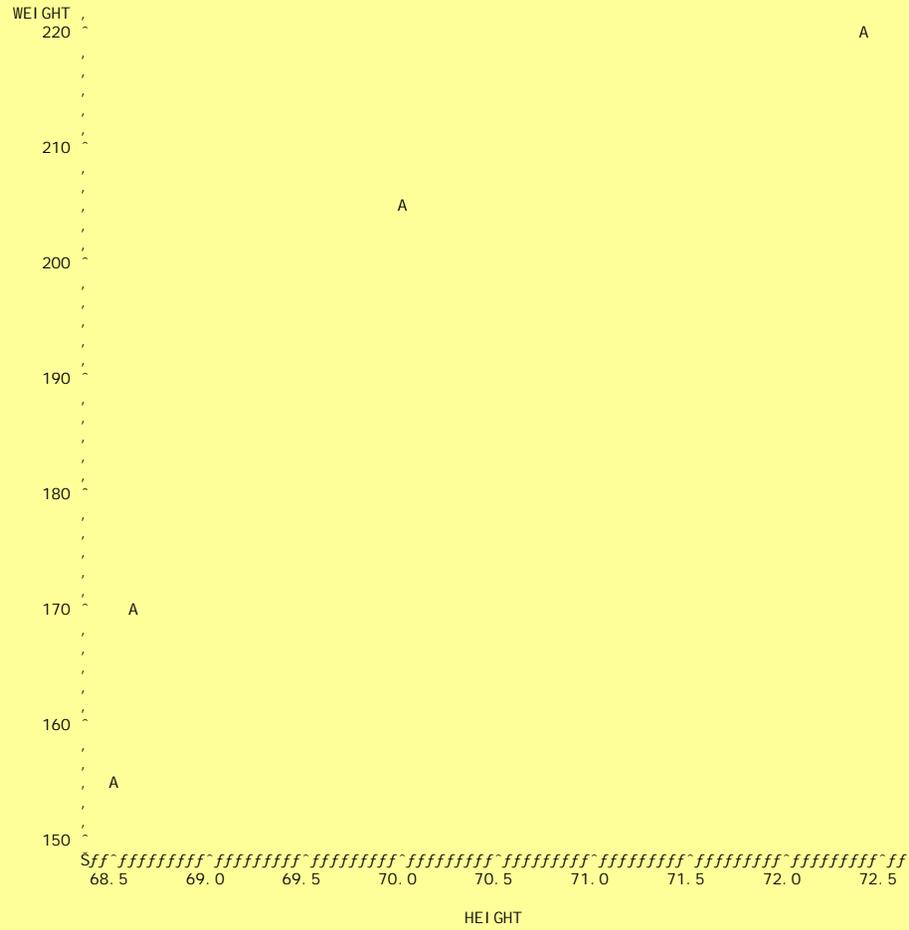
Plot of WEIGHT*HEIGHT. Legend: A = 1 obs, B = 2 obs, etc.



PLOT OF HEIGHT VERSUS WEIGHT BY GENDER

GENDER=M

Plot of WEIGHT*HEIGHT. Legend: A = 1 obs, B = 2 obs, etc.

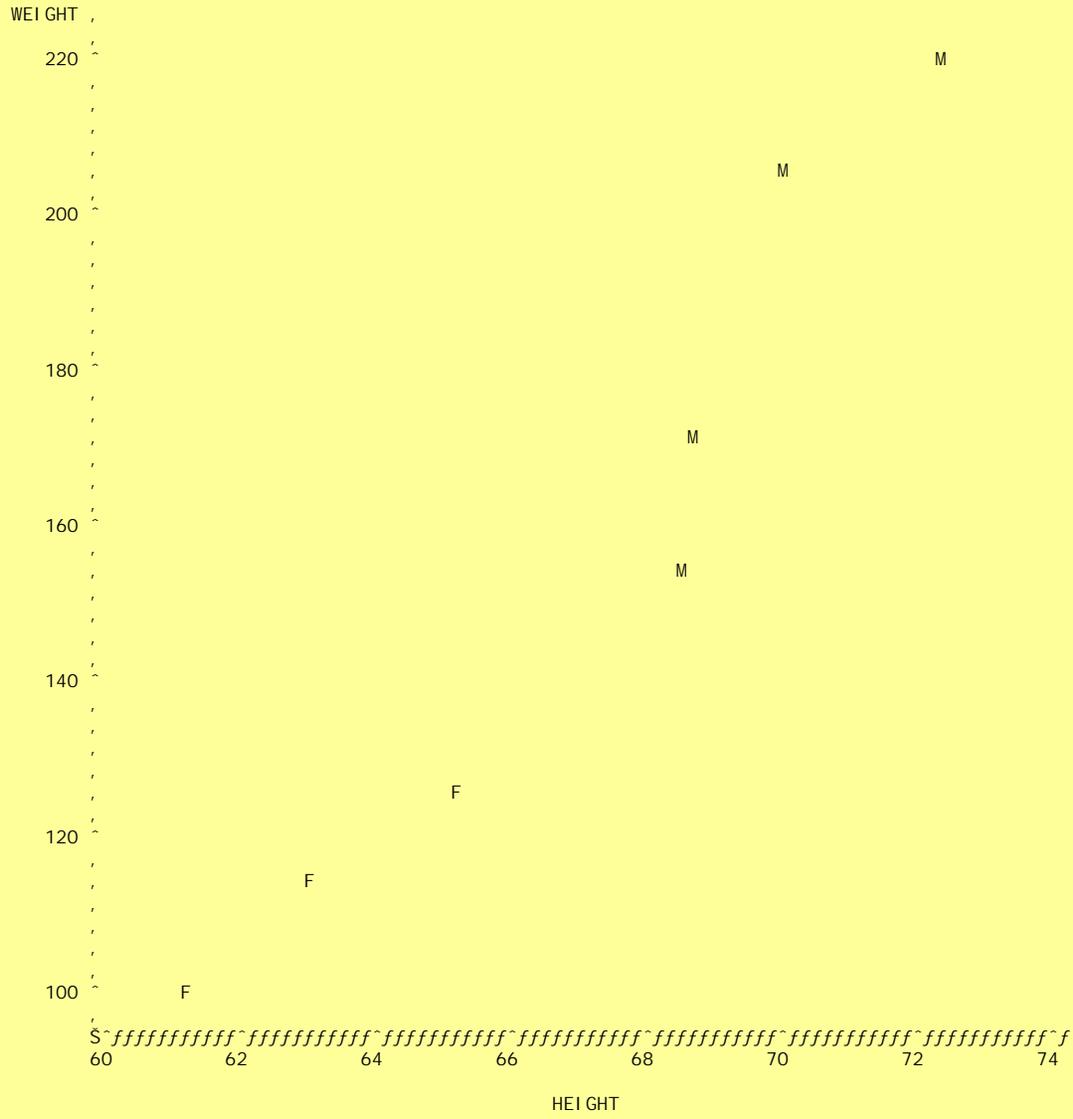


- Generate a Graph of HEIGHT versus WEIGHT using the data in example 1 and using gender as the symbol (F as female and M as male).

```
PROC PLOT DATA=HTWT;  
    TITLE ' PLOT OF HEIGHT VERSUS WEIGHT' ;  
    PLOT WEIGHT*HEIGHT=GENDER;  
RUN;
```

PLOT OF HEIGHT VERSUS WEIGHT

Plot of WEIGHT*HEIGHT. Symbol is value of GENDER.



Creating Summary Data Sets with *PROC MEANS* and *PROC UNIVARIATE*

```
PROC MEANS DATA=HTWT NOPRINT;  
  TITLE' OUTPUT DATA SUMMARY BY PROC MEANS' ;  
VAR HEIGHT WEIGHT;  
OUTPUT OUT=SUMMARY MEAN= HMEAN WMEAN;  
RUN;  
  
PROC PRINT DATA=SUMMARY;  
RUN;
```

OUTPUT DATA SUMMARY BY PROC MEANS

OBS	_TYPE_	_FREQ_	HMEAN	WMEAN
1	0	7	66.9714	155.571

```

PROC MEANS DATA=HTWT NOPRINT;
  TITLE' OUTPUT DATA SUMMARY BY PROC MEANS' ;
VAR HEIGHT WEIGHT;
CLASS GENDER;
OUTPUT OUT=SUMMARY MEAN= HMEAN WMEAN;
RUN;

PROC PRINT DATA=SUMMARY;
RUN;

```

OUTPUT DATA SUMMARY BY PROC MEANS

OBS	GENDER	_TYPE_	_FREQ_	HMEAN	WMEAN
1		0	7	66.9714	155.571
2	F	1	3	63.1000	113.000
3	M	1	4	69.8750	187.500

```

PROC MEANS DATA=HTWT NOPRINT NWAY;
  TITLE' OUTPUT DATA SUMMARY BY PROC MEANS' ;
VAR HEIGHT WEIGHT;
CLASS GENDER;
OUTPUT OUT=SUMMARY MEAN= HMEAN WMEAN;
RUN;

PROC PRINT DATA=SUMMARY;
RUN;

```

OUTPUT DATA SUMMARY BY PROC MEANS

OBS	GENDER	_TYPE_	_FREQ_	HMEAN	WMEAN
1	F	1	3	63.100	113.0
2	M	1	4	69.875	187.5

```

PROC MEANS DATA=HTWT NOPRINT NWAY;
  TITLE' OUTPUT DATA SUMMARY BY PROC MEANS' ;
VAR HEIGHT WEIGHT;
CLASS GENDER;
OUTPUT OUT=SUMMARY MEAN= HMEAN WMEAN
                  N = H_NUM W_NUM
                  STD = H_STD W_STD
                  MAX =H_MAX W_MAX;

RUN;

PROC PRINT DATA=SUMMARY;
RUN;

```

OUTPUT DATA SUMMARY BY PROC MEANS

OBS	GENDER	_TYPE_	_FREQ_	HMEAN	WMEAN	H_NUM	W_NUM	H_STD	W_STD	H_MAX	W_MAX
1	F	1	3	63.100	113.0	3	3	1.95192	13.1149	65.1	125
2	M	1	4	69.875	187.5	4	4	1.81728	30.1386	72.4	220

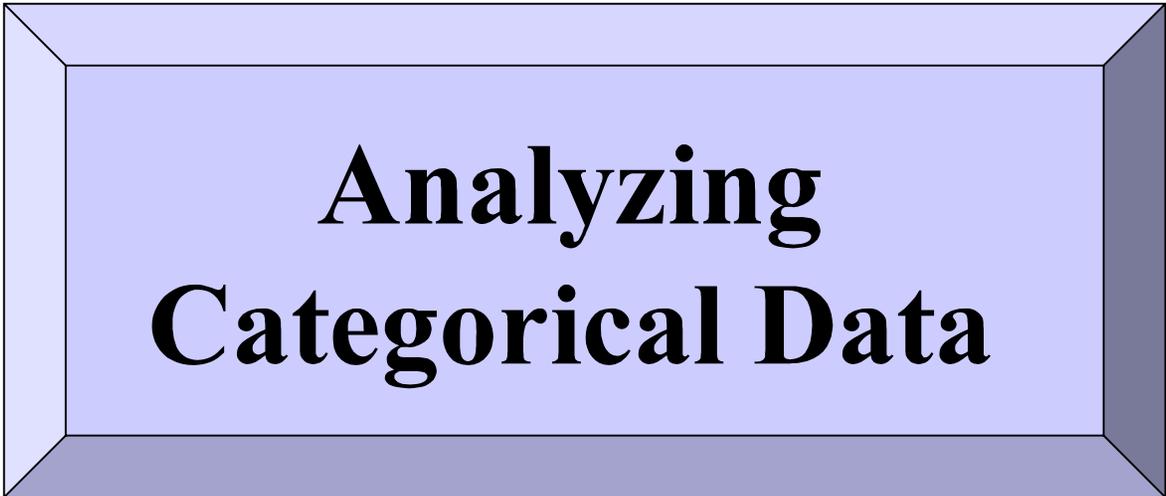
```

PROC SORT DATA=HTWT;
BY GENDER;
RUN;
PROC UNIVARIATE DATA=HTWT NOPRINT ;
  TITLE' OUTPUT DATA SUMMARY BY PROC UNIVARIATE' ;
VAR HEIGHT WEIGHT;
BY GENDER;
OUTPUT OUT=UNIVOUT
      N = H_NUM W_NUM
      MEAN= HMEAN WMEAN
      MEDIAN =H_MED W_MED
      PCTLPRE=HPER_ WPER_
      PCTLPTS=30 70;
RUN;
PROC PRINT DATA=UNIVOUT;
RUN;

```

OUTPUT DATA SUMMARY BY PROC UNIVARIATE

OBS	GENDER	H_NUM	W_NUM	HMEAN	WMEAN	H_MED	W_MED	HPER_30	HPER_70	WPER_30	WPER_70
1	F	3	3	63.100	113.0	63.0	115.0	61.2	65.1	99	125
2	M	4	4	69.875	187.5	69.3	187.5	68.6	70.0	170	205



**Analyzing
Categorical Data**

Contents:

- Description of the data Used in this section
- Adding Variables Labels
- Adding “Value Labels” (Format)
- Recoding Data
- Using a Format to Recode a Variable
- Two-way Frequency Tables
- McNemar’s Test for Paired Data
- Odds Ratio
- Chi-square Test for Trend

- McNemar's Test for Paired Data
- Odds Ratio
- Chi-square Test for Trend

DATA & Data Description

Column	Description	Variable Name
1-3	Subject ID	ID
4-5	Age in years	AGE
6	Gender	GENDER
7	Race	RACE
8	Marital status	MARITAL
9	Education level	EDUC
10	President doing good job	PRES
11	Arms budget increased	ARMS
12	Federal aid to cities	CITIES

001091111232
002452222422
003351324442
004271111121
005682132333
006651243425

- Gender
 - 1=Male
 - 2=Female
- Race
 - 1=White
 - 2=African American
 - 3=Hispanic
 - 4=Other
- Marital Status
 - 1=Single
 - 2=Married
 - 3=Widowed
 - 4=Divorced
- Education Levels
 - 1=High school or less
 - 2=Two year college
 - 3=Four year college (B.A. or B.S.)
 - 4=Post graduate degree
- President doing good job
- Arms budget increased
- Federal aid to cities
 - 1=Strongly disagree
 - 2=Disagree
 - 3=Neutral
 - 4=Agree
 - 5=Strongly agree

DATA QUEST;

INPUT ID	1-3
AGE	4-5
GENDER	\$ 6
RACE	\$ 7
MARITAL	\$ 8
EDUC	\$ 9
PRES	10
ARMS	11
CITIES	12;

CARDS;

001091111232

002452222422

003351324442

004271111121

005682132333

006651243425

;

PROC MEANS DATA=QUEST MAXDEC=2 N MEAN STD;

TITLE ' SOME DESCRIPTIVE STATISTICS FOR THE AGE VARIABLE' ;

VAR AGE;

RUN;

PROC FREQ DATA=QUEST;

TITEL ' FREQUENCY COUNTS FOR CATEGORICAL VARIABLES' ;

TABLES GENDER RACE MARITAL EDUC PRES ARMS CITIES;

RUN;

SOME DESCRIPTIVE STATISTICS FOR THE AGE VARIABLE

Analysis Variable : AGE

N	Mean	Std Dev
6	41.50	22.70

FREQUENCY COUNTS FOR CATEGORICAL VARIABLES

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	66.7	4	66.7
2	2	33.3	6	100.0

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	50.0	3	50.0
2	2	33.3	5	83.3
3	1	16.7	6	100.0

MARITAL	Frequency	Percent	Cumul ati ve Frequency	Cumul ati ve Percent
1	2	33.3	2	33.3
2	2	33.3	4	66.7
3	1	16.7	5	83.3
4	1	16.7	6	100.0

EDUC	Frequency	Percent	Cumul ati ve Frequency	Cumul ati ve Percent
1	2	33.3	2	33.3
2	2	33.3	4	66.7
3	1	16.7	5	83.3
4	1	16.7	6	100.0

PRES	Frequency	Percent	Cumul ati ve Frequency	Cumul ati ve Percent
1	1	16.7	1	16.7
2	1	16.7	2	33.3
3	1	16.7	3	50.0
4	3	50.0	6	100.0

Adding Variable Labels

LABEL

```
variable1='label1' <...variablen='labeln'>;
```

- The 'labeln' can contain up to 40 characters (each blank counts as a character)
- Must be enclosed in a single or double quotes (but not a mixture)
- It could be placed any place in the DATA Step.

Example:

```
DATA QUEST;  
  INPUT ID      1-3  
    AGE        4-5  
    GENDER    $  6  
    RACE      $  7  
    MARI TAL  $  8  
    EDUC      $  9  
    PRES      10  
    ARMS      11  
    CIT IES   12;  
  LABEL MARI TAL=' Marital Status'  
    EDUC=' Education Level '  
    PRES=' President Doing a Good Job'  
    ARMS=' Arms Budget Increase'  
    CIT IES=' Federal Aid to Cities';  
  
CARDS;  
001091111232  
002452222422  
003351324442  
004271111121  
005682132333  
006651243425  
;
```

```
PROC MEANS DATA=QUEST MAXDEC=2 N MEAN STD;  
  TITLE ' SOME DESCRIPTIVE STATISTICS FOR THE AGE VARIABLE' ;  
VAR AGE;  
RUN;  
  
PROC FREQ DATA=QUEST;  
  TITEL ' FREQUENCY COUNTS FOR CATEGORICAL VARIABLES' ;  
  TABLES GENDER RACE MARITAL EDUC PRES ARMS CITIES;  
RUN;
```

FREQUENCY COUNTS FOR CATEGORICAL VARIABLES

GENDER	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
1	4	66.7	4	66.7
2	2	33.3	6	100.0

RACE	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
1	3	50.0	3	50.0
2	2	33.3	5	83.3
3	1	16.7	6	100.0

Marital Status

MARITAL	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
1	2	33.3	2	33.3
2	2	33.3	4	66.7
3	1	16.7	5	83.3
4	1	16.7	6	100.0

Education Level

EDUC	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
1	2	33.3	2	33.3
2	2	33.3	4	66.7
3	1	16.7	5	83.3
4	1	16.7	6	100.0

President Doing a Good Job

PRES	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
1	1	16.7	1	16.7
2	1	16.7	2	33.3
3	1	16.7	3	50.0
4	3	50.0	6	100.0

Adding “Value Labels” (Format)

- This can be done using PROC FORMAT
- Then using the FORMAT Statement: this statement could be placed within a DATA step or as a statement in a PROC step

FORMAT variables < format > <DEFAULT=default format>

... ;

where,

- **variables**
 - Names the variable(s) you want to associate with a format. To disassociate a format from a variable, use the variable's name in a FORMAT statement with no format.
- **format**
 - Specifies the format for writing the values of the variable(s).

PROC FORMAT: *Introduction*

- The **FORMAT** procedure provides a simple method for creating your own formats and informats. The SAS System uses informats and formats to read, display, and print data stored in a SAS data set. With **PROC FORMAT**, you can design informats for reading and interpreting non-standard data, and you can design formats for displaying data in non-standard ways.

PROC FORMAT: *Syntax*

```
PROC FORMAT CNTLIN= SAS-data-set
  CNTLOUT= SAS-data-set
  FMTLIB
  LIBRARY= libref<.catalog-name>
  MAXLABLEN= n
  MAXSELEN= n
  NOREPLACE
  NOTSORTED
  PAGE;
VALUE name <(format-option-list)> range-='formatted-value-1'
  <...range-n='formatted-value-n'>;
PICTURE name <(format-option-list)> range-1='picture-1'
  <(picture-option-list)> <...range-n='picture-n'
  <(picture-option-list-n)>>;
INVALUE name <(informat-option-list)> <'>range1<'> =informatted-
  value-1 <...<'>range-n<'>=informatted-value-n>;
SELECT entry_list;
EXCLUDE entry_list;
```

VALUE name <(format-option-list)> range1 ='formatted-value-1'
< ...range-n='formatted-value-n'>;

The VALUE statement defines a format that writes variable values within certain ranges as different formatted-values. This format is given a name, and may take on several format options.

Example:

```
PROC FORMAT;
  VALUE $SEXFMT ' 1' = ' Mal e'
                ' 2' = ' Femal e' ;
  VALUE $RACE   ' 1' = ' Whi te'
                ' 2' = ' Afri can Ameri can'
                ' 3' = ' Hi spani c'
                ' 4' = ' Other' ;
  VALUE $OSCAR  ' 1' = ' Si ngl e'
                ' 2' = ' Marri ed'
                ' 3' = ' Wi dowed'
                ' 4' = ' Di vorced' ;
  VALUE $EDUC   ' 1' = ' Hi gh school or l ess'
                ' 2' = ' Two year col l ege'
                ' 3' = ' Four year col l ege'
                ' 4' = ' Post graduate degree' ;
  VALUE LIKERT  1 = ' Strongl y di sagree'
                2 = ' Di sagree'
                3 = ' Neutral '
                4 = ' Agree'
                5 = ' Strongl y agree' ;
RUN;
```

```
DATA QUEST;
```

```
INPUT ID 1-3
```

```
AGE 4-5
```

```
GENDER $ 6
```

```
RACE $ 7
```

```
MARITAL $ 8
```

```
EDUC $ 9
```

```
PRES 10
```

```
ARMS 11
```

```
CITIES 12;
```

```
LABEL MARITAL='Marital Status'
```

```
EDUC='Education Level'
```

```
PRES='President Doing a Good Job'
```

```
ARMS='Arms Budget Increase'
```

```
CITIES='Federal Aid to Cities';
```

```
FORMAT GENDER $SEXFMT.
```

```
RACE $RACE.
```

```
MARITAL $OSCAR.
```

```
EDUC $EDUC.
```

```
PRES ARMS CITIES LIKERT.;
```

FREQUENCY COUNTS FOR CATEGORICAL VARIABLES

GENDER	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
Mal e	4	66.7	4	66.7
Femal e	2	33.3	6	100.0

RACE	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
Whi te	3	50.0	3	50.0
Afri can Ameri can	2	33.3	5	83.3
Hi spani c	1	16.7	6	100.0

Mari tal Status

MARI TAL	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
Si ngl e	2	33.3	2	33.3
Marri ed	2	33.3	4	66.7
Wi dowed	1	16.7	5	83.3
Di vorced	1	16.7	6	100.0

Education Level

EDUC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High school or less	2	33.3	2	33.3
Two year college	2	33.3	4	66.7
Four year college	1	16.7	5	83.3
Post graduate degree	1	16.7	6	100.0

President Doing a Good Job

PRES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Strongly disagree	1	16.7	1	16.7
Disagree	1	16.7	2	33.3
Neutral	1	16.7	3	50.0
Agree	3	50.0	6	100.0

Recoding Data

- Suppose that we want to create a new variable that represent the age group denoted by **AGEGRP**. Such that
 - if the person is 20 years old or less then **AGEGRP=1**
 - if the person is greater than 20 years old and less than or equal to 40 then **AGEGRP=2**
 - if the person is greater than 40 years old and less than or equal to 60 then **AGEGRP=3**
 - if the person is greater than 60 years old then **AGEGRP=4**

The SAS statement is

```
IF 0 <= AGE <= 20 THEN AGEGRP=1;  
IF 20 < AGE <= 40 THEN AGEGRP=2;  
IF 40 < AGE <= 60 THEN AGEGRP=3;  
IF AGE > 60 THEN AGEGRP=4;
```

The output from PROC FREQ for AGEGRP is

Age Group					
	AGEGRP	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
<i>ff</i>					
0-20		1	16.7	1	16.7
21-40		2	33.3	3	50.0
41-60		1	16.7	4	66.7
Greater than 60		2	33.3	6	100.0

Using a Format to Recode a Variable

```
PROC FORMAT;
VALUE AGROUP      LOW-20  =' 0-20'
                  21-40   =' 21-40'
                  41-60   =' 41-60'
                  60-HIGH=' Greater than 60' ;

RUN;
.
.
.
PROC FREQ DATA=QUEST;
  TITLE ' FREQUENCY COUNTS FOR AGE USING FORMAT' ;
  TABLES AGE;
  FORMAT AGE AGROUP. ;
RUN;
```

FREQUENCY COUNTS FOR AGE USING FORMAT

AGE	Frequency	Percent	Cumul ative Frequency	Cumul ative Percent
<i>ff</i> 0-20	1	16.7	1	16.7
21-40	2	33.3	3	50.0
41-60	1	16.7	4	66.7
Greater than 60	2	33.3	6	100.0

NOTE:

The actual value of the original value is not changed . So, if you place a format statement within a DATA step and associate a format with a variable, all computations regarding that variable still involve the original value.

Two-way Frequency Table

```
PROC FREQ DATA=QUEST;
  TITLE 'TWO WAY FREQUENCY TABLE' ;
  TABLES ARMS*GENDER;
RUN;
```

TABLE OF ARMS BY GENDER

ARMS(Arms Budget Increase)
GENDER

Frequency	GENDER		Total
Percent	Male	Female	
Row Pct			
Col Pct	Male	Female	Total
=====			
Disagree	2	1	3
	33.33	16.67	50.00
	66.67	33.33	
	50.00	50.00	
=====			
Neutral	1	1	2
	16.67	16.67	33.33
	50.00	50.00	
	25.00	50.00	
=====			
Agree	1	0	1
	16.67	0.00	16.67
	100.00	0.00	
	25.00	0.00	
=====			
Total	4	2	6
	66.67	33.33	100.00

```

PROC FREQ DATA=QUEST;
  TITLE 'Multiple TWO WAY FREQUENCY TABLE';
  TABLES (ARMS PRES)*GENDER;
RUN;

```

TWO WAY FREQUENCY TABLE

TABLE OF ARMS BY GENDER

ARMS(Arms Budget Increase)
GENDER

	Male	Female	Total
Di disagree	2	1	3
	33.33	16.67	50.00
	66.67	33.33	
	50.00	50.00	
Neutral	1	1	2
	16.67	16.67	33.33
	50.00	50.00	
	25.00	50.00	
Agree	1	0	1
	16.67	0.00	16.67
	100.00	0.00	
	25.00	0.00	
Total	4	2	6
	66.67	33.33	100.00

TWO WAY FREQUENCY TABLE

TABLE OF PRES BY GENDER

PRES(President Doing a Good Job)
GENDER

	Male	Female	Total
Strongly disagree	1	0	1
	16.67	0.00	16.67
	100.00	0.00	
	25.00	0.00	
Disagree	1	0	1
	16.67	0.00	16.67
	100.00	0.00	
	25.00	0.00	
Neutral	0	1	1
	0.00	16.67	16.67
	0.00	100.00	
	0.00	50.00	
Agree	2	1	3
	33.33	16.67	50.00
	66.67	33.33	
	50.00	50.00	
Total	4	2	6
	66.67	33.33	100.00

Computing Chi-Square from Frequency Count

```
DATA CHI SQ;  
  INPUT GROUP $ OUTCOME $ COUNT;  
CARDS;  
CONTROL DEAD 20  
CONTROL ALIVE 80  
DRUG DEAD 10  
DRUG ALIVE 90  
;  
PROC FREQ DATA=CHI SQ;  
  TITLE ' TWO WAY FREQUENCY TABLE' ;  
  TABLES GROUP*OUTCOME / CHI SQ;  
  WEIGHT COUNT;  
RUN;
```

TWO WAY FREQUENCY TABLE

TABLE OF GROUP BY OUTCOME

GROUP	OUTCOME		Total
Frequency,			
Percent ,			
Row Pct ,			
Col Pct ,	ALIVE	DEAD	
CONTROL	80	20	100
	40.00	10.00	50.00
	80.00	20.00	
	47.06	66.67	
DRUG	90	10	100
	45.00	5.00	50.00
	90.00	10.00	
	52.94	33.33	
Total	170	30	200
	85.00	15.00	100.00

STATISTICS FOR TABLE OF GROUP BY OUTCOME

Statistic	DF	Value	Prob
Chi-Square	1	3.922	0.048
Likelihood Ratio Chi-Square	1	3.987	0.046
Continuity Adj. Chi-Square	1	3.176	0.075
Mantel-Haenszel Chi-Square	1	3.902	0.048
Fisher's Exact Test (Left)			0.037
(Right)			0.986
(2-Tail)			0.073
Phi Coefficient		-0.140	
Contingency Coefficient		0.139	
Cramer's V		-0.140	

Sample Size = 200

Let us look at the columns of the table one at a time and pay attention at the column percentage. It shows that 47.06% of the Alive were in the control group while 52.94 % were in the Drug group. Next, 66.67% of the Dead were in the Control group while 33.33% were in the Drug Group. We conclude that there is a relationship between Group and the outcome. (Discuss more)

The chi-square tests of no association computed by PROC FREQ test the null hypothesis that there is no association between the row variables and the column variables. Or that the row and the column variables are independent.

So the Chi-square statistic

- Measure the strength of the evidence that an association exists.
- Does not measure the strength of the association
- Depend on the sample size

- If the chi-square table has 1 degree of freedom and the expected cell value is less than 5, a 2-tail Fisher exact test could be used. A correction for continuity called “Yates” could be used too.
- In the case where the degrees of freedom are greater than 1, it is desirable that no more than 20% of the cells have expected value less than 5. You need to combine cells.

McNemar's Test for Paired Data

- Example

- Determine the effect of an anti-cigarette advertisement on people's attitude towards smoking.
- Data
 - 30 people who had negative attitude toward cigarette and stayed negative after the advertisement.
 - 10 people who had negative attitude toward cigarette and had positive attitude after the advertisement.
 - 45 people who had positive attitude toward cigarette and changed to negative after the advertisement.
 - 30 people who had positive attitude toward cigarette and stayed positive after the advertisement.

```
DATA CI GAR;  
  INPUT BEFORE $ AFTER $ COUNT;  
CARDS;  
N N 30  
N P 10  
P N 45  
P P 15  
;  
PROC FREQ DATA=CI GAR;  
  TITLE ' McNEMARS TEST' ;  
  TABLES BEFORE*AFTE R / AGREE;  
  WEI GHT COUNT;  
RUN;
```

TABLE OF BEFORE BY AFTER

	BEFORE	AFTER	Total
Frequency,			
Percent ,			
Row Pct ,			
Col Pct , N		P	
~~~~~			
N	30	10	40
	30.00	10.00	40.00
	75.00	25.00	
	40.00	40.00	
~~~~~			
P	45	15	60
	45.00	15.00	60.00
	75.00	25.00	
	60.00	60.00	
~~~~~			
Total	75	25	100
	75.00	25.00	100.00

STATISTICS FOR TABLE OF BEFORE BY AFTER

McNemar's Test

Statistic = 22.273      DF = 1      Prob = 0.001

Simple Kappa Coefficient

Kappa = -0.000      ASE = 0.077      95% Confidence Bounds  
 -0.151      0.151

Sample Size = 100

The McNemar's chi square statistic is 22.273 with p-value of 0.001. So the anti-cigarette advertisement was effective in changing people's attitude towards smoking.

# Odds Ratio

## Example:

Do you think that people with brain tumor are more likely to have been exposed to benzene than people without brain Tumor. A case-Control design experiment is used to answer this question. The outcome is given in the following table:

The odds of a Case being exposed to benzene is 50/100.

The odds of a Control being exposed to benzene is 20/130.

Therefore  $.5/.155=3.25$  is the odds Ratio. So case is 3.25 more likely to be exposed to benzene than the control group.

## Outcome

## Exposure

	Case	Control	Total
Yes	50	20	70
No	100	130	230
Total	150	150	300

To check this we can use CHISQ and CMH (Cochran-Mantel-Haenszel) options

```
DATA ODDS ;  
  INPUT OUTCOME $ EXPOSURE $ COUNT;  
CARDS;  
CASE 1-YES 50  
CASE 2-NO 100  
CONTROL 1-YES 20  
CONTROL 2-NO 130  
;  
  
PROC FREQ DATA=ODDS;  
  TITLE 'ODDS RATIO' ;  
  TABLES EXPOSURE*OUTCOME / CHI SQ CMH;  
  WEIGHT COUNT;  
RUN;
```

# ODDS RATIO

## SUMMARY STATISTICS FOR EXPOSURE BY OUTCOME

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	16.714	0.001
2	Row Mean Scores Differ	1	16.714	0.001
3	General Association	1	16.714	0.001

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control (Odds Ratio)	Mantel-Haenszel	3.250	1.847	5.719
	Logit	3.250	1.819	5.807
Cohort (Col 1 Risk)	Mantel-Haenszel	1.643	1.295	2.084
	Logit	1.643	1.333	2.025
Cohort (Col 2 Risk)	Mantel-Haenszel	0.505	0.364	0.701
	Logit	0.505	0.343	0.745

The confidence bounds for the M-H estimates are test-based.

The odds ratio is 3.25 as was shown earlier and the 95% confidence interval is [1.847,5.719] which does not include 1 so the odds ratio 3.25 is significant at 0.05 alpha level.

# Chi-square Test For Trend

		<b>Group</b>			
		A	B	C	D
<b>Test Result</b>	Fail	10	15	14	25
	Pass	90	85	86	75
		100	100	100	100

Notice that the proportions of “Fail” in group A through Group D is increasing except form group B to C. To test if there is a linear trend in proportions, we can use CHISQ option and look at the statistics labeled “Mantel Haenszel chi-square”.

```
DATA TREND;  
  INPUT RESULT $ GROUP $ COUNT;  
CARDS;  
FAIL A 10  
FAIL B 15  
FAIL C 14  
FAIL D 25  
PASS A 90  
PASS B 85  
PASS C 86  
PASS D 75  
;  
  
PROC FREQ DATA=TREND;  
  TITLE 'TREND TEST' ;  
  TABLES RESULT*GROUP / CHISQ;  
  WEIGHT COUNT;  
RUN
```

# TREND TEST

## TABLE OF RESULT BY GROUP

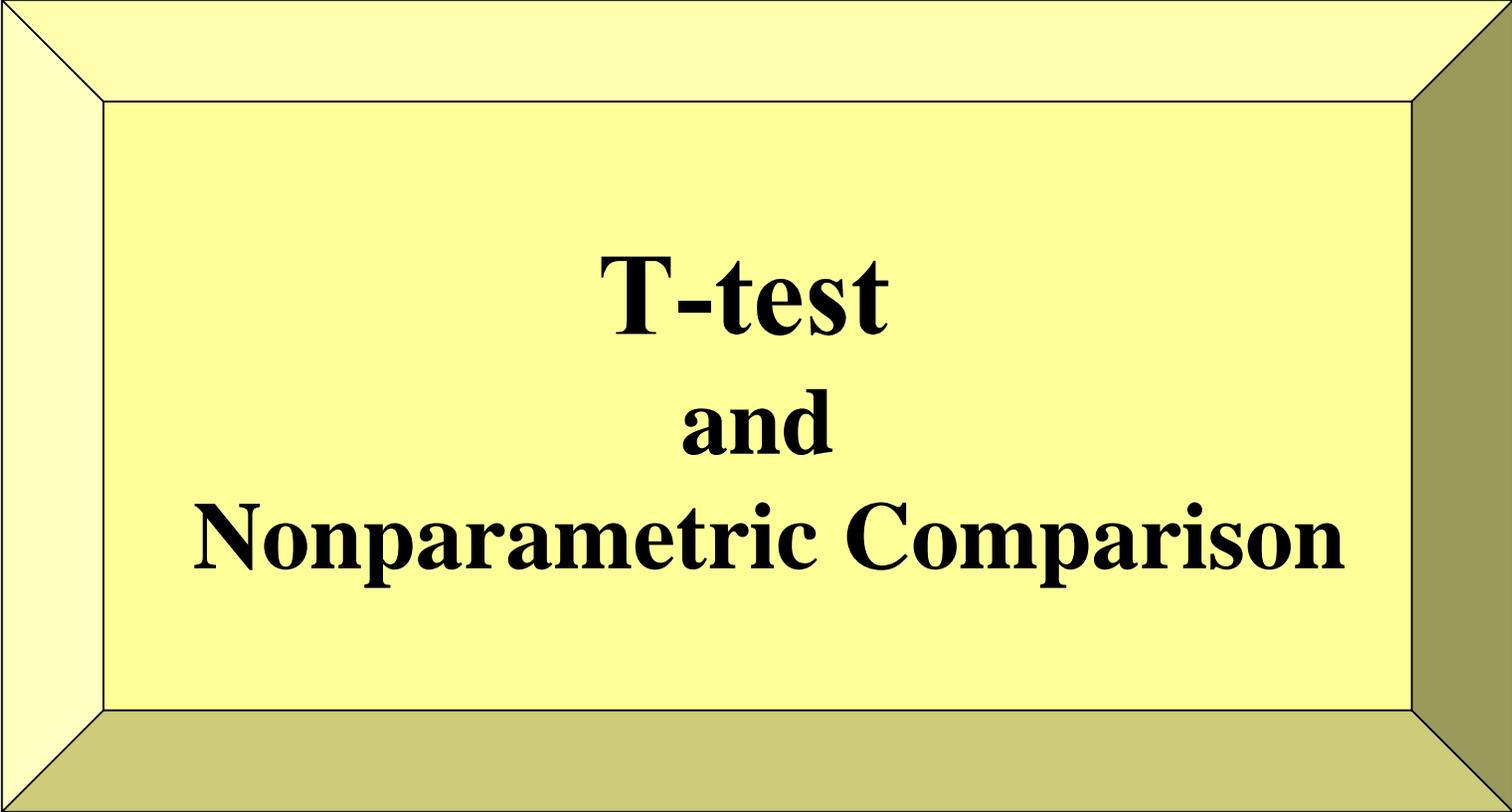
RESULT	GROUP				Total
Frequency,	A	B	C	D	
Percent ,					
Row Pct ,					
Col Pct ,					
FAIL	10	15	14	25	64
	2.50	3.75	3.50	6.25	16.00
	15.63	23.44	21.88	39.06	
	10.00	15.00	14.00	25.00	
PASS	90	85	86	75	336
	22.50	21.25	21.50	18.75	84.00
	26.79	25.30	25.60	22.32	
	90.00	85.00	86.00	75.00	
Total	100	100	100	100	400
	25.00	25.00	25.00	25.00	100.00

## STATISTICS FOR TABLE OF RESULT BY GROUP

Statistic	DF	Value	Prob
Chi-Square	3	9.077	0.028
Likelihood Ratio Chi-Square	3	8.718	0.033
Mantel-Haenszel Chi-Square	1	7.184	0.007
Phi Coefficient		0.151	
Contingency Coefficient		0.149	
Cramer's V		0.151	

Sample Size = 400

From the output the M-H- chi square test for trend is 7.184 with a p-value of 0.007 so there is a significant linear trend.



**T-test  
and  
Nonparametric Comparison**



# T-tests and Nonparametric Comparison

- T-test: Testing Differences between Two Means
- Random Assignment of Subjects
- Two Independent Sample: Distribution Free Tests
- Paired T-tests (Related Samples)

# PROC TTEST: *Introduction*

- The t test tests the hypothesis that the true means of two groups of observations are the same. This analysis can be considered a special case of a one-way analysis of variance with two levels of classification.
- PROC TTEST computes the t statistic based on the assumption that the variances of the two groups are equal, and it computes an approximate t based on the assumption that the variances are unequal. For each t, the degrees of freedom and probability levels are given.
- Note that the underlying assumption of the t test computed by the TTEST procedure is that the variables are normally and independently distributed within each group.

# PROC TTEST: *Syntax*

```
PROC TTEST DATA= SAS-data-set COCHRAN;  
  CLASS variable;                               /* required */  
  VAR variables;  
  BY variables;
```

- A CLASS statement specifying the name of the grouping variable must accompany the PROC TTEST statement. The grouping variable must have two, and only two, levels. PROC TTEST divides the observations into the two groups for the t test using the levels of this variable.
- The VAR statement specifies the names of the dependent variables whose means are to be compared.
  - If the VAR statement is omitted, all numeric variables in the input data set (except a numeric variable appearing in the CLASS statement) are included in the analysis.
- The COCHRAN option requests the Cochran and Cox approximation of the probability level of the approximate t statistic for the unequal variance situation.

## Example:

Patients are randomly assigned to a control or a treatment group (where a drug is administered). Their response time to stimulate is measured. Do the treatment scores comes from a population whose mean is different from the mean of the population from which the control scores were drawn?

<b>Control</b>	<b>Treatment</b>
80	100
93	103
83	104
89	99
98	102

# SAS Statements

```
DATA RESPONSE;  
  INPUT GROUP $ TIME;  
CARDS;  
C 80  
C 93  
C 83  
C 89  
C 98  
T 100  
T 103  
T 104  
T 99  
T 102  
;  
PROC TTEST DATA=RESPONSE;  
  TITLE 'T-test Example';  
  CLASS GROUP;  
  VAR TIME;  
RUN;
```

## T-test Example

TTEST PROCEDURE

Variable: TIME

GROUP	N	Mean	Std Dev	Std Error	Minimum	Maximum
C	5	88.60000000	7.30068490	3.26496554	80.00000000	98.00000000
T	5	101.60000000	2.07364414	0.92736185	99.00000000	104.00000000

Variances	T	DF	Prob> T
Unequal	-3.8302	4.6	0.0145
Equal	-3.8302	8.0	0.0050

For H0: Variances are equal,  $F' = 12.40$      $DF = (4, 4)$      $Prob>F' = 0.0318$

First we look at the hypothesis that the two variances are equal to determine which t-value to use. Here the F ratio (the larger variance divided by the smaller variance) is 12.4 with  $Prob>F = 0.0318$ . So at 0.05 level we reject the null hypothesis that the two variances are equal. Hence, we should choose the t-test for the unequal variances which equal to -3.8302 with  $Prob > |T| = 0.0145$ . The conclusion is we do reject the null hypothesis that the mean for the control group equal the mean for the treatment group.

# Paired T-Test

- It is used in the experimental situations where each subject receives both treatments. In this case the t-test we used before can't be used since the groups are no longer independent. Instead we compute the difference between the control and the treatment for each subject. And then test whether the mean of the difference is equal, less or greater than zero.

## Example:

In an experiment the response time for each patient was measured in the absence of a drug (control value) and then after having received the drug (treatment value). Would you expect the response time after taking the drug to be different?

Subject	Control	Treatment
1	90	95
2	87	92
3	100	104
4	80	89
5	95	101
6	90	105

```

DATA PAIRED;
  INPUT CTIME TTIME;
  DIFF = TTIME - CTIME;
CARDS;
90 95
87 92
100 104
80 89
95 101
90 105
;
PROC MEANS DATA=PAIRED N MEAN STDERR T PRT;
  TITLE 'PAIRED T-TEST EXAMPLE' ;
  VAR DIFF;
RUN;

```

### PAIRED T-TEST EXAMPLE

Analysis Variable : DIFF

N	Mean	Std Error	T	Prob> T
6	7.3333333	1.6865481	4.3481318	0.0074

Looking at the results we see that the T-value = 4.348 with  $\text{Prob}>|T| = 0.0074$ . So the hypothesis that the mean value for the difference is rejected. Also looking at the mean value of the difference we could see that it has a positive value so we conclude that the response time are longer under the drug treatment.

# Random Assignments of Subjects

To use SAS to assign subjects to either a treatment or control group (This method will work if you have more than one group.) we would use the random number function **RANUNI(seed)** which generates a pseudorandom number in the interval 0 to 1.

- A zero seed specifies that the function use the time clock to generate a random seed to initiate the random number sequence.
- If you use a zero seed, you will obtain a different series of random numbers every time the program is run unless you run the program at the same time everyday.
- You could supply the seed as any number you wish. In this case you will generate the same series of random numbers.
- Then to split the group in half (or as many as you want) you could use PROC RANK in SAS.

# PROC RANK: *Introduction*

The RANK procedure ranks values from smallest to largest, assigning the rank 1 to the smallest number, 2 to the next largest, and so on up to rank n, the number of nonmissing observations. Tied values are given averaged ranks. Several options are available to request other ranking and tie-handling rules.

# PROC RANK: *Syntax*

```
PROC RANK DATA= SAS-data-set  
    TIES= MEAN|HIGH|LOW  
    DESCENDING  
    OUT= SAS-data-set  
    FRACTION  
    NPLUS1  
    PERCENT  
    GROUPS= n  
    NORMAL= BLOM|TUKEY|VW  
    SAVAGE;  
VAR variable-list;  
RANKS new-variable-list;  
BY variable-list;
```

# Example:

- Assign the following people randomly into two groups using SAS.

CODY, GREGORY, SAM, SMITH,  
HELM, ANDY

```

PROC FORMAT;
  VALUE GRPFRM  0 = 'CONTROL'
                1 = 'TREATMENT' ;
RUN;

DATA RANDOM;
  INPUT SUBJ NAME $20. ;
  GROUP=RANUNI (0);
CARDS;
1 CODY
2 GREGORY
3 SAM
4 SMI TH
5 HELM
6 ANDY
;
PROC RANK DATA=RANDOM GROUP=2 OUT=SPLIT;
  VAR GROUP;
RUN;

PROC SORT DATA=SPLIT;
  BY NAME;
RUN;

PROC PRINT DATA=SPLIT;
  TITLE 'SUBJECT GROUP ASSIGNMENT' ;
  ID NAME;
  VAR SUBJ GROUP;
  FORMAT GROUP GRPFRM. ;
RUN;

```

## SUBJECT GROUP ASSIGNMENT

NAME	SUBJ	GROUP
ANDY	6	CONTROL
CODY	1	TREATMENT
GREGORY	2	CONTROL
HELM	5	TREATMENT
SAM	3	CONTROL
SMI TH	4	TREATMENT

# Two Independent Samples

## *Distribution Free Tests*

A t-test is not appropriate

- If the data does not come from a normal distribution
- If the data values represent ordered categories (I.e from ordinal scaled data)
- If we have a small sample size

The solution is to use a nonparametric test ( a test that does not assume a distribution for the data).

## Example:

Consider the following experiment. Two groups A and B. Group B has been treated with a drug to prevent tumor formation. Both groups are exposed to a chemical that encourages tumor growth. The masses (in grams) of tumors in groups A and B are

**A: 3.1 2.2 1.7 2.7 2.5**

**B: 0.0 0.0 1.0 2.3**

**Are there any difference in tumor mass between group A and B?**

Since the data does not come from a normal distribution and the sample size is small then we need to use the nonparametric test.

The Wilcoxon test first puts all the data into an increasing order retaining the group identity as follows

MASS	0.0	0.0	1.0	1.7	2.2	2.3	2.5	2.7	3.1
GROUP	B	B	B	A	A	B	A	A	A
RANK	1.5	1.5	3	4	5	6	7	8	9

Then the sum of the ranks for A and B are compared. We have

$$\text{SUM RANKS A} = 4+5+7+8+9 = 33$$

$$\text{SUM RANKS B} = 1.5+1.5+3+6 = 12$$

If there is smaller tumors in group B, we would expect B to be at the lower end of the rank ordering and therefore have smaller sum of ranks than the A's.

So the question now. Is the sum of ranks for group A is sufficiently different than the sum of ranks for group B so the probability of the difference by chance alone is small?

# PROC NPAR1WAY: *Introduction*

The NPAR1WAY procedure performs analysis of variance on ranks and computes statistics based on the empirical distribution function and certain rank scores of a response variable across a one-way classification.

# PROC NPAR1WAY: *Syntax*

```
PROC NPAR1WAY DATA= SAS-data-set
    MISSING
    ANOVA
    EDF
    MEDIAN
    NOPRINT
    SAVAGE
    VW
    WILCOXON;
CLASS variable;          /* required */
EXACT <keywords>;
OUTPUT <OUT=SAS-data-set> <options>;
VAR variable-list;
BY variable-list;
```

- The **MISSING** option interprets missing class values as non missing and includes them in calculations as valid class levels.
- The **ANOVA** option performs a standard analysis of variance in addition to the nonparametric ANOVA performed on the ranks.
- The **WILCOXON** option requests an analysis of the ranks of the data or the Wilcoxon Scores.
  - For two levels, this is the same as a Wilcoxon rank-sum test. For any number of levels, this is a Kruskal-Wallis test.
- The **MEDIAN** option requests an analysis of the median scores. The median score is 1 for points above the median, 0 otherwise.
  - For two samples, this produces a median test. For more than two samples, this is the Brown-Mood test.
- **The EXACT statement specifies analyses for which exact p-values are needed, in addition to the asymptotic p-values already given by PROC NPAR1WAY. This is needed when the sample size is small (<10) in each group.**

NONPARAMETRIC TEST TO COMPARE TUMOR MASSES

N P A R 1 W A Y P R O C E D U R E

Wilcoxon Scores (Rank Sums) for Variable MASS  
Classified by Variable GROUP

GROUP	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	5	33.0	25.0	4.06543697	6.60000000
B	4	12.0	20.0	4.06543697	3.00000000

Average Scores Were Used for Ties

Wilcoxon 2-Sample Test      S = 12.0000

Exact P-Values

(One-sided) Prob  $\leq$  S = 0.0317

(Two-sided) Prob  $\geq$  |S - Mean| = 0.0635

Normal Approximation (with Continuity Correction of .5)

Z = -1.84482      Prob  $>$  |Z| = 0.0651

T-Test Approx. Significance = 0.1023

Kruskal-Wallis Test (Chi-Square Approximation)

CHISQ = 3.8723      DF = 1      Prob  $>$  CHISQ = 0.0491

Here we are testing the null hypothesis that the two groups have the same rank sum against the alternative that they are different. Since the sample size is small we use the **Exact two-tailed** p-value which equal 0.0635. So at 0.05 we do not reject the null hypothesis that the two are equal.

But if the alternative hypothesis is that the rank sum for group A is greater than the rank sum for group B we use the **Exact one-tailed** p-value which equal to 0.0317. So at 0.05 we do reject the null and conclude that the rank sum for group A is larger than the rank sum for group B.

## Example:

Consider the following experiment. Two drugs 1 and 2 is given to patients. **Are there any difference in heart rate mass between group 1 and 2?**

Since the data does not come from a normal distribution then we need to use the nonparametric test.

```
PROC NPAR1WAY DATA=DRUGDATA;  
CLASS DRUG;  
VAR CHANG_BP;  
RUN;
```

N P A R 1 W A Y P R O C E D U R E

Wilcoxon Scores (Rank Sums) for Variable CHANG_BP  
Classified by Variable DRUG

DRUG	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	18	324.500000	333.0	31.5540353	18.0277778
2	18	341.500000	333.0	31.5540353	18.9722222

Average Scores Were Used for Ties

Wilcoxon 2-Sample Test (Normal Approximation)  
(with Continuity Correction of .5)

S = 324.500                      Z = -.253533                      Prob > |Z| = 0.7999

T-Test Approx. Significance = 0.8013

Kruskal-Wallis Test (Chi-Square Approximation)

CHISQ = 0.07257                      DF = 1                      Prob > CHISQ = 0.7876

From the Wilcoxon 2-sample test (normal approximation), the p-value is 0.7999. Thus, we do not reject the null hypothesis at 0.05 level and, assuming that the distributions have the same shape, we conclude that the treatment means are different.

N P A R 1 W A Y P R O C E D U R E

Medi an Scores (Number of Poi nts Above Medi an)  
for Vari abl e CHANG_BP  
Cl assi fi ed by Vari abl e DRUG

DRUG	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	18	8.0	9.0	1.52127766	0.444444444
2	18	10.0	9.0	1.52127766	0.555555556

Average Scores Were Used for Ties

Medi an 2-Sampl e Test (Normal Approxi mati on)

S = 8.00000                      Z = -.657342                      Prob > |Z| = 0.5110

Medi an 1-Way Anal ysi s (Chi -Square Approxi mati on)

CHI SQ = 0.43210                      DF = 1                      Prob > CHI SQ = 0.5110

From the median 2-sample test, the p-value is 0.5110. Thus, we do not reject the null hypothesis at 0.05 level and, assuming that the distributions have the same shape, we conclude that the treatment means are different.

N P A R 1 W A Y P R O C E D U R E

Van der Waerden Scores (Normal) for Variable CHANG_BP  
Classified by Variable DRUG

DRUG	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	18	-.250899181	0.0	2.77649479	-.013938843
2	18	0.250899181	0.0	2.77649479	0.013938843

Average Scores Were Used for Ties

Van der Waerden 2-Sample Test (Normal Approximation)

S = -.250899                      Z = -.090365                      Prob > |Z| = 0.9280

Van der Waerden 1-Way Analysis (Chi-Square Approximation)

CHI SQ = 0.00817                      DF = 1                      Prob > CHI SQ = 0.9280

From the Van der Waerden 2- sample test, the p-value is 0.9280. Thus, we do not reject the null hypothesis at 0.05 level and, assuming that the distributions have the same shape, we conclude that the treatment means are different.

N P A R 1 W A Y P R O C E D U R E

Savage Scores (Exponential) for Variable CHANG_BP  
Classified by Variable DRUG

DRUG	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	18	0.711274469	0.0	2.84968075	0.039515248
2	18	-.711274469	0.0	2.84968075	-.039515248

Average Scores Were Used for Ties

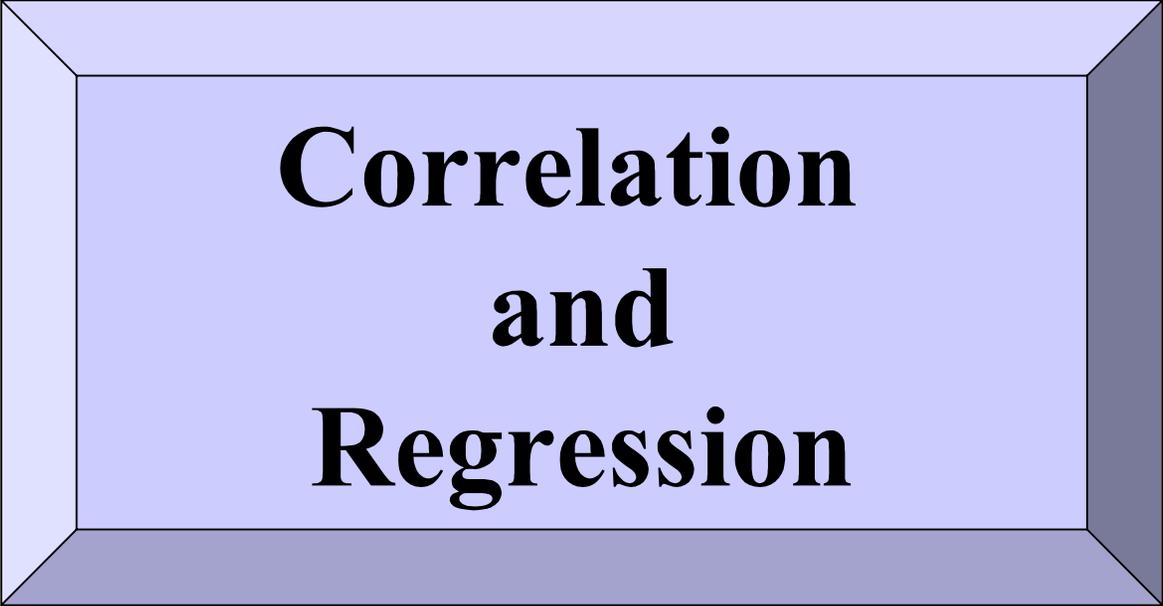
Savage 2-Sample Test (Normal Approximation)

S = 0.711274                      Z = 0.249598                      Prob > |Z| = 0.8029

Savage 1-Way Analysis (Chi-Square Approximation)

CHI SQ = 0.06230                      DF = 1                      Prob > CHI SQ = 0.8029

From the Van Savage 2- sample test, the p-value is 0.8029.  
Thus, we do not reject the null hypothesis at 0.05 level and, assuming that the distributions have the same shape, we conclude that the treatment means are different.



**Correlation  
and  
Regression**

# Contents:

- Correlation
- Significant of Correlation Coefficient
- How to interpret a Correlation Coefficient
- Partial Correlation
- Linear regression
- Partitioning the Total sum of squares
- Plotting the points on a regression line

# Correlation

- The correlation coefficient is a statistics that measures the strength of a linear relationship between two variables
- This number ranges between -1 and 1
  - A small or a zero value means that the two values are unrelated.
  - A positive correlation means that when values of one variable increase, values of the other variable tend to increase also.
  - A negative correlation means that when values of one variable increase, values of the other variable tend to decrease.

# Pearson Correlation

- Pearson Correlation is a measure of association between two variables. It can be used only
  - when both variables are assessed on either an interval or ratio level of measurement.
  - And that the observations have been drawn from normally distributed populations

# Spearman correlation

- Spearman Rank-Correlation Coefficient: is a measure of association between two variables. It can be used in the following situations
  - when both variables are assessed on an ordinal level measurement.
  - When one of the variables is an ordinal level variable and the other is an interval/ratio level variable.
  - Also used with interval/ratio variables if markedly non-normal since the Spearman coefficient is a distribution free test.

# PROC CORR: *Syntax*

PROC CORR <option-list>;

VAR variable-list;

WITH variable-list;

PARTIAL variable-list;

WEIGHT variable;

FREQ variable;

BY variable-list;

# <option-list>

ALPHA	BEST= number	COV
CSSCP	DATA= SAS-data-set	HOEFFDING
KENDALL	NOCORR	NOMISS
NOPRINT	NOPROB	NOSIMPLE
OUTH= SAS-data-set	OUTK= SAS-data-set	OUTP= SAS-data-set
OUTS= SAS-data-set	PEARSON	RANK
SINGULAR= p	SPEARMAN	SSCP
VARDEF= DF N WDF WEIGHT WGT		

These options can appear in the PROC CORR statement. If no options are specified, CORR calculates Pearson product-moment correlations and significance probabilities, printing them in a rectangular table along with univariate statistics.

- ALPHA

- The ALPHA option requests that Cronbach's coefficient alpha be calculated and printed. Separate coefficients are generated using the raw variables and the standardized variables (each variable is scaled to have unit variance). Only the variables in the VAR statement are used.
- For each variable, the correlation between the variable and the total of the remaining variables is calculated. Also, CORR calculates Cronbach's coefficient alpha using only the remaining variables.
- When a PARTIAL statement is also used, the coefficient alpha is calculated by using the variables after partialling.
- Requesting this option also activates the PEARSON option.

- **BEST= number**
  - The **BEST=** option prints the **n** correlations for each variable with the largest absolute values; the coefficients are printed in descending order.
- **COV**
  - The **COV** option requests that covariances be printed.
  - If **COV** and **OUTP=** are both specified, the output data set also contains the covariance matrix.
  - Specifying the **COV** option activates the **PEARSON** option.
  - When a **PARTIAL** statement is used, the partial covariance matrix is printed.

- **CSSCP**
  - The CSSCP option requests that the corrected sums of squares and crossproducts be printed.
  - If CSSCP and OUTP= are both specified, the output data set also contains the CSSCP matrix.
  - Specifying the CSSCP option activates the PEARSON option.
  - When a PARTIAL statement is used, both unpartialled and partialled CSSCP matrices are printed. When the OUTP= option is used, the partialled, but not the unpartialled, CSSCP matrix is placed in the output data set.
- **DATA= SAS-data-set**
  - The DATA= option names the SAS data set used by PROC CORR. If the DATA= option is omitted, the most recently created SAS data set is used .

- **HOEFFDING**

- The HOEFFDING option requests that the Hoeffding's D statistic be calculated and printed.
- This option is not valid if either a WEIGHT or a PARTIAL statement is included.
- Note: The statistic calculated by CORR is 30 times the usual definition. This scales the statistic to a range between -0.5 and 1, with only large positive values indicating dependence.

- **KENDALL**

- The KENDALL option requests that the Kendall's tau-b coefficients be calculated and printed.
- Kendall's tau-b is based on the number of concordant and discordant pairs of observations and uses a correction for tied pairs (pairs of observations that have equal values of X and equal values of Y).
- This option is not valid if a WEIGHT statement is included.
- Range:  $-1 \leq \text{tau-b} \leq 1$

- **NOCORR**
  - The NOCORR option specifies that Pearson correlations not be calculated or printed.
  - If both NOCORR and OUTP= are specified, the output data set does not contain correlations, but PROC CORR still makes the data set TYPE=CORR. To change the data set type to COV, CSSCP, or SSCP, use the TYPE= data set option:
    - `proc corr nocorr cov outp=b(type=cov);`
- **NOMISS**
  - The NOMISS option drops from analysis observations with missing values.
- **NOPRINT**
  - The NOPRINT option suppresses all printed output. Use this option when you want to create only an output data set containing observations.

- **NOPROB**
  - The NOPROB option suppresses the printing of significance probabilities associated with the correlations.
- **NOSIMPLE**
  - The NOSIMPLE option suppresses the printing of simple descriptive statistics for each variable.
  - However, when an output data set is requested, the simple descriptive statistics for each variable specified in the VAR statement are output to the specified data set.
  - When a PARTIAL statement is used and the PEARSON option is used, then for each variable in the VAR or WITH statements, CORR also prints the simple statistics (variance and standard deviation) after partialling.

- **OUTH= SAS-data-set**
  - The OUTH= option requests that CORR create a new SAS data set containing the Hoeffding statistics.
  - Requesting this data set activates the HOEFFDING option.
- **OUTK= SAS-data-set**
  - The OUTK= option requests that CORR create a new SAS data set containing the Kendall statistics.
  - Requesting this data set activates the KENDALL option.
- **OUTP= SAS-data-set**
  - The OUTP= option requests that CORR create a new SAS data set containing the Pearson statistics.
  - Requesting this data set activates the PEARSON option.
- **OUTS= SAS-data-set**
  - The OUTS= option requests that CORR create a new SAS data set containing the Spearman statistics.
  - Requesting this data set activates the SPEARMAN option.

- PEARSON
  - The PEARSON option requests the usual Pearson product-moment correlations.
  - Since these are the default statistics, specify this option only if you are also requesting Spearman, Kendall, or Hoeffding statistics.
- RANK
  - The RANK option prints correlation coefficients for each variable in order of highest to lowest in absolute value. When RANK is omitted, the correlations are printed in a rectangular table defined by variable names at the top and side.
- SINGULAR= p
  - The SINGULAR option specifies the criterion for determining the singularity of a variable when a PARTIAL statement is used, where  $0 < p < 1$ .
  - Default: 1E-8

- SPEARMAN
  - The SPEARMAN option requests that Spearman coefficients be calculated and printed. These are the correlations of the ranks of the variables.
  - This option is not valid if a WEIGHT statement is included.
  - Range:  $-1 \leq r \leq 1$
- VARDEF= DF | N | WEIGHT | WGT | WDF
  - The VARDEF= option specifies the divisor to be used in the calculation of variances and covariances. Possible values are:
    - DF uses the degrees of freedom.
    - N uses the number of observations.
    - WEIGHT | WGT uses the sum of the weights.
    - WDF uses the sum of the weights minus one.
    - Default: DF
    - Note: Partial variances and covariances can also be calculated.

- SSCP
  - The SSCP option requests that the sums of squares and crossproducts be printed.
  - If both SSCP and OUTP= are specified, the output data set also contains the SSCP matrix.
  - Specifying the SSCP option activates the PEARSON option.
  - When a PARTIAL statement is used, the unpartialled SSCP matrix is printed. If OUTP= is also specified, the output data set does not include the partial SSCP matrix.

- VAR variable-list;
  - The VAR statement lists the names of the variables for which correlations are to be evaluated. If omitted, CORR calculates correlations between all the numeric variables in the input data set that are not listed in other statements.
  - For example, the following statements produce correlation coefficients between three pairs of variables: A and B, A and C, and B and C:

```
proc corr;
```

```
var a b c;
```

- **WITH variable-list;**
  - The **WITH** statement enables correlations for specific combinations of variables to be requested. List the variables to appear on the top of the printed correlation matrix in the **VAR** statement, and list variables to appear on the side of the correlation matrix in the **WITH** statement.

- PARTIAL variable-list;
  - The PARTIAL option specifies variables to be partialled in Pearson's partial correlation, Spearman's partial rank-order correlation, or Kendall's partial tau-b.
  - The HOEFFDING option is not valid with a PARTIAL statement.
  - Note: Specifying the PARTIAL statement also activates the NOMISS option.

- **WEIGHT** variable;
  - The **WEIGHT** option specifies the name of the weighting variable to be used in computing weighted product-moment correlation coefficients.
  - If the value of the **WEIGHT** variable is missing or less than zero, then a value of zero for the weight is used.
  - The **SPEARMAN**, **KENDALL**, and **HOEFFDING** options are not valid with a **WEIGHT** statement.

- **FREQ variable;**
  - When a FREQ statement is specified, each observation in the input data set is assumed to represent 'n' observations, where 'n' is the value of the FREQ variable for the observation. The total number of observations is considered equal to the sum of the FREQ variable.
  - If the value of the FREQ variable is missing or less than one, the observation is not used in the analysis. If the value is not an integer, the value is truncated to an integer.

# Example:

For the following data, find the Pearson Correlation. This data is saved in a:\HTWT.txt

<b>Gender</b>	<b>Height</b>	<b>Weight</b>	<b>Age</b>
M	68	155	23
F	61	99	20
F	63	115	21
M	70	205	45
M	69	170	.
F	65	125	30
M	72	220	48

# Input the data into SAS

```
DATA CORR_EX1;  
  INFILE 'a:\HTWT.txt';  
  INPUT SUBJECT GENDER $  
  HEIGHT WEIGHT AGE;  
RUN;
```

## Default output when using PROC CORR

```
PROC CORR DATA=CORR_EX1;  
  TITLE 'EXAMPLE OF CORRELATION MATRIX' ;  
RUN;
```

EXAMPLE OF CORRELATION MATRIX

Correlation Analysis

4 'VAR' Variables: SUBJECT HEIGHT WEIGHT AGE

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SUBJECT	7	4.000000	2.160247	28.000000	1.000000	7.000000
HEIGHT	7	66.714286	3.903600	467.000000	61.000000	72.000000
WEIGHT	7	155.571429	45.796132	1089.000000	99.000000	220.000000
AGE	6	31.166667	12.416387	187.000000	20.000000	48.000000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

	SUBJECT	HEIGHT	WEIGHT	AGE
SUBJECT	1.00000 0.0 7	0.49411 0.2597 7	0.50877 0.2436 7	0.73124 0.0986 6
HEIGHT	0.49411 0.2597 7	1.00000 0.0 7	0.97625 0.0002 7	0.86614 0.0257 6
WEIGHT	0.50877 0.2436 7	0.97625 0.0002 7	1.00000 0.0 7	0.92496 0.0082 6
AGE	0.73124 0.0986 6	0.86614 0.0257 6	0.92496 0.0082 6	1.00000 0.0 6

## Using VAR statement

```
PROC CORR DATA=CORR_EX1;  
  TITLE 'CORRELATION MATRIX FOR HEIGHT  
        AND WEIGHT';  
  VAR HEIGHT WEIGHT;  
RUN;
```

CORRELATION MATRIX FOR HEIGHT AND WEIGHT

Correlation Analysis

2 'VAR' Variables: HEIGHT WEIGHT

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	7	66.71429	3.90360	467.00000	61.00000	72.00000
WEIGHT	7	155.57143	45.79613	1089	99.00000	220.00000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 7

	HEIGHT	WEIGHT
HEIGHT	1.00000 0.0	0.97625 0.0002
WEIGHT	0.97625 0.0002	1.00000 0.0

# Using WITH statement

```
PROC CORR DATA=CORR_EX1;  
  TITLE 'CORRELATION MATRIX FOR HEIGHT AGAINST  
        WEIGHT AND AGE USING THE WITH STATEMENT';  
  VAR HEIGHT ;  
  WITH WEIGHT AGE;  
RUN;
```

## Correlation Analysis

2 ' WITH' Variables: WEIGHT AGE  
 1 ' VAR' Variables: HEIGHT

### Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
WEIGHT	7	155.57143	45.79613	1089	99.00000	220.00000
AGE	6	31.16667	12.41639	187.00000	20.00000	48.00000
HEIGHT	7	66.71429	3.90360	467.00000	61.00000	72.00000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0  
 / Number of Observations

	HEIGHT
WEIGHT	0.97625 0.0002 7
AGE	0.86614 0.0257 6

# Using NOSIMPLE statement

```
PROC CORR DATA=CORR_EX1 NOSIMPLE;  
  TITLE 'EXAMPLE OF CORRELATION MATRIX WITHOUT SIMPLE  
        DESCRIPTIVE STATISTICS';  
RUN;
```

# EXAMPLE OF CORRELATION MATRIX WITHOUT SIMPLE

## Correlation Analysis

4 'VAR' Variables: SUBJECT HEIGHT WEIGHT AGE

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0  
/ Number of Observations

	SUBJECT	HEIGHT	WEIGHT	AGE
SUBJECT	1.00000 0.0 7	0.49411 0.2597 7	0.50877 0.2436 7	0.73124 0.0986 6
HEIGHT	0.49411 0.2597 7	1.00000 0.0 7	0.97625 0.0002 7	0.86614 0.0257 6
WEIGHT	0.50877 0.2436 7	0.97625 0.0002 7	1.00000 0.0 7	0.92496 0.0082 6
AGE	0.73124 0.0986 6	0.86614 0.0257 6	0.92496 0.0082 6	1.00000 0.0 6

# Using COV and NOCORR statement

```
PROC CORR DATA=CORR_EX1 COV NOCORR;  
    TITLE ' FINDING COVARIANCE AND NO CORRELATION' ;  
VAR HEIGHT WEIGHT AGE;  
RUN;
```

FINDING COVARIANCE AND NO CORRELATION

Correlation Analysis

3 'VAR' Variables: HEIGHT WEIGHT AGE

Variances and Covariances

COV('W', 'V') / VAR('W') / VAR('V') / DF('W', 'V')

'W' \ 'V'	HEIGHT	WEIGHT	AGE
HEIGHT	15.238095 15.238095 15.238095 6	174.523810 15.238095 2097.285714 6	45.500000 17.900000 154.166667 5
WEIGHT	174.523810 2097.285714 15.238095 6	2097.285714 2097.285714 2097.285714 6	570.566667 2468.166667 154.166667 5
AGE	45.500000 154.166667 17.900000 5	570.566667 154.166667 2468.166667 5	154.166667 154.166667 154.166667 5

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	7	66.71429	3.90360	467.00000	61.00000	72.00000
WEIGHT	7	155.57143	45.79613	1089	99.00000	220.00000
AGE	6	31.16667	12.41639	187.00000	20.00000	48.0000

## Using RANK statement

```
PROC CORR DATA=CORR_EX1 RANK;  
    TITLE 'EXAMPLE OF CORRELATION MATRIX WITH  
          THE RANK OPTION';  
VAR HEIGHT WEIGHT AGE;  
RUN;
```

Variabl e	N	Mean	Std Dev	Sum	Mi ni mum	Maxi mum
HEI GHT	7	66. 71429	3. 90360	467. 00000	61. 00000	72. 00000
WEI GHT	7	155. 57143	45. 79613	1089	99. 00000	220. 00000
AGE	6	31. 16667	12. 41639	187. 00000	20. 00000	48. 00000

Pearson Correlati on Coeffi ci ents / Prob > |R| under Ho: Rho=0  
/ Number of Observations

HEI GHT

HEI GHT	WEI GHT	AGE
1. 00000	0. 97625	0. 86614
0. 0	0. 0002	0. 0257
7	7	6

WEI GHT

WEI GHT	HEI GHT	AGE
1. 00000	0. 97625	0. 92496
0. 0	0. 0002	0. 0082
7	7	6

AGE

AGE	WEI GHT	HEI GHT
1. 00000	0. 92496	0. 86614
0. 0	0. 0082	0. 0257
6	6	6

## Using Best statement

```
PROC CORR DATA=CORR_EX1 BEST=2;  
  TITLE 'EXAMPLE OF CORRELATION MATRIX WITH  
        THE BEST OPTION' ;  
VAR HEIGHT WEIGHT AGE;  
RUN;
```

Variabl e	N	Mean	Std Dev	Sum	Mi ni mum	Maxi mum
HEI GHT	7	66. 71429	3. 90360	467. 00000	61. 00000	72. 00000
WEI GHT	7	155. 57143	45. 79613	1089	99. 00000	220. 00000
AGE	6	31. 16667	12. 41639	187. 00000	20. 00000	48. 00000

Pearson Correlati on Coeffi ci ents / Prob > |R| under Ho: Rho=0  
/ Number of Observations

HEI GHT	HEI GHT	WEI GHT
	1. 00000	0. 97625
	0. 0	0. 0002
	7	7

WEI GHT	WEI GHT	HEI GHT
	1. 00000	0. 97625
	0. 0	0. 0002
	7	7

AGE	AGE	WEI GHT
	1. 00000	0. 92496
	0. 0	0. 0082
	6	6

## Using OUTPUT= statement

```
PROC CORR DATA=CORR_EX1 OUTPUT=PEAR_COR NOPRINT;  
    TITLE 'CREATE A DATA SET THAT CONTAINS THE OUTPUT' ;  
VAR HEIGHT WEIGHT AGE;  
RUN;  
  
PROC PRINT DATA=PEAR_COR;  
RUN;
```

CREATE A DATA SET THAT CONTAINS THE OUTPUT

OBS	_TYPE_	_NAME_	HEI GHT	WEI GHT	AGE
1	MEAN		66.7143	155.571	31.1667
2	STD		3.9036	45.796	12.4164
3	N		7.0000	7.000	6.0000
4	CORR	HEI GHT	1.0000	0.976	0.8661
5	CORR	WEI GHT	0.9762	1.000	0.9250
6	CORR	AGE	0.8661	0.925	1.0000

# Example:

In a poll of men and women television viewers, preferences for the top 10 shows led to the following rankings. Is there a relationship between the ranking provided for the two groups?

Television Show	Ranking by Men	Ranking by Women
1	1	5
2	5	10
3	8	6
4	7	4
5	2	7
6	3	2
7	10	9
8	4	8
9	6	1
10	9	3

# Input the data into SAS

```
DATA CORR_EX2;  
  INFILE 'a:\TVRANK.TXT' ;  
  INPUT TELSHOW MEN WOMEN;  
RUN;
```

# Using Spearman Option

```
PROC CORR DATA=CORR_EX2 SPEARMAN;  
VAR MEN WOMEN;  
RUN;
```

CORRELATION MATRIX FOR HEIGHT AND WEIGHT

Correlation Analysis

2 'VAR' Variables: MEN WOMEN

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
MEN	10	5.50000	3.02765	5.50000	1.00000	10.00000
WOMEN	10	5.50000	3.02765	5.50000	1.00000	10.00000

Spearman Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 10

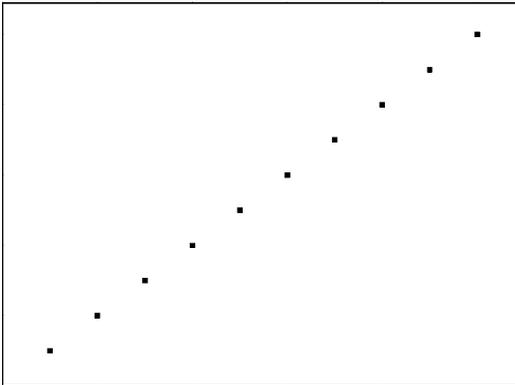
	MEN	WOMEN
MEN	1.00000 0.0	0.04242 0.9074
WOMEN	0.04242 0.9074	1.00000 0.0

# Significance Of a Correlation Coefficient

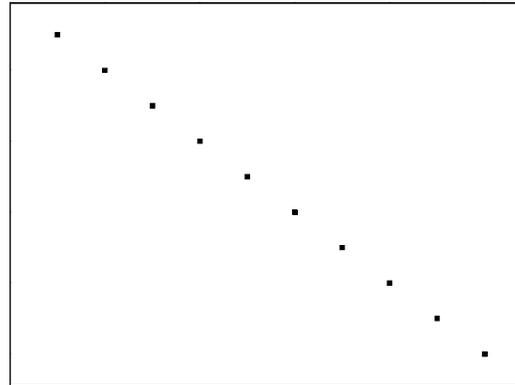
- How large a correlation coefficient should be to say that the two variables are correlated?
  - When PROC CORR is used to find the correlation coefficient, it also prints a probability associated with it. This probability is the probability of obtaining a correlation coefficient as large or larger than the one obtained by chance alone. (Discuss more)

# Values of $r$ and Their Implications

$r = +1$  perfect positive  
linear relationship  
between  $x$  and  $y$



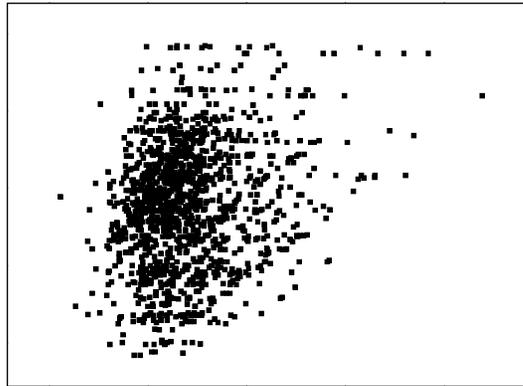
$r = -1$  perfect negative  
linear relationship  
between  $x$  and  $y$



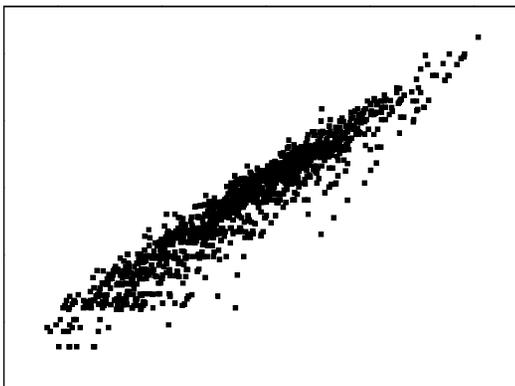
$r = -0.06$  virtually no  
correlation



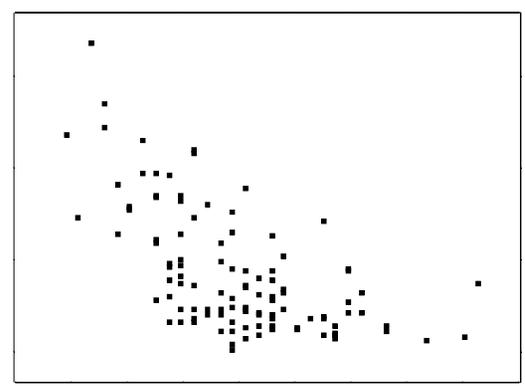
$r = +0.20$  weak positive linear relationship



$r = +0.96$  strong positive linear relationship



$r = -0.60$  moderate negative linear relationship



## How to interpret a Correlation Coefficient?

- Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. Specifically, a correlation should not and cannot be interpreted as proof of a cause-and-effect relation between the two variables.
- The value of a correlation can be affected greatly by the range of scores represented in the data.
- One or two extreme data points can have a dramatic effect on the value of a correlation

- When judging “how good” a relationship is, it is tempting to focus on the numerical value of the correlation. For example, a correlation of +0.5 is half way between 0 and 1.00 and therefore appears to represent a moderate degree of relation. However, a correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is 100% perfectly predictable relation between X and Y, a correlation of 0.5 does not mean that you can make prediction with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus, a correlation of  $r = 0.5$  provides  $r^2 = 0.25$  or 25% accuracy.

## Coefficient of Determination?

- The best way to interpret a correlation coefficient is to look at the square of the coefficient since it gives the proportion of variance in one of the variables that can be explained by variation in the other variable.

# Partial Correlation

- To compute the strength of the relationship between two variables when the effect of the other variable is removed can be done by computing a Partial Correlation.

```
PROC CORR DATA=CORR_EX1 NOSIMPLE ;  
    TITLE 'EXAMPLE OF PARTIAL CORRELATION MATRIX ' ;  
VAR HEIGHT WEIGHT ;  
PARTIAL AGE;  
RUN;
```

## EXAMPLE OF PARTIAL CORRELATION MATRIX

### Correlation Analysis

1 'PARTIAL' Variables: AGE  
2 'VAR' Variables: HEIGHT WEIGHT

Pearson Partial Correlation Coefficients / Prob > |R| under Ho: Partial Rho=0  
/ N = 6

	HEIGHT	WEIGHT
HEIGHT	1.00000 0.0	0.91934 0.0272
WEIGHT	0.91934 0.0272	1.00000 0.0

# Regression Analysis

- Regression Analysis enables you to describe the relationship between variables.
- Two types of variables are important in regression analysis
  - dependent variable (response variable): is the variable in interest.
  - independent variable (predictors, regressors and explanatory variables): explains the variability of the dependent variable

- Here we are interested in Simple Linear regression
- The simple linear regression model is given by

$$Y = b_0 + b_1X + \varepsilon$$

where

$Y$  is the dependent variable

$X$  is the independent variable

$b_0$  is the intercept parameter

$b_1$  is the slope parameter

$b_0 + b_1X$  is the mean of all  $Y$ 's associated with  $X$

$\varepsilon$  is the error term representing the deviation of  $Y$  about  $b_0 + b_1X$

- The error term  $\varepsilon$ , has the following properties
  - mean of 0 at each  $X$ .
  - normally distributed at each value of  $X$
  - the same variance at each value of  $X$
  - independent

# PROC REG: *Introduction*

- **PROC REG** is a general-purpose procedure for regression, while other regression procedures in the SAS System implement more specialized applications. **PROC REG** provides nine model-selection methods, tests linear hypotheses and multivariate hypotheses, generates scatter plots of data and various statistics, computes collinearity diagnostics and influence statistics, produces partial leverage plots, and outputs statistics to a SAS data set, including predicted values, residuals, ridge regression estimates and confidence limits.

# PROC REG: *Syntax*

PROC REG options;

label: MODEL dependents= regressors / <options>;

BY variable-list;

FREQ variable;

ID variable;

VAR variable-list;

ADD variable-list;

DELETE variable-list;

REWEIGHT <condition|ALLOBS> </options> | <STATUS|UNDO>;

WEIGHT variable;

label: MTEST <equation1, ... equationk / options>;

OUTPUT OUT= SAS-data-set keyword= names ...;

PAINT <condition|ALLOBS> </options> | <STATUS|UNDO>;

PLOT <yvariable1*xvariable1> <=symbol1>,...

<yvariablek*xvariablek> <=symbolk> </options>;

PRINT <options ANOVA MODELDATA>;

REFIT;

RESTRICT equation1, ... equationk;

label: TEST equation1, ... equationk / option;

## PROC REG OPTIONS;

**MODEL DEPENDENT VARIABLE (S) =  
INDEPENDENT VARIABLE / OPTION;**

**RUN;**

PROC REG options;

The following options can appear in the PROC REG statement.

ALL	ANNOTATE= SAS-data-set
CORR	COVOUT
DATA= SAS-data-set	GOUT= graphics-catalog
GRAPHICS	NOPRINT
OUTEST= SAS-data-set	OUTSEB
OUTSSCP= SAS-data-set	OUTSTB
OUTVIF	PCOMIT= values
PRESS	RIDGE= values
SIMPLE	SINGULAR= n
USSCP	

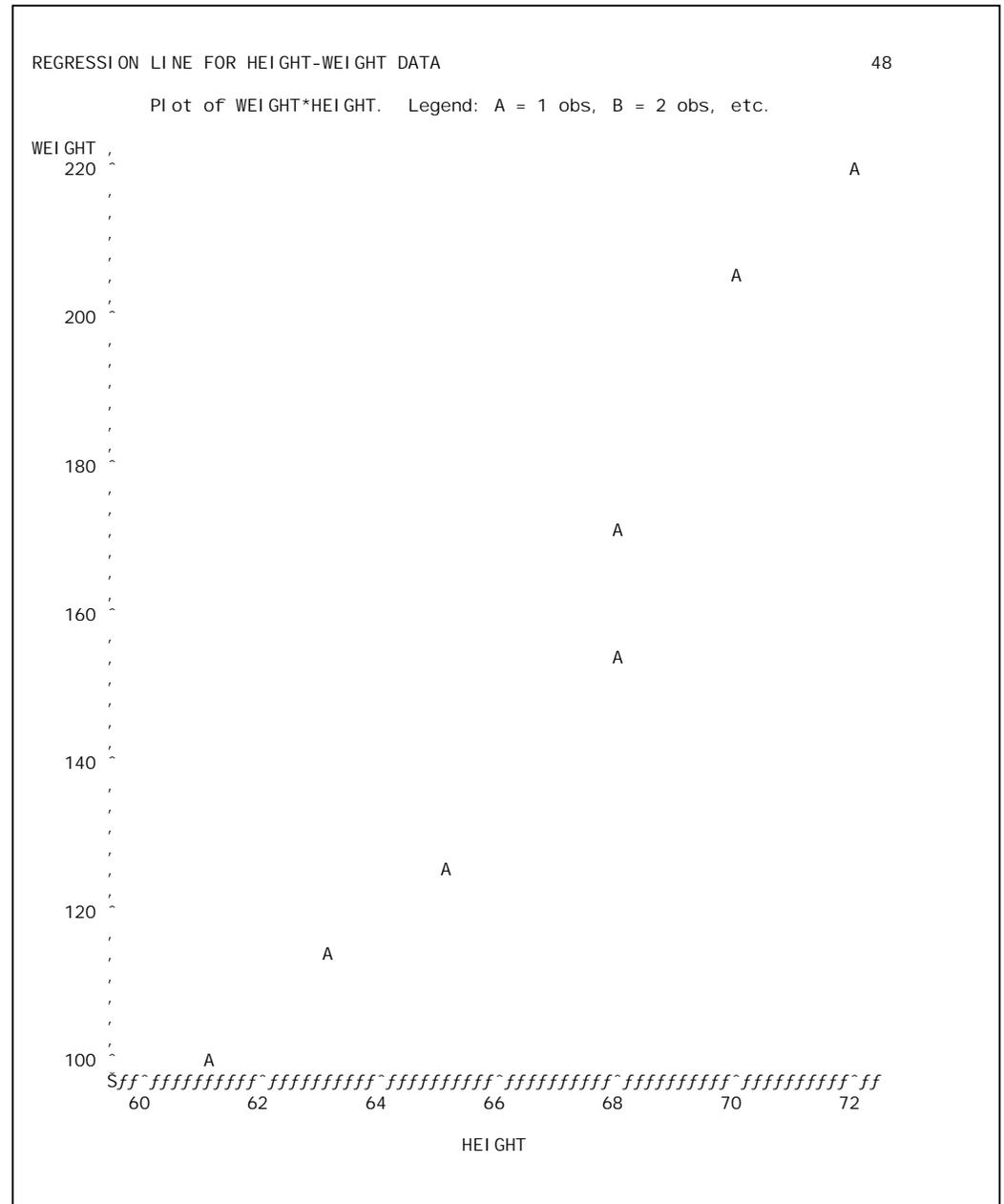
# Example:

Given the person's height, what would be the person weight.

How could we best determine the relationship between height and weight?

Looking at the graph we see that the relationship is linear.

So we are going to fit a linear model.



```
PROC REG DATA=CORR_EX1;  
  TITLE 'REGRESSION LINE FOR HEIGHT-WEIGHT DATA' ;  
  MODEL WEIGHT=HEIGHT;  
RUN;
```

REGRESSION LINE FOR HEIGHT-WEIGHT DATA

Model : MODEL1

Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	11993.05804	11993.05804	101.523	0.0002
Error	5	590.65625	118.13125		
C Total	6	12583.71429			

Root MSE	10.86882	R-square	0.9531
Dep Mean	155.57143	Adj R-sq	0.9437
C. V.	6.98639		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	-608.515625	75.94454359	-8.013	0.0005
HEIGHT	1	11.453125	1.13668841	10.076	0.0002

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	11993.05804	11993.05804	101.523	0.0002
Error	5	590.65625	118.13125		
C Total	6	12583.71429			

•In the Analysis of Variance table, the model, error, and corrected total sums of squares are provided.

**SSM** : gives the amount of variability in weight explained by height

**SSE** : gives the amount of unexplained variability in weight

**SST** : gives the total amount of variability

Note that

$$\mathbf{SST = SSM + SSE}$$

- The degrees of freedom (DF) is also provides which represent the amount of independent information in each sum.
- The model mean square **MSM** and the error mean square **MSE** is also provided.

$$\mathbf{MSM = SSM/df}$$

$$\mathbf{MSE = SSE/df}$$

- MSE represent the estimate of the variance of the Weight at each value of Height.
- The model **F** statistic and the **p-value** to test the significance of the model is given too. (Here the p-value is  $0.0002 < 0.05$  so the model is significant so we conclude that height explains a significant portion of variability in weight so it might be useful in the analysis)

Root MSE	10.86882	R-square	0.9531
Dep Mean	155.57143	Adj R-sq	0.9437
C. V.	6.98639		

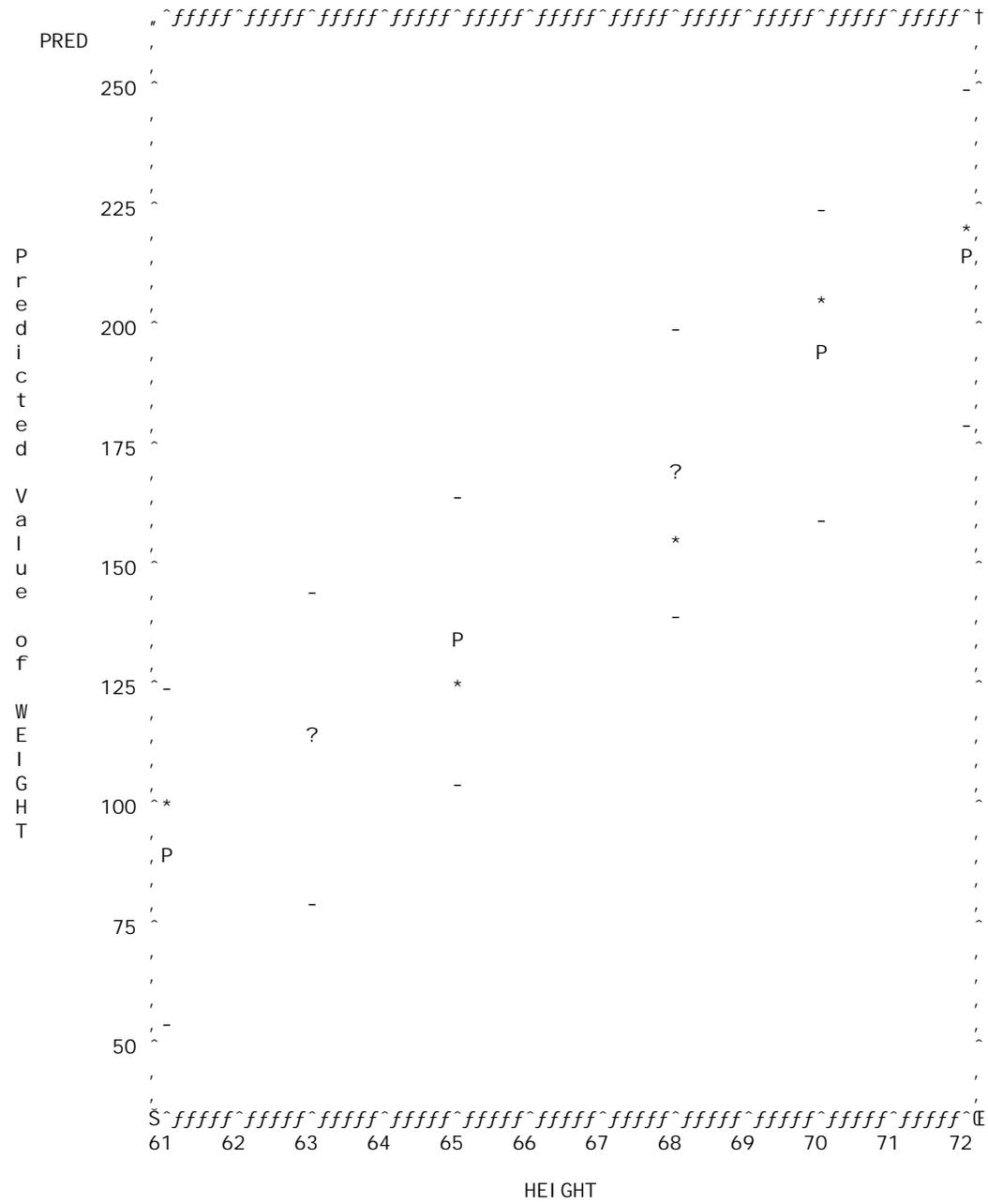
- Root MSE : the estimate of the standard deviation of weight at each value of the height (10.86882).
- Dep Mean : is the mean of the dependent variable (155.57143)
- C.V. : coefficient of variation which is a unitless measure and it measures the amount of variability compared to the magnitude of the mean of the dependent variable (6.98639).
- R-square : the coefficient of determination ( $SSM/SST = 0.9531$  i.e 95% of the variability in weight is explained by height).
- Adj R-sq : Adjusted r squared which is used to compare two models for the same data (The larger the better).

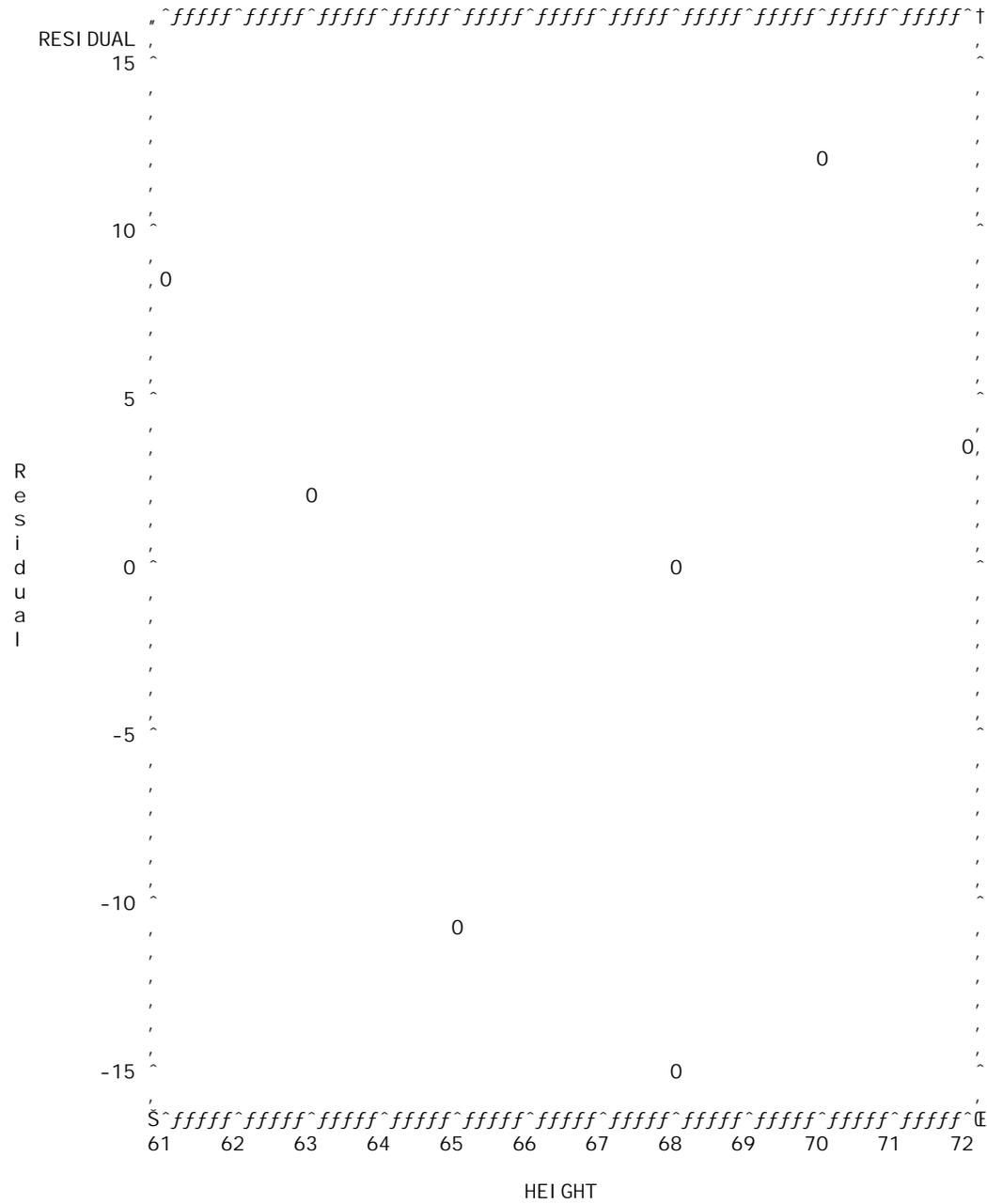
Parameter Estimates					
Variabl e	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
I NTERCEP	1	-608. 515625	75. 94454359	-8. 013	0. 0005
HEI GHT	1	11. 453125	1. 13668841	10. 076	0. 0002

- The estimate of the intercept parameter is -608.515625
- The estimate of the slop parameter is 11.453125
- The t-statistic an the p-value for testing the significance of the parameter estimate is given. Both of them look significant.

## Plotting Predicted, Residuals and Confidence Limits

```
PROC REG DATA=CORR_EX1;  
  TITLE ' PLOT PREDICTED, RESIDUALS CONFIDENT LIMITS' ;  
  MODEL WEI GHT=HEI GHT;  
  PLOT PREDI CTED. *HEI GHT=' P'  
        U95. *HEI GHT=' -' L95. *HEI GHT=' -'  
        WEI GHT*HEI GHT=' *' / OVERLAY;  
  PLOT RESI DUAL. *HEI GHT=' O' ;  
RUN;
```





Looking at the residual plot it seems that there is a curve pattern .  
So adding a new quadratic term may improve the model.

```
DATA REG_QUAD;  
  INFILE 'a:\HTWT.TXT' ;  
  INPUT SUBJECT GENDER $ HEIGHT WEIGHT AGE;  
  HEIGHT2 = HEIGHT**2 ;  
CARDS;  
  
PROC REG DATA=REG_QUAD;  
  MODEL WEIGHT=HEIGHT HEIGHT2;  
  PLOT RESIDUAL.*HEIGHT='0' ;  
RUN;
```

Model : MODEL1

Dependent Variable: WEIGHT

### Analysis of Variance

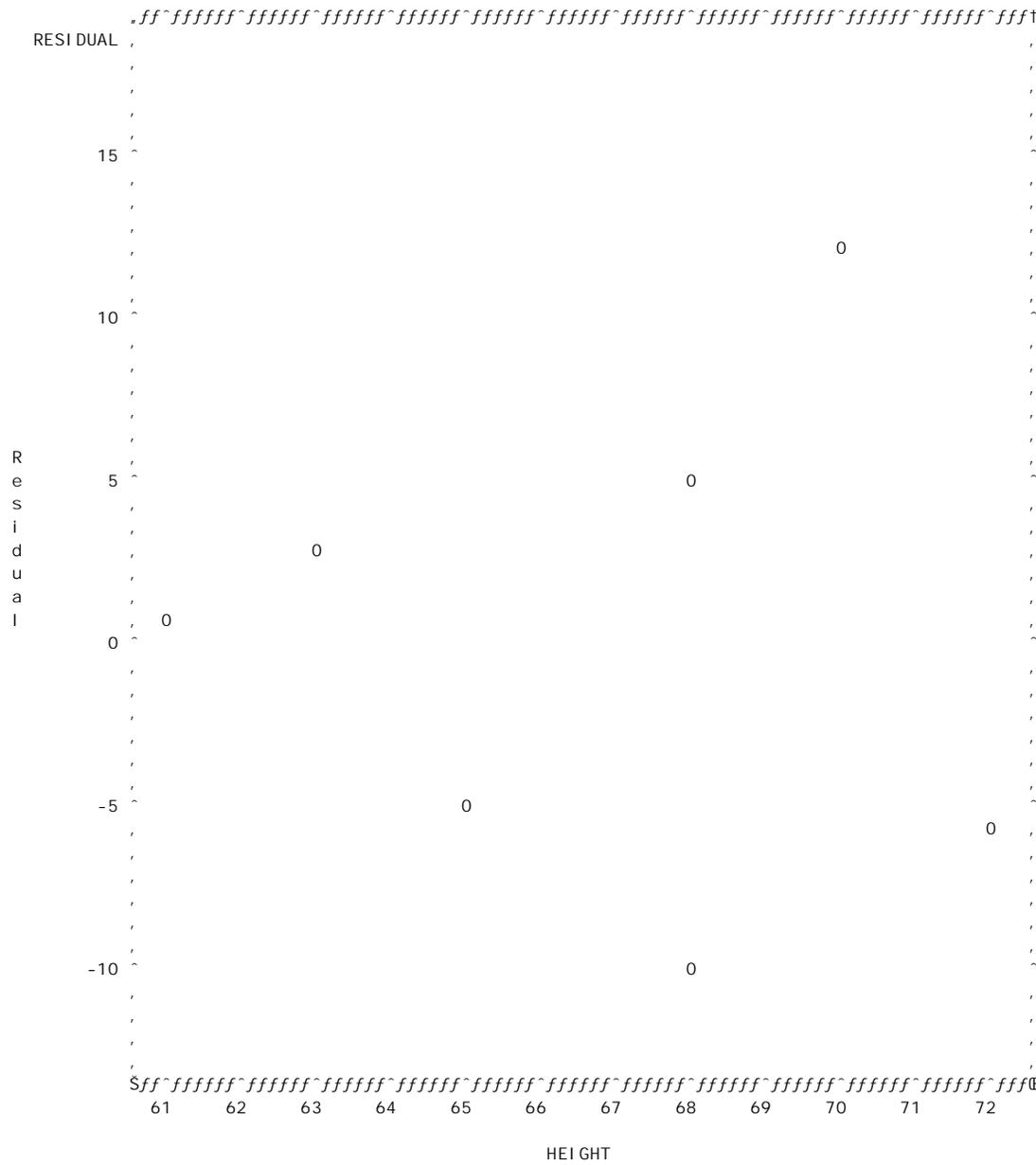
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	12253.83381	6126.91691	74.293	0.0007
Error	4	329.88047	82.47012		
C Total	6	12583.71429			

Root MSE	9.08131	R-square	0.9738
Dep Mean	155.57143	Adj R-sq	0.9607
C. V.	5.83739		

### Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	1707.608805	1304.0412816	1.309	0.2605
HEIGHT	1	-58.506812	39.35415596	-1.487	0.2113
HEIGHT2	1	0.526720	0.29620643	1.778	0.1500



Notice now that the residual plot shows that the residuals are more random.

# Multiple Regression

## Example:

```
PROC REG DATA=CORR_EX1;  
  TITLE 'REGRESSION LINE FOR HEIGHT-WEIGHT DATA';  
  MODEL WEIGHT=HEIGHT AGE;  
RUN;
```

REGRESSION LINE FOR HEIGHT-WEIGHT DATA

Model : MODEL1

Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	12064.86242	6032.43121	65.577	0.0033
Error	3	275.97091	91.99030		
C Total	5	12340.83333			
Root MSE	9.59116	R-square	0.9776		
Dep Mean	153.16667	Adj R-sq	0.9627		
C. V.	6.26191				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	-432.574545	116.79733503	-3.704	0.0342
HEIGHT	1	8.209091	2.02846082	4.047	0.0272
AGE	1	1.278182	0.69119072	1.849	0.1615

## Some Selection Methods for PROC REG

- **FORWARD:** Start with the best single regressor, then finds the best one to add to what exist; the next best, etc.
- **BACKWARD:** Starts with all variables in the equation, then drops the worst one, then the next, etc.
- **STEPWISE:** Similar to **FORWARD** except that there is an additional step where all variables in each equation are checked again to see if they remain significant after the new variables has been entered.

# Example:

```
PROC REG DATA=CORR_EX1;  
  TITLE 'REGRESSION LINE FOR HEIGHT-WEIGHT DATA';  
  MODEL WEIGHT=HEIGHT AGE / selection=stepwise;  
RUN;
```

REGRESSION LINE FOR HEIGHT-WEIGHT DATA

Stepwise Procedure for Dependent Variable WEIGHT

Step 1 Variable HEIGHT Entered R-square = 0.95214657 C(p) = 4.41971155

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	11750.28212291	11750.28212291	79.59	0.0009
Error	4	590.55121043	147.63780261		
Total	5	12340.83333333			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-608.79702048	85.55396457	7475.88014592	50.64	0.0021
HEIGHT	11.45810056	1.28436147	11750.28212291	79.59	0.0009

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1500 level.  
 No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable WEIGHT

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	HEIGHT		1	0.9521	0.9521	4.4197	79.5886	0.0009

# Question

# Describing Data

## Question 1:

Given the data in the following table:

Use PROC UNIVARIATE to produce normal probability plots and box plot and test the distributions for normality.

Do this for the variables REACT, LIVER_WT, and SPLEEN, first for all subject and then for the two doses.

Create a summary data set that contains the mean, std, median, 20th and 80th percentile.

S U B J E C T	D O S E	R E A C T	L I V E R _ W T	S P L E E N
1	1	5 . 4	1 0 . 2	8 . 9
2	1	5 . 9	9 . 8	7 . 3
3	1	4 . 8	1 2 . 2	9 . 1
4	1	6 . 9	1 1 . 8	8 . 8
5	1	1 5 . 8	1 0 . 9	9 . 0
6	2	4 . 9	1 3 . 8	6 . 6
7	2	5 . 0	1 2 . 0	7 . 9
8	2	6 . 7	1 0 . 5	8 . 0
9	2	1 8 . 2	1 1 . 9	6 . 9
1 0	2	5 . 5	9 . 9	9 . 1

## Question 2:

Given the data in the following table:

Write a SAS program to compare the sales records of the company's three sales people. (Compute the sum and the mean number of visits, phone calls, and units sold for each salesman.)

Plot the number of visits against the number of phone calls. Use the value of Sales person as the plotting symbol.

Make a frequency bar chart for each Salesperson for the sum of "units sold."

Salesperson	Target company	Number of Visits	Number of phone calls	Units Sold
Brown	American	3	12	28,000
Johnson	VRW	6	14	33,000
Rivera	Texam	2	6	8,000
Brown	Standard	0	22	0
Brown	Knowles	2	19	12,000
Rivera	Metro	4	8	13,000
Rivera	Uniman	8	7	27,000
Johnson	Oldham	3	16	8,000
Johnson	Rondo	2	14	2,000

# Categorical Data Analysis

## Question 1:

Run the program below to create a SAS data set called DEMOG:

```
DATA DEMOG;
```

```
    INPUT WEIGHT HEIGHT GENDER $;
```

```
CARDS;
```

```
155 68 M
```

```
98 60 F
```

```
202 72 M
```

```
289 75 M
```

```
130 63 F
```

```
:
```

**WEIGHT 0 - 100 = 1**

**101 - 150 = 2**

**151 - 200 = 3**

**> 200 = 4**

**HEIGHT 0 - 70 = 1**

**> 70**

**We then want to generate a table of WEIGHT categories (rows) by HEIGHT categories (columns). Recode these variables in two ways: (1) with “IF” statements; (2) with formats. Then write the necessary statements statements to generate the table.**

## Question2 :

A school administrator believes that larger class sizes lead to more discipline problems. Four class sizes (small, medium, large and gigantic) are tested. The table below summarizes the problems recorded for each of the class sizes. Treating class size as an ordinal variable, test if there is a linear increase in the proportion of class problems.

Class Size	Small	Medium	Large	Gigantic
Problem	3	6	17	80
No Problem	12	22	38	120
Total	15	28	55	200

### Question3 :

A friend gives you some summary data on the relationship between socio-economic status (SES) and asthma, as follows:

Asthma	Yes	No
Low SES	40	100
High SES	30	130

Create a SAS data set from these data and compute chi-square.

# Correlation and Regression

## Question 1 :

Given the following data:

X	Y	Z
1	3	1 5
7	1 3	7
8	1 2	5
3	4	1 4
4	7	1 0

- Write a SAS program and compute the Pearson correlation coefficient between X and Y; X and Z. What is the significance of each.

- Change the correlation request to produce a correlation coefficient between each variable versus every other variable.
- Compute a regression line (Y on X). Y is the dependent variable, X is the independent variable.
- What is the slope and the intercept:
- Are they significantly different from zero?
- Generate a plot of Y versus X.
- A plot of the regression line and the original data on the same set of axis.

## Question 2:

Given the data set

County	POP	HOSPITAL	Fire_CO	RURAL
1	35	1	2	YES
2	88	5	8	NO
3	5	0	1	YES
4	55	3	3	YES
5	75	4	5	NO
6	125	5	8	NO
7	225	7	9	YES
8	500	10	11	NO

- a. Write a SAS program to create a SAS data set of the data above.
- b. Run PROC UNIVARIATE to check the distributions for the variables POP, HOSPITAL, and FIRE_CO.
- c. Compute a correlation matrix for the variables POP, HOSPITAL, and FIRE_CO. Produce both Pearson and Spearman Correlation. Which is more appropriate?
- d. Recode POP, HOSPITAL, and FIRE_CO so that they each have two levels (use median cut or a value somewhere near the 50th percentile). Compute crosstabulations between the variable RURAL and the recorded variables.

# T-test & Nonparametric Comparison

## Question 1:

The following table shows the time for subjects to feel relief from headache pain

Aspirin	Tylenol
4 0	3 5
4 2	3 7
4 8	4 2
3 5	2 2
6 2	3 8
3 5	2 9

- A.** Write a SAS program to read these data and perform a t-test. Is either product significantly faster than the other (at 0.05 level).
- B.** Perform a Wilcoxon rank-sum test. Include a request for an exact p-value.

## Question 2:

Eight subjects are tested to see which of two medications (A or B) works best for headaches. Each subject tries each of the two drugs (for two different headaches) and the time span to pain relief is measured. Ignoring an order effect, what type of test would you use to test if the drug is faster than the other? Write a SAS statements to run the appropriate analysis.

Subject	Drug A	Drug B
1	20	18
2	40	36
3	30	30
4	45	46
5	19	15
6	27	22
7	32	29
8	26	25

# Solution

# Describing Data

# Question 1

```
DATA PATIENT;  
  INPUT SUBJECT DOSE REACT LIVER_WT SPLEEN;  
CARDS;  
1  1  5.40 10.2 8.9  
2  1  5.90 9.80 7.3  
3  1  4.80 12.2 9.1  
4  1  6.90 11.8 8.8  
5  1 15.8 10.9 9.0  
6  2  4.90 13.8 6.6  
7  2  5.00 12.0 7.9  
8  2  6.70 10.5 8.0  
9  2 18.2 11.9 6.9  
10 2  5.50 9.90 9.1  
;  
  
PROC UNIVARIATE DATA=PATIENT NORMAL PLOT;  
  VAR REACT LIVER_WT SPLEEN;  
RUN;
```

```
PROC SORT DATA=PATIENT;  
  BY DOSE;  
RUN;
```

```
PROC UNIVARIATE DATA=PATIENT NORMAL PLOT;  
  VAR REACT LIVER_WT SPLEEN;  
  BY DOSE;  
RUN;
```

```
PROC UNIVARIATE DATA=PATIENT NOPRINT;  
  VAR REACT LIVER_WT SPLEEN;  
  BY DOSE;  
  OUTPUT OUT=UNIVOUT MEAN=RMEAN LMEAN SMEAN  
    MEDIAN=RMEDIAN LMEDIAN SMEDIAN  
    PCTLPRE=R_ L_ S_  
    PCTLPTS=20 80;  
RUN;
```

```
PROC PRINT DATA=UNIVOUT;  
RUN;
```

## Question 2

```
DATA SALES;  
  INPUT NAME $ COMPANY $ N_VISITS N_PHONE N_UNITS;  
CARDS;  
BROWN   AMERICAN   3   12  28000  
JOHNSON VRW        6   14  33000  
RIVERA  TEXAM      2    6   8000  
BROWN   STANDARD   0   22    0  
BROWN   KNOWLES    2   19  12000  
RIVERA  METRO      4    8  13000  
RIVERA  UNIMAN     8    7  27000  
JOHNSON OLDHAM     3   16   8000  
JOHNSON RONDO      2   14   2000  
;  
  
PROC MEANS DATA=SALES N SUM MEAN;  
  VAR N_VISITS N_PHONE N_UNITS;  
  CLASS NAME;  
RUN;
```

```
PROC PLOT DATA=SALES;  
  PLOT N_VISITS*N_PHONE=NAME;  
RUN;
```

```
PROC CHART DATA=SALES;  
  VBAR N_UNITS / GROUP=NAME;  
RUN;
```

# Categorical Data Analysis

# Question 1

```
*using IF satement*;
DATA DEMOG;
  INPUT WEIGHT HEIGHT GENDER $;
  IF 0 LE WEIGHT LT 101 THEN WTGRP=1;
  ELSE IF 101 LE WEIGHT LT 151 THEN WTGRP=2;
  ELSE IF 151 LE WEIGHT LT 200 THEN WTGRP=3;
  ELSE WTGRP=4;
  IF 0 LE HEIGHT LE 70 THEN HTGRP=1;
  ELSE HTGRP=2;
CARDS;
155 68 M
 98 60 F
202 72 M
280 75 M
130 63 F
;
PROC FREQ DATA=DEMOG;
  TABLES WTGRP*HTGRP;
RUN;
```

# Question 1

```
* Using PROC FORMAT*;
PROC FORMAT;
  VALUE WET      LOW-100=' 1'
                101-150=' 2'
                151-200=' 3'
                200-HIGH=' 4' ;
  VALUE HET      0-70=' 1'
                70-HIGH=' 2' ;

RUN;
DATA DEMOG;
  INPUT WEIGHT HEIGHT GENDER $;
CARDS;
155 68 M
 98 60 F
202 72 M
280 75 M
130 63 F
;
PROC FREQ DATA=DEMOG;
  TABLES WEIGHT*HEIGHT;
  FORMAT WEIGHT WET.  HEIGHT HET. ;
RUN;
```

## Question 2

```
DATA TEST;  
  INPUT SES $ ASTHMA $ COUNT;  
CARDS;  
L Y 40  
L N 100  
H Y 30  
H N 130  
;  
  
PROC FREQ DATA=TEST;  
  TABLES SES*ASTHMA/CHI SQ;  
  WEIGHT COUNT;  
RUN;
```

# Question 3

```
PROC FORMAT;
  VALUE SIZE 1=' SMALL'
           2=' MEDI AN'
           3=' LARGE'
           4=' GI GANTI C' ;

DATA CLASS;
  INPUT SIZE PROBLEM $ COUNT;
  FORMAT SIZE SIZE. ;

CARDS ;
1 1-YES 3
1 2-NO 12
2 1-YES 6
2 2-NO 22
3 1-YES 17
3 2-NO 38
4 1-YES 80
4 2-NO 120
;

PROC FREQ DATA=CLASS;
  TABLES PROBLEM*SI ZE/CHI SQ;
WEI GHT COUNT;
RUN;
```

# Correlation & Regression

# Question 1

```
DATA PROB1;  
  INPUT X Y Z;
```

```
CARDS;
```

```
1 3 15
```

```
7 13 7
```

```
8 12 5
```

```
3 4 14
```

```
4 7 10
```

```
;
```

```
PROC CORR DATA=PROB1;
```

```
  VAR X;
```

```
  WITH Y Z;
```

```
RUN;
```

```
PROC CORR DATA=PROB1;
```

```
  VAR X Y Z;
```

```
RUN;
```

```
PROC REG DATA=PROB1;
```

```
  MODEL Y = X;
```

```
  PLOT PREDICTED. *X=' P' Y*X='*' /OVERLAY;
```

```
RUN;
```

## Question 2

```
DATA PROB2;
  INPUT COUNTY POP HOSPITAL FIRE_CO RURAL $;
CARDS;
1 35 1 2 YES
2 88 5 8 NO
3 5 0 1 YES
4 55 3 3 YES
5 75 4 5 NO
6 125 5 8 NO
7 225 7 9 YES
8 500 10 11 NO
;

PROC UNIVARIATE DATA=PROB2 NORMAL PLOT;
  VAR POP HOSPITAL FIRE_CO;
RUN;

PROC CORR DATA=PROB2 NOSIMPLE PEARSON SPEARMAN; ;
  VAR POP HOSPITAL FIRE_CO;
RUN;
```

## Question 2

```
PROC FORMAT;
  VALUE POP          LOW-81  =' BELOW MWDI AN'
                    82-HI GH=' ABOVE MEDI AN' ;
  VALUE HOSPI TAL   LOW-4    =' BELOW MEDI AN'
                    5-HI GH  =' ABOVE MEDI AN' ;
  VALUE FI RE_CO    LOW-6    =' BELOW MEDI AN'
                    7-HI GH  =' ABOVE MEDI AN' ;

RUN;

PROC FREQ DATA=PROB2;
  FORMAT POP POP.
        HOSPI TAL HOSPI TAL.
        FI RE_CO FI RE_CO. ;
  TABLES RURAL*(POP HOSPI TAL FI RE_CO) /CHI SQ;

RUN;
```

# T-test & Nonparametric Comparison

# Question 1

```
PROC FORMAT;  
  VALUE $MEDFMT ' A' =' ASPIRIN'  
              ' B' =' TYLENOL' ;
```

```
DATA HEADACE;  
  INPUT GROUP $ TIME;
```

```
CARDS;
```

```
A 40
```

```
A 42
```

```
A 48
```

```
A 35
```

```
A 62
```

```
A 35
```

```
B 35
```

```
B 37
```

```
B 42
```

```
B 22
```

```
B 38
```

```
B 29
```

```
;
```

```
PROC TTEST DATA=HEADACE;  
  CLASS GROUP;  
  VAR TIME;  
  FORMAT GROUP $MEDFMT. ;  
RUN;
```

```
PROC NPAR1WAY DATA=HEADACE ;  
  CLASS GROUP;  
  VAR TIME;  
  EXACT WILCOXON;  
RUN;
```

## Question 2

```
DATA PHEAD;
```

```
    INPUT SUBJECT DRUG_A DRUG_B;
```

```
        DIFF=DRUG_A-DRUG_B;
```

```
CARDS;
```

```
1 20 18
```

```
2 40 36
```

```
3 30 30
```

```
4 45 46
```

```
5 19 15
```

```
6 27 22
```

```
7 32 29
```

```
8 26 25
```

```
;
```

```
PROC MEANS DATA=PHEAD N MEAN STDERR T PRT;
```

```
    VAR DIFF ;
```

```
RUN;
```

## References:

- Cody, R.P. and Smith, J.K. (1997), Applied Statistics and the SAS Programming Language, New Jersey Prentice-Hall, Inc.