Fakulteit Natuur- & Landbouwetenskappe

Faculty of Natural & Agricultural Sciences

Department of Statistics

WST121

SAS Notes

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Leading Minds

CONTENTS

Introduction to SAS for Windows
(Statistical Analysis System)

Activating SAS

- Click on the START menu at the bottom left hand side of the screen
- Go to → Programs
  → SAS system
  → SAS system for windows

SAS Windows

Interaction with SAS can be accomplished through the use of three windows namely the PROGRAM EDITOR window, the LOG window and the OUTPUT window.

Activate a certain window by VIEW
→ Enhanced Editor or Program Editor (or F5)
→ Log (or F6)
→ Output (or F7)

Program Editor Window

- contains the SAS program (type in a new program or open an existing program)
- is used for writing and editing of SAS statements
- contents can be altered and filed
- to run program, use runner-icon (or F8)
- to recall program after running, use F4

LOG Window

- use this window to identify your errors in the SAS program
- contains SAS statements submitted previously
- contains error messages in red, warning messages in green
- contents cannot be altered, but can be filed
- always clear contents of the LOG window before rerunning a program
  (use EDIT → clear all)

Output Window

- contains output generated by a SAS-program
- the contents cannot be altered, but can be filed
- always clear contents of the OUTPUT window before rerunning a program
  (use EDIT → clear all)

PART 1: The DATA step

Using SAS to calculate probabilities and quantiles for known statistical distributions

Example

For $X_1 \sim \chi^2(20)$, $P(X_1 < 25.04) = p$ $p = 0.79857$

For $X_2 \sim bi(5,0.3)$, $P(X_2 \leq 3) = p$ $p = 0.96922$

For $X_3 \sim n(10,16)$, $P(X_3 \leq 5) = p$ $p = 0.10565$

SAS Program

*

```
data func1;
x1=probchi(25,20);
x2=probbnml(0.3,5,3);
x3=cdf('normal',5,10,4);
proc print;
run;
```

SAS Output

```
OBS       X1          X2          X3
 1     0.79857     0.96922     0.10565
```

DATA step	Comment
DATA FUNC1;	The keyword DATA starts the data step and indicates that a SAS data set with the name FUNC1 is to be created. Contains 1 to 8 characters. *Starts* with letters A-Z or _ (not numbers). SAS is not case sensitive.

x1=probchi(25,20);
x2=probbnml(0.3,5,3);
x3=cdf('normal',5,10,4);
	Variables x1,x2 and x3 are created using built-in SAS functions (see list on the next page)

PROC PRINT DATA=FUNC1;	The data set is printed with the PROC PRINT procedure.

Probability and Density Functions

Calculates the area on the left of a value from a specified distribution.

List of some functions

CDF
left cumulative distribution function,
e.g. CDF('NORMAL',x,μ,σ)
returns the probability that an observation from a $n(\mu,\sigma.^2)$ distribution is less than or equal to x.

POISSON((lambda,n)
Poisson probability distribution function

PROBBNML(p,n,m)
binomial probability distribution function

PROBCHI(x,df<,nc>)
chi-squared probability distribution function

PROBF(x,ndf,ddf<,nc>)
F distribution function

PROBHYPR(N,K,n,x<,r>)
hypergeometric probability distribution function

PROBNORM(x)
standard normal probability distribution function

PROBT(x,df<,nc>)
Student's t distribution function

2. Quantile Functions

Inverse distributions.

List of some functions

CINV(p,df<,nc>)
the quantile for the chi-square distribution

FINV(p,ndf,ddf<,nc>)
the quantile for the F distribution

PROBIT(argument)
inverse of the **standard** normal distribution function

TINV(p,df<,nc>)
the quantile for the t distribution

Example

For $X \sim F(7,14)$, $P(X < x) = 0.9$ $x = 2.19313$

For $Y \sim n(10,16)$ and $Z \sim n(0,1)$, P(Y<y)=*P(Z<z)=0.25* *z=-0.674* and $y=\mu+\sigma z=7.304$

SAS Program

```
data func2;
x=finv(0.9,7,14);
z=probit(0.25);
y=10+z*4;
proc print;
run;
```

SAS Output

```
OBS       X              Z            Y
 1     2.19313       -0.67449     7.30204
```

DATA step	Comment
DATA FUNC2;	The keyword DATA starts the data step and indicates that a SAS data set with the name FUNC2 is to be created. Contains 1 to 8 characters. *Starts* with letters A-Z or _ (not numbers). SAS is not case sensitive.

x=finv(0.9,7,14);
z=probit(0.25);
y=10+z*4;
	Variables x, and z are created using built-in SAS functions (see list on the next page)

PROC PRINT DATA=FUNC2;	The data set is printed with the PROC PRINT procedure.

Entering a data set directly into SAS, creating new variables and printing the data set

Consider the questionnaire below:

Questionnaire : TV-channel preferences

FOR OFFICE USE

Respno V1 ☐☐ 1-2

1 **Home Language**

Afrikaans	1
English	2
Sepedi	3
Tswana	4
Venda	5
Xhosa	6
Zulu	7

V2 ☐ 3

2 **Gender**

Male	M
Female	F

V3 ☐ 4

3 **What is your favourite TV-channel?**

SABC1	1
SABC2	2
SABC3	3
MNET	4

V4 ☐ 5

4 **How many hours do you watch TV per day?**

_____

V5 ☐☐☐☐ 6-9

5 **Age in years**

_____

V6 ☐☐ 10-11

From 12 questionnaires the following data was captured:

Respno	Language	Gender	TVChan	Time_D	Age
1	1	M	1	5	26
2	2	M	3	1	40
3	2	F	4	6.5	26
4	6	F	2	1	39
5	1	M	2	6	18
6	7	M	3	3	36
7	5	M	2	7	31
8	3	M	1	2.5	28
9	4	F	2	3.5	23
10	2	M	1	2	21
11	1	F	4	8	34
12	1	M	4	.	20

The data is entered into SAS with the program given below. This program also creates the following variables:

(i) TIME_W, which calculates the time that a respondent watches TV per week.

(ii) TGROUP, which is defined as follows:
TGROUP=H if the time that a respondent watches TV per day is 4 hours or more.
TGROUP=L if the time that a respondent watches TV per day is less than 4 hours.

(iii) LGROUP, which is defined as follows:
LGROUP=Afrikaans if a repondent's language=1.
LGROUP=English if a repondent's language=2.
LGROUP=African if a repondent's language is between 3 and 7.

The data set is then printed with the PROC PRINT procedure.

SAS Program

```
data tv;
input respno language gender $ tvchan time_d age @@;
time_w=time_d*7; *Mathematical expression;
if time_d>=4 then tgroup='H'; *recoding;
if time_d<4 then tgroup='L';
if time_d=. then tgroup=' ';
if language=1 then lgroup='Afrikaans';
if language=2 then lgroup='English';
if 3<=language<=7 then lgroup='African';
cards;
1  1       M       1       5       26
2  2       M       3       1       40
3  2       F       4       6.5     26
4  6       F       2       1       39
5  1       M       2       6       18
6  7       M       3       3       36
7  5       M       2       7       31
8  3       M       1       2.5     28
9  4       F       2       3.5     23
10 2       M       1       2       21
11 1       F       4       8       34
12 1       M       4       .       20
;
PROC PRINT DATA=TV;
RUN;
```

SAS Output

Obs	respno	language	gender	tvchan	time_d	age	time_w	tgroup	lgroup
1	1	1	M	1	5.0	26	35.0	H	Afrikaans
2	2	2	M	3	1.0	40	7.0	L	English
3	3	2	F	4	6.5	26	45.5	H	English
4	4	6	F	2	1.0	39	7.0	L	African
5	5	1	M	2	6.0	18	42.0	H	Afrikaans
6	6	7	M	3	3.0	36	21.0	L	African
7	7	5	M	2	7.0	31	49.0	H	African
8	8	3	M	1	2.5	28	17.5	L	African
9	9	4	F	2	3.5	23	24.5	L	African
10	10	2	M	1	2.0	21	14.0	L	English
11	11	1	F	4	8.0	34	56.0	H	Afrikaans
12	12	1	M	4	.	20	.		Afrikaans

DATA step	**Comment**
DATA TV;	The keyword DATA starts the data step and indicates that a SAS data set with the name TV is to be created. Contains 1 to 8 characters. *Starts* with letters A-Z or _ (not numbers).
INPUT RESPNO LANGUAGE GENDER$ TVCHAN TIME_D AGE @@;	INPUT gives names to all the variables in the data set. Contains 1 to 8 characters. *Starts* with letters A-Z or _ (not numbers). Alpha numerical data is indicated with a $ sign. @@ indicates that more than one respondent's data can be entered into a line.

TIME_W=TIME_D*7;
IF TIME_D >= 4 THEN TGROUP='H';
IF TIME_D < 4 THEN TGROUP='L';
IF TIME_D = '.' THEN TGROUP=' ';
IF LANGUAGE=1 THEN
LGROUP='AFRIKAANS';
IF LANGUAGE=2 THEN
LGROUP='ENGLISH';
IF 3<=LANGUAGE<=7 THEN
LGROUP='AFRICAN';
	New variables are created

CARDS; 1 1 M 1 5 26 2 2 M 3 1 40 3 2 F 4 6.5 26 4 6 F 2 1 39 5 1 M 2 6 18 6 7 M 3 3 36 7 5 M 2 7 31 8 3 M 1 2.5 28 9 4 F 2 3.5 23 10 2 M 1 2 21 11 1 F 4 8 34 12 1 M 4 . 20	CARDS indicates that the data follows. Data is entered in SAS. In this case only one respondent's data is entered into a line. The @@ could be omitted in the INPUT statement. No commas are allowed in SAS. Use a decimal point in stead of a comma for decimal digits. A non-response is indicated by a .
;	The data entered must be followed by a semicolon at the end.

Creating a new library in SAS

Data sets in SAS are organized in libraries. They are stored in the WORK library of SAS by default. On exiting the SAS program all the data in the WORK library is lost. Other libraries can be created and used to store data. The data in these libraries will not be lost when the program is terminated.

- Create a directory C:\WST121 **or** use the existing directory S:\WST121. The name of this directory need not be the same as that of the library in SAS.
- Activate SAS
- Select → New Library (icon)
  - → Enter the name of the new library, e.g. WST121
  - → Enable at startup
  - → Specify the Path:, e.g. C:WST121 **or** S:\WST121
  - → OK

How to import an Excel file into SAS

- Activate SAS
- Select → File
  - → Import Data
  - → Select type of format to import, e.g. Microsoft Excel 97, 2000 or 20002 Workbook
  - → Next
  - → Give the location of the Excel file, e.g. Workbook: S:\WST121
  - → OK
  - → Select the specific worksheet in the file, e.g.
    What table do you want to import? Sheet 1
  - → Next
  - → Choose the SAS destination, e.g. Library: WST121
  - → Give a name for the data set, e.g. Member: D1
  - → Next
  - → Finish
- Check LOG window
- Check data set:
  - → Explorer
  - → WST121
  - → D1

Using the SET statement to create a new data set from an existing one.

A temporary data file can be stored in a permanent library using the SET statement.

e.g. The temporary data file TV can be stored in the library WST121:

SAS Program

```
data wst121.tv;
set tv;
proc print data=wst121.tv;
```

SAS Output

The same output as that on p.9, is obtained.

Using the SET statement to create a new data set from an existing one in order to recode variables.

A temporary or permanent data file with new variables can be created from an existing one by using the SET statement.

e.g. A new temporary data file D1 with new variables can be created from WST121.TV:

SAS Program

```
data d1;
set wst121.tv;
if age<20 then agecat="  <20";
if 20<=age<30 then agecat="20-29";
if age>=30 then agecat="  30+";
run;

proc print data=d1;
var age agecat;
run;
```

SAS Output

Obs	age	agecat
1	26	20-29
2	40	30+
3	26	20-29
4	39	30+
5	18	<20
6	36	30+
7	31	30+
8	28	20-29
9	23	20-29
10	21	20-29
11	34	30+
12	20	20-29

PART 2: PROC SURVEYSELECT

Drawing a simple random sample using SAS

The following program draws a simple random sample from a given dataset.

SAS Program

```
proc surveyselect data=d1
      method=srs n=10
      out=SampleofD1;
run;
```

Note on the SAS Program

The DATA option specifies the population from which the sample must be drawn, e.g. dataset WORK.D1.
The METHOD=option specifies the sampling method, e.g. Simple Random Sampling.
The N=option specifies the size of the sample, e.g. 10.
The OUT=options names the output data set, e.g. WORK.SAMPLEOFD1.

SAS Output

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Input Data Set	D1
Random Number Seed	280921001
Sample Size	10
Selection Probability	0.059524
Sampling Weight	16.8
Output Data Set	SAMPLEOFD1

The data set can be viewed with Explorer, printed and used for further computations.

PART 3: SAS Base Procedures

1. PROC FREQ

(a) The following program constructs one way frequency tables for the categorical variables LGROUP, GENDER and TVCHAN.

SAS Program

```
proc freq data=tv;
tables lgroup gender tvchan;
run;
```

Note on SAS Program

If the TABLES statement is omitted, one-way frequency tables will be constructed for all the variables, including the continuous variables.

SAS Output

The FREQ Procedure

lgroup	Frequency	Percent	Cumulative Frequency	Cumulative Percent
African	5	41.67	5	41.67
Afrikaans	4	33.33	9	75.00
English	3	25.00	12	100.00

gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	4	33.33	4	33.33
M	8	66.67	12	100.00

tvchan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	25.00	3	25.00
2	4	33.33	7	58.33
3	2	16.67	9	75.00
4	3	25.00	12	100.00

(b) The following program constructs a one-way frequency table for the variable TVCHAN by GENDER. Note that the data is first sorted by the variable GENDER. The BY statement in PROC FREQ specifies the variable by which the table must be constructed.

SAS Program

```
proc sort data=tv;
by gender;
proc freq data=tv;
tables tvchan;
by gender;
run;
```

SAS Output

```
---------------------------------------- gender=F ----------------------------------------

                              The FREQ Procedure

                                                Cumulative    Cumulative
tvchan     Frequency     Percent     Frequency      Percent
-------------------------------------------------------------------------
     2             2       50.00             2        50.00
     4             2       50.00             4       100.00

---------------------------------------- gender=M ----------------------------------------

                              The FREQ Procedure

                                                Cumulative    Cumulative
tvchan     Frequency     Percent     Frequency      Percent
-------------------------------------------------------------------------
     1             3       37.50             3        37.50
     2             2       25.00             5        62.50
     3             2       25.00             7        87.50
     4             1       12.50             8       100.00
```

(c) The following program will construct a two-way frequency table for the variables LGROUP and TVCHAN.

SAS Program

```
proc freq data=tv;
tables language*tvchan;
run;
```

SAS Output

The FREQ Procedure

Table of lgroup by tvchan

lgroup / tvchan Frequency Percent Row Pct Col Pct	1	2	3	4	Total
African	1 8.33 20.00 33.33	3 25.00 60.00 75.00	1 8.33 20.00 50.00	0 0.00 0.00 0.00	5 41.67
Afrikaans	1 8.33 25.00 33.33	1 8.33 25.00 25.00	0 0.00 0.00 0.00	2 16.67 50.00 66.67	4 33.33
English	1 8.33 33.33 33.33	0 0.00 0.00 0.00	1 8.33 33.33 50.00	1 8.33 33.33 33.33	3 25.00
Total	3 25.00	4 33.33	2 16.67	3 25.00	12 100.00

(d) The following program will construct a three-way frequency table for the variables LGROUP, GENDER and TVCHAN.

SAS Program

```
proc freq data=tv;
tables gender*lgroup*tvchan;
run;
```

SAS Output

Table 1 of lgroup by tvchan
Controlling for gender=F

lgroup / Frequency / Percent / Row Pct / Col Pct	tvchan 1	2	3	4	Total
African	0	2	0	0	2
	0.00	50.00	0.00	0.00	50.00
	0.00	100.00	0.00	0.00	
	.	100.00	.	0.00	
Afrikaans	0	0	0	1	1
	0.00	0.00	0.00	25.00	25.00
	0.00	0.00	0.00	100.00	
	.	0.00	.	50.00	
English	0	0	0	1	1
	0.00	0.00	0.00	25.00	25.00
	0.00	0.00	0.00	100.00	
	.	0.00	.	50.00	
Total	0	2	0	2	4
	0.00	50.00	0.00	50.00	100.00

Table 2 of lgroup by tvchan
Controlling for gender=M

lgroup / Frequency / Percent / Row Pct / Col Pct	tvchan 1	2	3	4	Total
African	1	1	1	0	3
	12.50	12.50	12.50	0.00	37.50
	33.33	33.33	33.33	0.00	
	33.33	50.00	50.00	0.00	
Afrikaans	1	1	0	1	3
	12.50	12.50	0.00	12.50	37.50
	33.33	33.33	0.00	33.33	
	33.33	50.00	0.00	100.00	
English	1	0	1	0	2
	12.50	0.00	12.50	0.00	25.00
	50.00	0.00	50.00	0.00	
	33.33	0.00	50.00	0.00	
Total	3	2	2	1	8
	37.50	25.00	25.00	12.50	100.00

2. PROC MEANS

(a) The following program gives summary statistics for the numerical variables TIME_D and AGE.

SAS Program

```
proc means data=tv;
var time_d age;
run;
```

Note on SAS Program

The VAR statement specifies the continuous variables for which descriptive statistics must be calculated.

SAS Output

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
time_d	11	4.1363636	2.4808356	1.0000000	8.0000000
age	12	28.5000000	7.4893864	18.0000000	40.0000000

(b) The following program gives summary statistics for the variable TIME_D for each of the four TV channels.
All results are rounded to two decimal places.

SAS Program

```
proc sort data=tv;
by tvchan;
proc means n mean std data=tv maxdec=2;
var time_d;
by tvchan;
run;
```

Note on SAS Program

Only the sample size, mean and standard deviation are printed. MAXDEC=2 rounds all numbers to 2 decimal places. The VAR statement specifies continuous variables for which descriptive statistics must be calculated.

SAS Output

```
---------------------------------------- tvchan=1 ----------------------------------------

                          The MEANS Procedure

                       Analysis Variable : time_d

 N            Mean         Std Dev
_____________________________________
 3            3.17            1.61
_____________________________________


---------------------------------------- tvchan=2 ----------------------------------------

                       Analysis Variable : time_d

 N            Mean         Std Dev
_____________________________________
 4            4.38            2.69
_____________________________________


---------------------------------------- tvchan=3 ----------------------------------------

                       Analysis Variable : time_d

 N            Mean         Std Dev
_____________________________________
 2            2.00            1.41
_____________________________________


---------------------------------------- tvchan=4 ----------------------------------------

                       Analysis Variable : time_d

 N            Mean         Std Dev
_____________________________________
 2            7.25            1.06
_____________________________________
```

2. PROC UNIVARIATE

(a) The following program gives extensive descriptive statistics for the variable TIME_D. A test for normality is also performed.

SAS Program

```
proc univariate normal plot data=tv;
var time_d;
run;
```

Note on the SAS Program

In PROC UNIVARIATE the following options can be specified (see SAS Help for more options):

NORMAL: gives a statistic which test for normality of the data

PLOT: gives a normal probability plot

FREQ: gives a one-way frequency table of the data.

The VAR statement specifies continuous variables for which descriptive statistics must be calculated.

SAS Output

The UNIVARIATE Procedure
Variable: time_d

Moments

N	11	Sum Weights	11
Mean	4.13636364	Sum Observations	45.5
Std Deviation	2.48083564	Variance	6.15454545
Skewness	0.1914727	Kurtosis	-1.4694539
Uncorrected SS	249.75	Corrected SS	61.5454545
Coeff Variation	59.9762462	Std Error Mean	0.74800009

Basic Statistical Measures

Location		Variability	
Mean	4.136364	Std Deviation	2.48084
Median	3.500000	Variance	6.15455
Mode	1.000000	Range	7.00000
		Interquartile Range	4.50000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	5.529897	Pr > |t|	0.0003
Sign	M	5.5	Pr >= |M|	0.0010
Signed Rank	S	33	Pr >= |S|	0.0010

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.930426	Pr < W	0.4152
Kolmogorov-Smirnov	D	0.146677	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.048759	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.30693	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Quantile	Estimate
100% Max	8.0
99%	8.0
95%	8.0
90%	7.0
75% Q3	6.5
50% Median	3.5
25% Q1	2.0
10%	1.0
5%	1.0
1%	1.0
0% Min	1.0

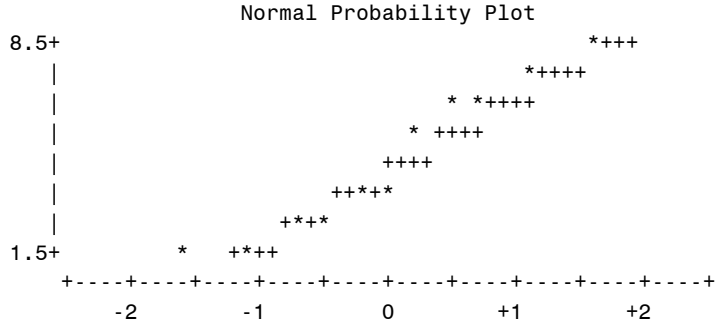
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
1.0	8	5.0	1
1.0	4	6.0	6
2.0	3	6.5	10
2.5	2	7.0	7
3.0	9	8.0	11

Missing Values

		-----Percent Of-----	
Missing Value	Count	All Obs	Missing Obs
.	1	8.33	100.00

Stem	Leaf	#	Boxplot
8	0	1	
7	0	1	
6	05	2	+-----+
5	0	1	| |
4			| + |
3	05	2	*-----*
2	05	2	+-----+
1	00	2	



Test for normality:
H_0: Data have a normal distribution
H_1: Data do not have a normal distribution
Since the *p*-value for the Shapiro-Wilk statistic is 0.4152 (> 0.05) the distribution of the data do not differ significantly from a normal distribution.

(a) The following program gives extensive descriptive statistics for the variable TIME_D by GENDER.

SAS Program

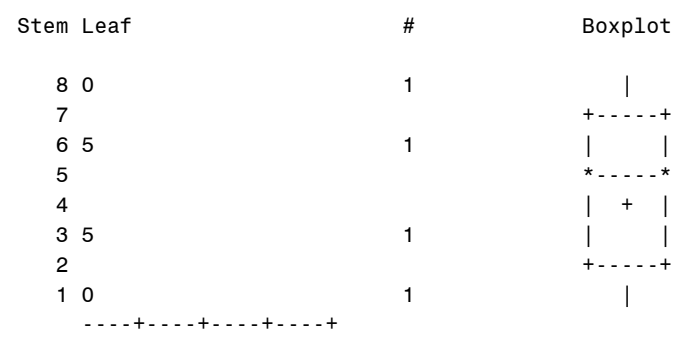
```
proc sort data=tv; by gender;
proc univariate plot data=tv;
var time_d;
by gender;
run;
```

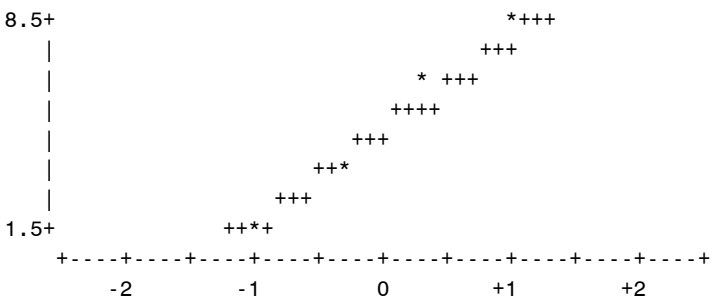
SAS Output

```
---------------------------------------- gender=F ----------------------------------------

                          The UNIVARIATE Procedure
                              Variable:  time_d

                                   Moments

N                          4    Sum Weights                4
Mean                    4.75    Sum Observations          19
Std Deviation       3.122499    Variance                9.75
Skewness           -0.328468    Kurtosis           -2.2393162
Uncorrected SS         119.5    Corrected SS            29.25
Coeff Variation    65.736821    Std Error Mean      1.5612495


                        Basic Statistical Measures

          Location                      Variability

      Mean     4.750000     Std Deviation            3.12250
      Median   5.000000     Variance                 9.75000
      Mode        .         Range                    7.00000
                            Interquartile Range      5.00000


                        Tests for Location: Mu0=0

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

      Student's t    t  3.042435    Pr > |t|    0.0558
      Sign           M         2    Pr >= |M|   0.1250
      Signed Rank    S         5    Pr >= |S|   0.1250


                      Quantiles (Definition 5)

                       Quantile      Estimate

                       100% Max          8.00
                       99%               8.00
                       95%               8.00
                       90%               8.00
                       75% Q3            7.25
                       50% Median        5.00
                       25% Q1            2.25
                       10%               1.00
                       5%                1.00
                       1%                1.00
                       0% Min            1.00
```

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
1.0	1	1.0	1
3.5	2	3.5	2
6.5	3	6.5	3
8.0	4	8.0	4





---------- gender=M ----------

The UNIVARIATE Procedure
Variable: time_d

Moments

N	7	Sum Weights	7
Mean	3.78571429	Sum Observations	26.5
Std Deviation	2.23340441	Variance	4.98809524
Skewness	0.33372642	Kurtosis	-1.5276707
Uncorrected SS	130.25	Corrected SS	29.9285714
Coeff Variation	58.9955881	Std Error Mean	0.84414752

Basic Statistical Measures

Location		Variability	
Mean	3.785714	Std Deviation	2.23340
Median	3.000000	Variance	4.98810
Mode	.	Range	6.00000
		Interquartile Range	4.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	4.48466	Pr > |t|	0.0042
Sign	M	3.5	Pr >= |M|	0.0156
Signed Rank	S	14	Pr >= |S|	0.0156

Quantiles (Definition 5)

Quantile	Estimate
100% Max	7
99%	7
95%	7
90%	7
75% Q3	6
50% Median	3
25% Q1	2
10%	1
5%	1
1%	1
0% Min	1

Extreme Observations

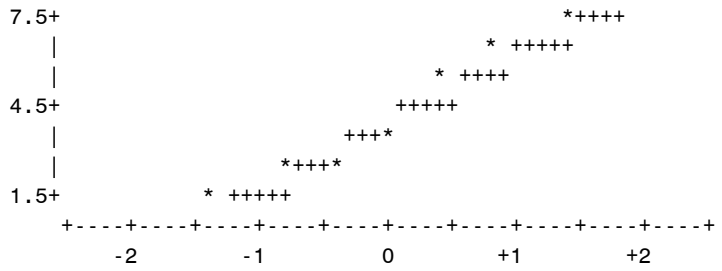
----Lowest----		----Highest---	
Value	Obs	Value	Obs
1.0	10	2.5	6
2.0	7	3.0	11
2.5	6	5.0	5
3.0	11	6.0	8
5.0	5	7.0	9

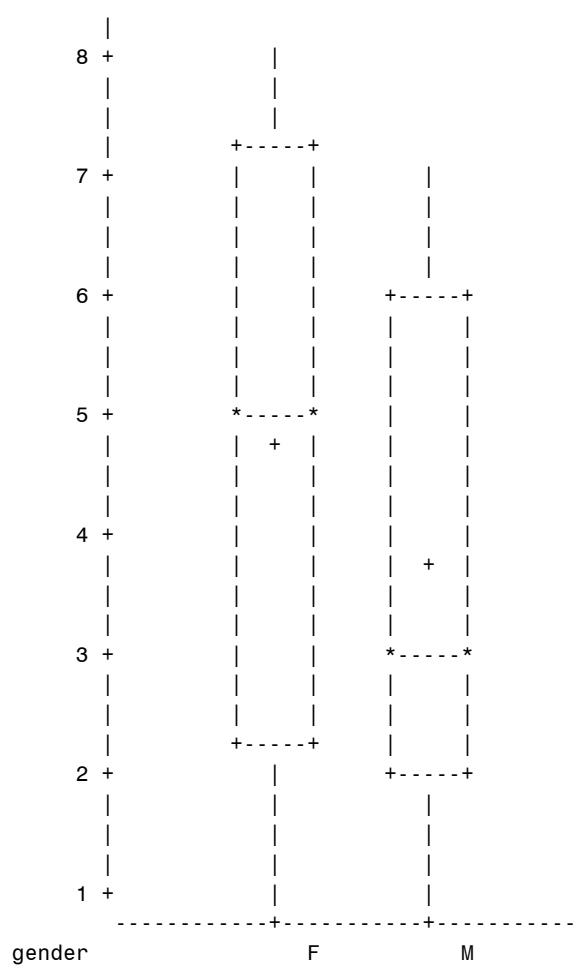
Missing Values

Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	1	12.50	100.00

Stem	Leaf	#	Boxplot
7	0	1	
6	0	1	
5	0	1	
4			
3	0	1	
2	05	2	
1	0	1	

Normal Probability Plot




Schematic Plots
8
7
6
5
4
3
2
1
gender
F
M

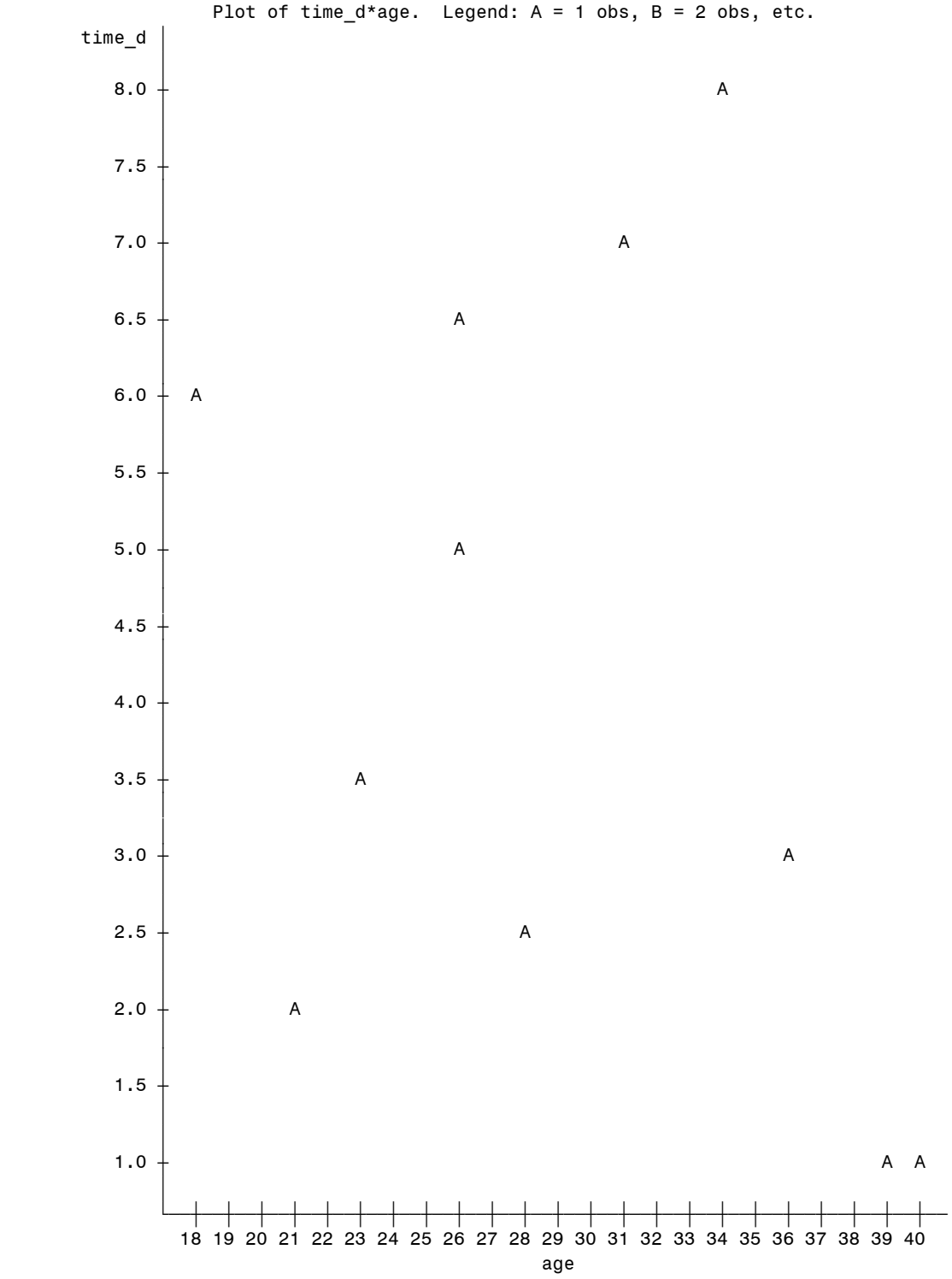
4. PROC PLOT

The following program plots the variable TIME_D against the variable AGE.

SAS Program

```
proc plot data=tv;
plot time_d*age;
run;
```

SAS Output



NOTE: 1 obs had missing values.

5. PROC CHART

The following program plots a vertical bar chart for the variable TVCHAN by GENDER.

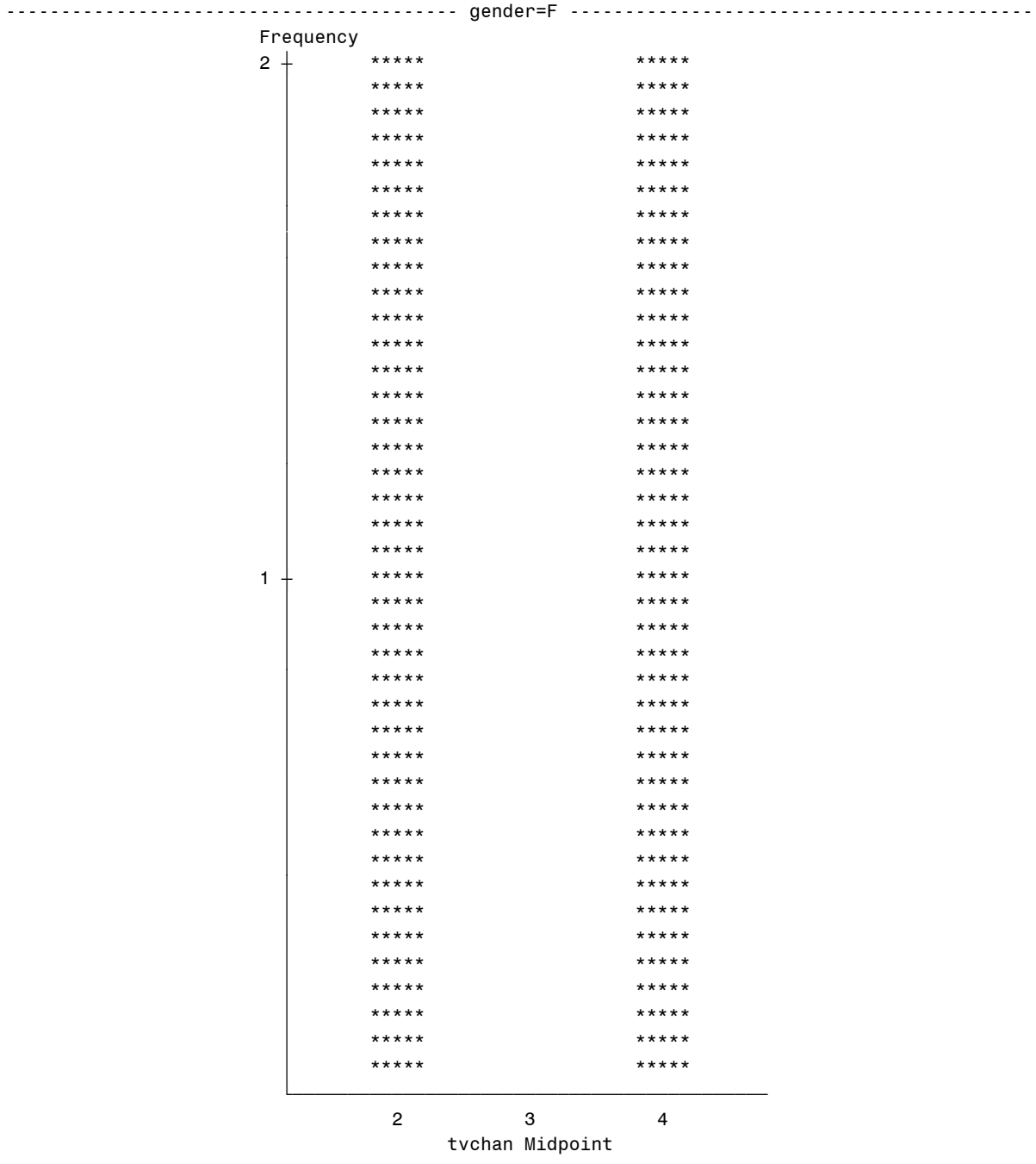
SAS Program

```
proc sort data=tv; by gender;
proc chart data=tv;
vbar tvchan; by gender;
run;
```

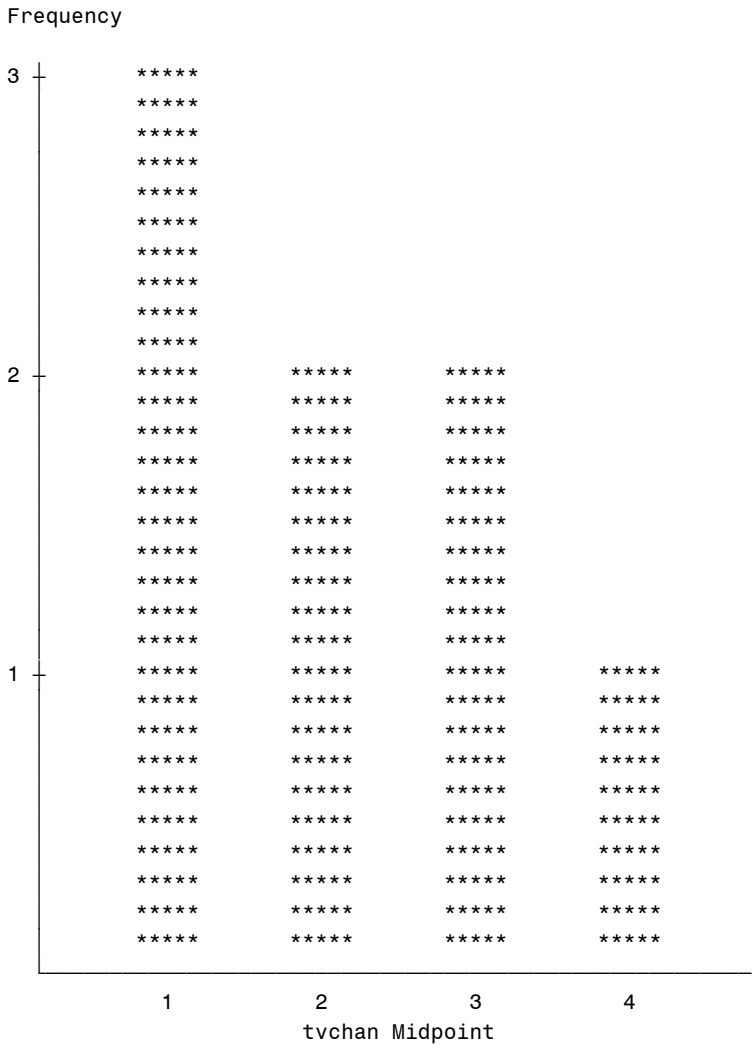
Note on the SAS Program

VBAR plots a vertical bar chart of the variable.
HBAR plots a horizontal bar chart of the variable.
PIE requests a pie chart for each variable listed.

SAS Output



```
---------------------------------------- gender=M ----------------------------------------
```



Simlar output, but with improved graphical quality, can be obtained by using PROC GPLOT (instead of PROC PLOT) and PROC GCHART (instead of PROC CHART).

Another statement that can be used with many PROCs is the following:

WHERE condition;

The WHERE clause enables you to restrict the data that you analyze by specifying a condition that each observation in your data set must satisfy, i.e. it can be used to select a subgroup in the data set.

Example

The following program plots a vertical bar chart for the variable TVCHAN for GENDER=M (produces output similar to the above – heading excluded):

SAS Program

```
proc chart data=tv;
vbar tvchan;
where gender='M';
run;
```

PART 4: PROC TTEST

PROC TTEST Syntax

PROC TTEST DATA= SAS-data-set <options>;
CLASS variable; /* required */
VAR variables;

The following options can appear in the PROC TTEST statement.

ALPHA=*p*
specifies that confidence intervals are to be 100(1-*p*)% confidence intervals, where 0<*p*<1. By default, PROC TTEST uses ALPHA=0.05. If *p* is 0 or less, or 1 or more, an error message is printed.

CI=EQUAL
specifies whether a confidence interval is displayed for σ and, if so, what kind. The CI=EQUAL option specifies an equal tailed confidence interval, and it is the default.

H0=*m*
requests tests against *m* instead of 0 in all three situations (one-sample, two-sample, and paired observation *t* tests). By default, PROC TTEST uses H0=0.

CLASS variable;
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.

VAR variables;
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.

1. HYPOTHESIS TESTING: ONE SAMPLE CASE

Testing H_0: $\mu = \mu_0$

Example:

A firm's board of directors has to decide whether newly appointed representatives would take the traditional course in sales techniques or, would instead, change to a new course offered by a consultant. Suppose the first-year sales figures of ten representatives selected at random, who completed the new course are as follows:

R287900 R419400 R338300 R287500 R310850
R292600 R390050 R369850 R430400 R338450

The mean first year sales of representatives who took the traditional course is R300000.

SAS Program

```
data ex_12_9;
input sales @@;
cards;
287900 419400 338300 287500 310850
292600 390050 369850 430400 338450
;
proc ttest data= ex_12_9 H0=300000 alpha=0.05 ci=equal;
var sales;
 run;
```

SAS Output

```
                    The TTEST Procedure
                     Variable:  sales

    N        Mean     Std Dev     Std Err     Minimum     Maximum
   10      346530     53767.6     17002.8      287500      430400

      Mean        95% CL Mean        Std Dev       95% CL Std Dev
    346530      308067   384993     53767.6     36983.2  98158.6

                    DF    t Value    Pr > |t|
                     9       2.74      0.0230
```

PROC TTEST Step	Comment
DATA EX_12_9;	The keyword DATA starts the data step and indicates that a SAS data set with the name EX_12_9 is to be created.
INPUT SALES @@;	INPUT gives names to all the variables in the data set. @@ indicates that more than one respondent's data will be entered into a line.
PROC TTEST; or PROC TTEST DATA=EX_12_9; PROC TTEST DATA=EX_12_9 H0=300000 ALPHA=0.05 CI=EQUAL;	The last data set is used to calculate descriptive statistics unless the name of the data set is specified in the DATA option of PROC TTEST. The mean, a 95% confidence interval for the mean, standard deviation, a 95% confidence interval for the standard deviation and the standard error of the estimate are given. The T-test test H0: μ=300000 at the 5% level. It gives DF, the degrees of freedom for T, the test statistic value, i.e. the *t*-value and PR > |t| the *p*-value for the *t*-test.
VAR SALES;	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.

Hypothesis test from SAS Output

$H_0 : \mu = 300000$ $H_1 : \mu > 300000$

Use $\alpha = 0.05$. Reject H_0 if p-value < 0.05.

Since p-value (Prob>|T|) = 0.0230/2 = 0.0115 H_0 is rejected.

∴The mean first-year sales of the ten representatives who attended the new course is significantly higher than R300 000.

Note: The p-value (Prob>|T|) for a two-sided hypothesis is given on the output. In the case of a one sided hypothesis the Prob>|T| value must be divided by 2.

2. HYPOTHESIS TESTING: TWO SAMPLE CASE

2.1. Testing H_0: $\mu_1 - \mu_2 = 0$ (Two Independent Samples from Normal Populations)

Example: Applied Statistics and the SAS Programming Language, Cody R.P. and Smith J.K.
Students are randomly assigned to a control or treatment group (where a drug is administered). Their response time to a stimulus is then measured. The times are as follows:

Control	Treatment
(response time in millisec)	
80	100
93	103
83	104
89	99
98	102

Do the treatment scores come from a popoulation whose mean is different from the mean of the population from which the control scores were drawn?

Solution

SAS Program

```
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;
 class group;
 var time;
 run;
```

SAS Output

The TTEST Procedure
Variable: time

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
c	5	88.6000	7.3007	3.2650	80.0000	98.0000
t	5	101.6	2.0736	0.9274	99.0000	104.0
Diff (1-2)		-13.0000	5.3666	3.3941		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
c		88.6000	79.5350	97.6650	7.3007	4.3741	20.9789
t		101.6	99.0252	104.2	2.0736	1.2424	5.9587
Diff (1-2)	Pooled	-13.0000	-20.8268	-5.1732	5.3666	3.6249	10.2811
Diff (1-2)	Satterthwaite	-13.0000	-21.9317	-4.0683			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	8	-3.83	0.0050
Satterthwaite	Unequal	4.6412	-3.83	0.0141

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	4	4	12.40	0.0318

Note:

The information in the bottom line of the output above is used to test the hypothesis of equal variances. If the *p*-value is small (say the Prob>F' value is less than 0.05) then the null hypothesis of equal variances is rejected. The *t*-value and *p*-value for unequal variances are used. If the Prob>F' value is greater than 0.05 the *t*-value and *p*-value for equal variances are used. In this example Prob>F' = 0.0318. The two samples come from populations with variances that differ significantly.

PROC TTEST Step	**Comment**
PROC TTEST; or PROC TTEST DATA=RESPONSE;	PROC TTEST computes a *t* statistic for testing the hypothesis that the means of two groups of observations in a SAS data set are equal. The last data set is used unless the name of the data set is specified in the DATA option of the PROC TTEST statement.
CLASS GROUP;	This statement identifies the independent variable; the variable that identifies the two groups of subjects.
VAR TIME;	Identifies the dependent variable. When more than one dependent variable is listed, a separate *t*-test is computed for each dependent variable in the list.

Hypothesis test from SAS Output

$H_0 : \sigma^2_{\text{control}} = \sigma^2_{\text{treatment}}$

$H_1 : \sigma^2_{\text{control}} \neq \sigma^2_{\text{treatment}}$

Use $\alpha = 0.05$. Reject H_0 if *p*-value < 0.05.

Since *p*-value (Prob>F') = 0.0318 H_0 is rejected.

$\therefore$ Population variances differ significantly.

$H_0 : \mu_{\text{control}} = \mu_{\text{treatment}}$

$H_1 : \mu_{\text{control}} \neq \mu_{\text{treatment}}$

Use $\alpha = 0.05$. Reject H_0 if *p*-value < 0.05.

Since *p*-value (Prob>|T|) = 0.0145 H_0 is rejected.

$\therefore$ The average response times for the two groups differ significantly.

Note:

The *p*-value for a two-sided hypothesis is given in the output. In the case of a one sided hypothesis the Prob>|T| value must be divided by 2.

Example:

Two risk factors that have a bearing on the condition of the heart are fitness and cholesterol level. In a research project on this subject the amounts of triglycerides (unsaturated fats) in the blood samples of nine coronary patients and 21 marathon athletes were measured. The observations in millimol per litre are as follows:

Coronary patients:	3.80	2.71	1.60	1.62	1.93	1.32	1.09	2.28	0.65
Marathon athletes:	0.86	0.84	1.15	1.12	0.72	1.62	1.23	1.22	1.13
	0.98	0.62	0.38	0.86	1.25	0.90	0.56	0.66	0.73
	0.73	0.50	0.92						

Test the hypothesis that, compared to marathon athletes, coronary patients have a significantly higher population mean triglyceride level.

SAS Program

```
data response;
input group$ triglyc @@;
cards;
c 3.80  c 2.71  c 1.60  c 1.62  c 1.93  c 1.32  c 1.09  c 2.28  c 0.65
m 0.86  m 0.84  m 1.15  m 1.12  m 0.72  m 1.62  m 1.23  m 1.22  m 1.13
m 0.98  m 0.62  m 0.38  m 0.86  m 1.25  m 0.90  m 0.56  m 0.66  m 0.73
m 0.73  m 0.50  m 0.92
;
proc ttest;
class group;
var triglyc;
run;
```

SAS Output

The TTEST Procedure

Variable: triglyc

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
c	9	1.8889	0.9443	0.3148	0.6500	3.8000
m	21	0.9038	0.3001	0.0655	0.3800	1.6200
Diff (1-2)		0.9851	0.5649	0.2251		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
c		1.8889	1.1630	2.6147	0.9443	0.6378	1.8091
m		0.9038	0.7672	1.0404	0.3001	0.2296	0.4334
Diff (1-2)	Pooled	0.9851	0.5241	1.4461	0.5649	0.4483	0.7640
Diff (1-2)	Satterthwaite	0.9851	0.2539	1.7162			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	4.38	0.0002
Satterthwaite	Unequal	8.7011	3.06	0.0140

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	20	9.90	<.0001

Hypothesis test from SAS Output

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Use $\alpha = 0.05$. Reject H_0 if p-value < 0.05.

Since p-value (Prob>F') = 0.0000 H_0 is rejected.

$\therefore$ Population variances differ significantly.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Use $\alpha = 0.01$. Reject H_0 if p-value < 0.01.

Since p-value (Prob>|T|) = 0.0141/2 = 0.00705 H_0 is rejected.

$\therefore$ The mean amount of triglycerides for coronary patients is significantly higher than for marathon athletes.

PART 5: PROC REG

PROC REG: Syntax

PROC REG DATA= SAS-data-set;
MODEL dependents = independent (s)/ options;
< label: > TEST equation,<, ...,equation> < / option > ;

DATA= SAS-data-set
The DATA= option names the SAS data set containing the data to be analyzed. If the DATA= option is omitted, the most recently created SAS data set is used .

MODEL dependents= independents / options;
The MODEL statement names the dependent variables and independent effects.
If no independent effects are specified, only an intercept term is fit. Many options can be specified in the MODEL statement after a slash (/), e.g.:
ALPHA= p
XPX displays sums-of-squares and crossproducts matrix (totals required in calculation of parameters)
NOINT fits a model without the intercept term
CLB computes100(1-α)% confidence limits for the parameter estimates
CLI computes100(1-α)% confidence limits for an individual predicted value
CLM computes100(1-α)% confidence limits for the expected value of the dependent variable
P computes predicted values
R produces analysis of residuals

< label: > TEST equation,<, ...,equation> < / option > ;

e.g.

```
model y=a1 a2 b1 b2;
   aplus: test a1+a2=1;
   b1:    test b1=0, b2=0;
   b2:    test b1, b2;
```

The TEST statement tests hypotheses about the parameters estimated in the preceding MODEL statement. Each equation specifies a linear hypothesis to be tested. The rows of the hypothesis are separated by commas.

Variable names must correspond to regressors, and each variable name represents the coefficient of the corresponding variable in the model. An optional label is useful to identify each test with a name. The keyword INTERCEPT can be used instead of a variable name to refer to the model's intercept.

The REG procedure performs an F test for the joint hypotheses specified in a single TEST statement. More than one TEST statement can accompany a MODEL statement. The relation between the t-test and F-test is t^2=F.

1. Fitting a simple regression line, $\hat{y} = \beta_0 + \beta_1 x$. Obtaining confidence intervals for the parameters. Testing H_0: $\beta_0 = 0$ and H_0: $\beta_1 = 0$.

Example:

It is common knowledge that it pays to advertise. An investigation is undertaken to determine the effect of advertisement on sales in pharmacies. Twelve pharmacies are selected at random and both the advertising expenditure and the business profit are determined. The data is reflected below:

Pharmacy	Advertising Expenditure (R), *x*	Business Profit (R), y
1	1581.61	6302.52
2	2292.80	6901.88
3	2595.47	7773.93
4	1062.19	4527.45
5	1930.86	5720.73
6	655.12	4660.12
7	1294.00	6507.83
8	1232.03	5348.99
9	2096.78	7578.83
10	1381.22	5125.67
11	1104.78	5258.17
12	403.62	4445.21

Fit a simple regression line to the data. Test whether the intercept is significantly different from zero. Test whether there is a significant relationship between advertising expenditure and business profit.

Solution

SAS Program

```
data d1;
input x y;
cards;
1581.61     6302.52
2292.80     6901.88
2595.47     7773.93
1062.19     4527.45
1930.86     5720.73
655.12      4660.12
1294.00     6507.83
1232.03     5348.99
2096.78     7578.83
1381.22     5125.67
1104.78     5258.17
403.62      4445.21
;
proc reg data=d1;
model y=x/ xpx clb;
run;
```

SAS Output

The REG Procedure
Model: MODEL1

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	x	y
Intercept	12	17630.48	70151.33
x	17630.48	30660573.634	110463001.84
y	70151.33	110463001.84	424815913.32

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11497813	11497813	35.74	0.0001
Error	10	3217342	321734		
Corrected Total	11	14715155			

Root MSE	567.21617	R-Square	0.7814
Dependent Mean	5845.94417	Adj R-Sq	0.7595
Coeff Var	9.70273		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	3561.97897	415.66847	8.57	<.0001	2635.81190	4488.14605
x	1	1.55456	0.26004	5.98	0.0001	0.97514	2.13397

2. Fitting a simple regression line, $\hat{y} = \beta_0 + \beta_1 x$.
Testing H_0: $\beta_0 = \beta_{00}$ and H_0: $\beta_1 = \beta_{10}$.

Example:

It is claimed that a certain device will offer a saving of more than 15% on fuel consumption of motor cars. Twelve motor cars of different makes and piston capacity are chosen. The distance (in km) travelled per litre of petrol, with and without the device, is given in the table below:

Vehicle	With device, *y*	Without device, *x*
1	9.11	7.63
2	8.24	7.20
3	13.47	11.05
4	13.43	10.95
5	9.98	8.45
6	14.50	11.25
7	9.10	7.60
8	12.67	10.68
9	10.96	9.07
10	11.74	9.46
11	12.74	9.69
12	14.51	13.11

Fit a simple regression line to the data. Test whether the new device offers a saving of more than 15%, i.e. test whether $\frac{y}{x} \geq 1.15$.

Note below that $t_b^2 = 0.226^2 = 0.05 = F$ and p-value $= P(T \geq t) = 0.4128 = \frac{1}{2} P(F \geq f) = \frac{1}{2} \times 0.8256$.

Solution

SAS Program

```
data  d2;
input y x;
cards;
 9.11  7.63
 8.24  7.20
13.47 11.05
13.43 10.95
 9.98  8.45
14.50 11.25
 9.10  7.60
12.67 10.68
10.96  9.07
11.74  9.46
12.74  9.69
14.51 13.11
;
proc reg data=d2;
model y=x;
b: test x=1.15;
run;
```

SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 12
Number of Observations Used 12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	48.72848	48.72848	118.76	<.0001
Error	10	4.10301	0.41030		
Corrected Total	11	52.83149			

Root MSE	0.64055	R-Square	0.9223
Dependent Mean	11.70417	Adj R-Sq	0.9146
Coeff Var	5.47282		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.33822	1.05922	0.32	0.7561
x	1	1.17437	0.10776	10.90	<.0001

Model: MODEL1

Test b Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.02098	0.05	0.8256
Denominator	10	0.41030		

3. Calculating:
(a)a confidence interval for the mean value of the dependent variable and
(b) aprediction interval
for the dependent variable for a particular value of the explanatory variable.

Example:

In an investigation into the effect of the landing speed on the number of landings made with a set of main landing tires on the Boeing 737, the following data was obtained:

Observation	Number of landings, *y*	Landing speed, *x*
1	84	112
2	85	114
3	84	116
4	80	118
5	81	120
6	76	122
7	78	124
8	71	126
9	72	128
10	68	130
11	69	132
12	64	134

(a) Fit a simple regression line to the data and determine a 95% confidence band for the population regression line.
(b) Also determine a 95% prediction band.

Solution
SAS Program

```
data boeing;
input y x;
cards;
84    112
85    114
84    116
80    118
81    120
76    122
78    124
71    126
72    128
68    130
69    132
64    134
;
Proc reg data=boeing;
model y=x / clm cli;
run;
```

SAS Output

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read 12
Number of Observations Used 12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	517.37063	517.37063	149.40	<.0001
Error	10	34.62937	3.46294		
Corrected Total	11	552.00000			

Root MSE	1.86090	R-Square	0.9373
Dependent Mean	76.00000	Adj R-Sq	0.9310
Coeff Var	2.44855		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	192.97902	9.58545	20.13	<.0001
x	1	-0.95105	0.07781	-12.22	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	84.0000	86.4615	1.0105	84.2100	88.7131	81.7433	91.1798	-2.4615
2	85.0000	84.5594	0.8826	82.5929	86.5260	79.9704	89.1485	0.4406
3	84.0000	82.6573	0.7650	80.9528	84.3619	78.1743	87.1404	1.3427
4	80.0000	80.7552	0.6633	79.2774	82.2331	76.3534	85.1571	-0.7552
5	81.0000	78.8531	0.5857	77.5481	80.1582	74.5063	83.2000	2.1469
6	76.0000	76.9510	0.5428	75.7416	78.1605	72.6319	81.2702	-0.9510
7	78.0000	75.0490	0.5428	73.8395	76.2584	70.7298	79.3681	2.9510
8	71.0000	73.1469	0.5857	71.8418	74.4519	68.8000	77.4937	-2.1469
9	72.0000	71.2448	0.6633	69.7669	72.7226	66.8429	75.6466	0.7552
10	68.0000	69.3427	0.7650	67.6381	71.0472	64.8596	73.8257	-1.3427
11	69.0000	67.4406	0.8826	65.4740	69.4071	62.8515	72.0296	1.5594
12	64.0000	65.5385	1.0105	63.2869	67.7900	60.8202	70.2567	-1.5385

Sum of Residuals	0
Sum of Squared Residuals	34.62937
Predicted Residual SS (PRESS)	50.87705

PART 6: PROC CORR

PROC CORR: Syntax

```
PROC CORR <option-list>;
    VAR variable-list;
    WITH variable-list;
    BY variable-list;
```

The CORR procedure computes Pearson correlation coefficients.

The FISHER options request that the Fisher's *z* transformation be used to derive confidence limits and a *p*-value under a specified null hypothesis $H_0: \rho = \rho_0$. Either a one-sided or a two-sided alternative is used for these statistics. The following FISHER-*options* are available:

ALPHA= α
specifies the level of the confidence limits for the correlation, $100(1-\alpha)\%$. The value of the ALPHA= option must be between 0 and 1, and the default is ALPHA=0.05.

BIASADJ= YES | NO
specifies whether or not the bias adjustment is used in constructing confidence limits. The BIASADJ=YES option also produces a new correlation estimate using the bias adjustment. By default, BIASADJ=YES.

RHO0= ρ_0
specifies the value ρ_0 in the null hypothesis $H_0: \rho = \rho_0$, where $-1 < \rho_0 < 1$. By default, RHO0=0.

TYPE= LOWER | UPPER | TWOSIDED
specifies the type of confidence limits. The TYPE=LOWER option requests a lower confidence limit from the lower alternative $H_1: \rho < \rho_0$, the TYPE=UPPER option requests an upper confidence limit from the upper alternative $H_1: \rho > \rho_0$, and the default TYPE=TWOSIDED option requests two-sided confidence limits from the two-sided alternative $H_1: \rho \neq \rho_0$.

The BY statement specifies groups in which separate correlation analyses are performed.

The VAR statement lists the numeric variables to be analyzed and their order in the correlation matrix. If you omit the VAR statement, all numeric variables not listed in other statements are used.

The WITH statement lists the numeric variables with which correlations are to be computed

1. **Calculating the correlation between two variables *X* and *Y*.**
   Testing H_0: ρ = 0.

Example 11.8: Wackerley, Mendenhall &Scheaffer

Table 11.3 Data for Example 11.8

Student	Mathematics Achievement Test Score	Final Calculus Grade
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75

Are the achievement scores and calculus grades independent? Use α=0.05. Identify the attained significance level.

SAS Program

```
data d1;
input student score grade;
cards;
 1 39 65
 2 43 78
 3 21 52
 4 64 82
 5 57 92
 6 47 89
 7 28 73
 8 75 98
 9 34 56
10 52 75
;
proc corr data=d1;
var score;
with grade;
run;
```

SAS Output

```
                    The CORR Procedure

                1 With Variables:    grade
                1      Variables:    score

                    Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

grade             10      76.00000      15.11438     760.00000      52.00000      98.00000
score             10      46.00000      16.57977     460.00000      21.00000      75.00000

          Pearson Correlation Coefficients, N = 10
                Prob > |r| under H0: Rho=0

                                 score

                grade         0.83979
                               0.0024
```

Hypothesis test from SAS Output

$H_0 : \rho = 0 \quad H_1 : \rho \neq 0$

Use $\alpha = 0.05$. Reject H_0 if p-value < 0.05.

Since p-value (Prob>|r|) = 0.0024/2 = 0.0021, H_0 is not rejected.

$\therefore$The population correlation coefficient differs significantly from 0.

Note: The p-value (Prob>|r|) for a two-sided hypothesis is given on the output. In the case of a one sided hypothesis the Prob>|r| value must be divided by 2.

2. Calculating the correlation between two variables *X* and *Y*. Obtaining confidence intervals for ρ. Testing H_0: ρ = ρ_0.

Example 11.8: Wackerley, Mendenhall &Scheaffer

Is there reason to believe that the correlation between achievement scores and calculus grade is greater than 0.80? Use α=0.10. Identify attained significance level.

SAS Program

```
proc corr data=d1 fisher (rho0=0.80 type=upper biasadj=no alpha=0.10);
var score;
with grade;
run;
```

SAS Output

The CORR Procedure

1 With Variables: grade
1 Variables: score

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Upper 90% CL	-----H0:Rho>=Rho0----- Rho0	p Value
score	grade	10	0.83979	1.22045	0.936010	0.80000	0.5811

Hypothesis test from SAS Output

$H_0 : \rho = 0.80 \qquad H_1 : \rho > 0.80$

Use $\alpha = 0.10$. Reject H_0 if p-value < 0.10.

Since p-value (Prob>z) = 0.5811, H_0 is not rejected.

$\therefore$The population correlation coefficient is not significantly larger than 0.80.

Note: The p-value (Prob>z) for a one-sided hypothesis is given on the output by specifying TYPE=UPPER.

Confidence Interval from SAS Output

90% Confidence Interval for ρ: (-∞,0.936)

Since 0.80 in 90% confidence interval for ρ, H_0 is not rejected.

PART 7: PROC MEANS and PROC TTEST(continued)

HYPOTHESIS TESTING: TWO DEPENDENT SAMPLES

Testing H_0: $\mu_1 - \mu_2 = 0$ (Two Dependent Samples from Normal Populations)

The following two examples illustrate how either PROC MEANS (by calculating differences) or PROC TTEST (using the PAIRED statement) can be used to test the hypothesis above.

Example: Applied Statistics and the SAS Programming Language, Cody R.P. and Smith J.K.

The response times to a stimulus of six students are each measured in the absence of a drug (control value) and after having received the drug (treatment value). The times are as follows:

Subject	Control	Treatment
	(response time in millisec)	
1	90	95
2	87	92
3	100	104
4	80	89
5	95	101
6	90	105

Do the average response times for the students differ between the control and treatment?

Solution

SAS Program

```
data paired;
input ctime ttime @@;
diff=ttime-ctime;
cards;
90 95  87  92  100 104
80 89  95 101   90 105
;
proc means n mean stderr t prt;
var diff;
 run;
```

SAS Output

```
Analysis Variable : DIFF
N          Mean     Std Error             T  Prob>|T|
-----------------------------------------------------
6     7.3333333     1.6865481     4.3481318    0.0074
-----------------------------------------------------
```

Hypothesis test from SAS Output

$H_0 : \mu_{\text{control}} = \mu_{\text{treatment}}$

$H_1 : \mu_{\text{control}} \neq \mu_{\text{treatment}}$

Reject H_0 if p-value < 0.05.

Since p-value (Prob>|T|) = 0.0074 H_0 is rejected.

$\therefore$ The average response times for the control and drug treatment differ significantly.

Note:

The p-value for a two-sided hypothesis is given in the output. In the case of a one sided hypothesis the Prob>|T| value must be divided by 2.

Example:

The time (in minutes) it takes operators to fit a certain part before and after completing a training program appears in the table below. Determine whether the training program significantly decreased the mean fitting time.

Time required by eight operators before and after training

Operator	Before training	After training
1	23	17
2	17	14
3	16	12
4	15	13
5	19	12
6	21	20
7	13	14
8	20	15

Solution

SAS Program

```
data d1;
input before after @@;
cards;
23 17  17 14  16 12  15 13
19 12  21 20  13 14  20 15
;
proc ttest;
paired before*after;
run;
```

SAS Output

The TTEST Procedure

Statistics

Difference	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
before - after	8	1.1434	3.375	5.6066	1.7649	2.6693	5.4327	0.9437	-1	7

T-Tests

Difference	DF	t Value	Pr > |t|
before - after	7	3.58	0.0090

Hypothesis test from SAS Output

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Use $\alpha = 0.01$. Reject H_0 if p-value < 0.01.

Since p-value (Prob>|T|) = 0.0090/2 = 0.0045 H_0 is rejected.

$\therefore$The training program significantly decreased the mean fitting time.

PART 8: PROC GLM

ANALYSIS OF VARIANCE

PROC GLM: Syntax
PROC GLM options ;
CLASS variable-list;
MODEL dependents= independents / options; /* required */
MEANS effects / options;

CLASS variable-list;
The CLASS statement names the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement.
Classification variables can be either character or numeric. Only the first sixteen characters of a character variable are used.

MODEL dependents= independents / options;
The MODEL statement names the dependent variables and independent effects.
If no independent effects are specified, only an intercept term is fit. Many options can be specified in the MODEL statement after a slash (/), e.g.: ALPHA= p

MEANS effects / options;
For any effect that appears on the right-hand side of the model and that does not contain any continuous variables, GLM can compute means of all continuous variables in the model.
You can use any number of MEANS statements, provided they appear after the MODEL statement.
These options can appear in the MEANS statement after a slash (/):

1. One-way Analysis of Variance (One-way ANOVA)

Example:

In order to compare four different textbooks covering the same topic, students are divided randomly into four classes and subsequently a different textbook is assigned to each of the four classes. The table below gives the final achievement of the 23 students who wrote the examination in the course. Test whether the mean achievement for the four groups differ significantly.

Textbook	Percentage
1	72 65 81 83 67 73
2	48 63 71 55 52 49 54
3	72 76 82 73 69
4	65 68 75 61 92

Solution

SAS Program

```
data d1;
input textbook obs @@;
cards;
1 72 1 65 1 81 1 83 1 67 1 73
2 48 2 63 2 71 2 55 2 52 2 49 2 54
3 72 3 76 3 82 3 73 3 69
4 65 4 68 4 75 4 61 4 92
;
proc glm;
class textbook;
model obs=textbook;
means textbook; run;
```

SAS Output

```
                          Class Level Information
                         Class    Levels    Values
                         TEXTBOOK      4    1 2 3 4
                    Number of observations in data set = 23
                        General Linear Models Procedure
Dependent Variable: OBS
Source                   DF         Sum of Squares          Mean Square   F Value     Pr > F
Model                     3          1482.32608696          494.10869565      6.89     0.0025
Error                    19          1363.50000000           71.76315789
Corrected Total          22          2845.82608696

                  R-Square                  C.V.              Root MSE             OBS Mean
                  0.520877              12.44190            8.47131382           68.08695652

Source                   DF             Type I SS          Mean Square   F Value     Pr > F
TEXTBOOK                  3          1482.32608696          494.10869565      6.89     0.0025

Source                   DF           Type III SS          Mean Square   F Value     Pr > F
TEXTBOOK                  3          1482.32608696          494.10869565      6.89     0.0025

   Level of           -------------obs-------------
   textbook     N             Mean          Std Dev

   1            6       73.5000000        7.2594766
   2            7       56.0000000        8.2462113
   3            5       74.4000000        4.9295030
   4            5       72.2000000       12.1942609
```

Hypothesis test from SAS Output

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : One of the μ's differ

Use $\alpha = 0.01$. Reject H_0 if p-value < 0.01.

Since p-value (Pr > F) = 0.0025 H_0 is rejected.

$\therefore$ The mean achievements differ significantly.

2. Two-way Analysis of Variance (Two-way ANOVA)

Example:

Suppose we wish to compare the average valuation of houses in the three areas of the Witwatersrand, namely the Johannesburg area, the East Rand and the West Rand. In order to ensure that differences in valuations are not due to differences in house size, we shall use a randomized block design. Three houses of comparable size are repeatedly selected randomly from the three areas and the valuations are noted. The table below gives the data:

	Block										
Area	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Jhb	320	160	195	165	300	315	120	170	120	300	210
East	430	105	170	160	206	190	165	215	159	150	320
West	515	185	95	185	220	160	205	240	305	240	235

Solution

SAS Program

```
data d1;
input treat$ block f @@;
cards;
j 1 320 j 2 160 j 3 195 j 4 165 j 5 300 j 6 315 j 7 120 j 8 170 j 9 120 j 10 300 j 11 210
e 1 430 e 2 105 e 3 170 e 4 160 e 5 206 e 6 190 e 7 165 e 8 215 e 9 159 e 10 150 e 11 320
w 1 515 w 2 185 w 3  95 w 4 185 w 5 220 w 6 160 w 7 205 w 8 240 w 9 305 w 10 240 w 11 235
;
proc glm;
class treat block;
model f=treat block;
run;
```

SAS Output

```
                          General Linear Models Procedure
                              Class Level Information
                    Class    Levels    Values
                    TREAT         3    e j w
                    BLOCK        11    1 2 3 4 5 6 7 8 9 10 11
                          Number of observations in data set = 33


                          General Linear Models Procedure
Dependent Variable: F
Source                   DF         Sum of Squares           Mean Square   F Value     Pr > F
Model                    12        179594.00000000        14966.16666667      3.50     0.0065
Error                    20         85420.72727273         4271.03636364
Corrected Total          32        265014.72727273

                  R-Square                  C.V.              Root MSE                F Mean
                  0.677676              29.82925           65.35316644         219.09090909

Source                   DF           Type III SS           Mean Square   F Value     Pr > F
TREAT                     2          4677.27272727         2338.63636364      0.55     0.5868
BLOCK                    10        174916.72727273        17491.67272727      4.10     0.0035
```

Hypothesis test from SAS Output

(a)

H_0 : The valuations are the same for the different areas.

H_1 : The valuations differ.

Use $\alpha = 0.05$. Reject H_0 if p-value < 0.05.

Since p-value (Pr > F) = 0.5868 H_0 is not rejected.

$\therefore$ The mean building valuations of houses of comparable floor area in the Johannesburg area, the East Rand and the West Rand do not differ significantly from one another.

(b)

H_0 : The block effects are the same.

H_1 : The block effects differ.

Use $\alpha = 0.05$. Reject H_0 if p-value < 0.05.

Since p-value (Pr > F) = 0.0035 H_0 is rejected.

$\therefore$ The inclusion of blocks was justified and led to a significant decrease in the mean error sum of squares.

PART 9: PROC FREQ (continued)

PROC FREQ: Syntax

PROC FREQ options;
OUTPUT <OUT= SAS-data-set><output-statistic-list>;
TABLES requests / options;
WEIGHT variable;
BY variable-list;

PROC FREQ options;

ORDER= specifies the order in which the values of the frequency and crosstabulation table variables are to be reported. The ORDER= DATA option orders values according to their order in the input data set.

TABLES requests / options;

The TABLES command requests tables be produced. Any number of TABLES statements can be included. If no TABLES statement is given, one-way frequencies for all of the variables in the data set are produced. To request a one-way frequency table for a variable, name the variable in a TABLES statement. For example: PROC FREQ; TABLES a;

For a crosstabulation table of two variables, give their names separated by an asterisk. The first variable's values form the rows of the table, and the second variable's values form the columns. For example: PROC FREQ; TABLES a*b;

For n-way crosstabulation tables, the last variable's values form the columns; the next-to-last variable's values form the rows. Each level (or combination of levels) of the other variables form one stratum. A contingency table is produced for each stratum.

Options that can be used in the TABLES statement:

Request Statistical Analysis		CHISQ	TESTF=	TESTP=	
Statistical Details		ALPHA=			
Request Additional Table Information		CELLCHI2			
		EXPECTED			
Suppress Printing	NOCOL	NOFREQ	NOPERCENT	NOPRINT	NOROW

TABLES requests / CHISQ EXPECTED CELLCHI2;

The CHISQ option requests tests of no association between the row variable and the column variable for each two-way table (or for each stratum in an n-way table). The tests include Pearson chi-square and for 2x2 tables, Fisher's exact test. Also printed are some measures of association based on chi-square. If null hypothesis proportions are given by the TESTP= option, then the CHISQ option performs a chi-square test for specified proportions. If null hypothesis frequencies are given by the TESTF= option, the CHISQ option performs a chi-square test for specified frequencies.

The EXPECTED option requests that the expected cell frequencies under the hypothesis of independence (or homogeneity) be printed. The CELLCHI2 option displays each cell's contribution to the total Pearson chi-square statistic.

1. The Chi-Square test

Example:

The following table was obtained from a table of random numbers.

Distribution of random numbers										
0	1	2	3	4	5	6	7	8	9	**Total**
7	11	7	12	14	12	6	9	8	14	100

Test at a 5% level of significance whether the random numbers do appear in the random table with equal probability.

Solution
SAS Program

```
data d1;
input digit f @@;
cards;
0  7
1 11
2  7
3 12
4 14
5 12
6  6
7  9
8  8
9 14
;
proc freq data=d1;
weight f;
tables digit /nocum testf=(10 10 10 10 10 10 10 10 10 10);
run;
```

SAS Output

```
                The FREQ Procedure
                                 Test
 digit     Frequency     Frequency     Percent
 ------------------------------------------------

     0             7            10        7.00
     1            11            10       11.00
     2             7            10        7.00
     3            12            10       12.00
     4            14            10       14.00
     5            12            10       12.00
     6             6            10        6.00
     7             9            10        9.00
     8             8            10        8.00
     9            14            10       14.00

          Chi-Square Test
        for Specified Frequencies
        -------------------------

        Chi-Square          8.0000
        DF                       9
        Pr > ChiSq          0.5341

             Sample Size = 100
```

2. The (I x J) independence test

Example:

Two-way contingency table of a random sample of 893 houses classified according to age and number of bedrooms

Bedrooms	Built before 1960	Built 1960 – 1969	Built 1970 - 1979	Built 1980 - 1989	Built 1990 or later	Total
2	41	11	5	9	7	73
3	135	70	92	169	167	633
4 and more	38	17	37	48	47	187
Total	214	98	134	226	221	893

Test whether there is a relationship between the age of the residence and the number of bedrooms.

Solution

SAS Program

```
proc format;
value $bb 1='built before 1960'
          2='built 1960-1969'
          3='built 1970-1979'
          4='built 1980-1989'
          5='built 1990 and later';
data ex_15_5;
input built$ bedrooms f @@;
format built bb.;
cards;
1 2 41   2 2 11 3 2 5  4 2 9   5 2 7
1 3 135  2 3 70 3 3 92 4 3 169 5 3 167
1 4 38   2 4 17 3 4 37 4 4 48  5 4 47
;
proc freq;
tables bedrooms*built/chisq expected nocol norow nopercent;
weight f;
run;
```

```
data ex_15_5;

do bedrooms=2 to 4;
  do built=1 to 5;
    input f @@;
    output;
  end;
end;

cards;
41  11 5  9   7
135 70 92 169 167
38  17 37 48  47
;
```

SAS Output

TABLE OF BEDROOMS BY BUILT

BEDROOMS / Frequency / Expected	built before 1960	built 1960-1969	built 1970-1979	built 1980-1989	built 1990 and later	Total
2	41 17.494	11 8.0112	5 10.954	9 18.475	7 18.066	73
3	135 151.69	70 69.467	92 94.985	169 160.2	167 156.66	633
4	38 44.813	17 20.522	37 28.06	48 47.326	47 46.279	187
Total	214	98	134	226	221	893

STATISTICS FOR TABLE OF BEDROOMS BY BUILT

Statistic	DF	Value	Prob
Chi-Square	8	55.184	0.001
Likelihood Ratio Chi-Square	8	50.247	0.001
Mantel-Haenszel Chi-Square	1	18.114	0.001
Phi Coefficient		0.249	
Contingency Coefficient		0.241	
Cramer's V		0.176	

Sample Size = 893

Hypothesis test from SAS Output

H_0 : The age of a residence and the number of bedrooms in the house are independent.

H_1 : The age of a residence does influence the number of bedrooms in the residence.

Use $\alpha = 0.01$. Reject H_0 if p-value < 0.01.

Since p-value (Prob) = 0.001 H_0 is rejected.

$\therefore$ The age of the residence and the number of bedrooms are dependent.