Part I - Fundamental Data Analysis

Chapter 1 - Introduction to SAS

The Statistical Analysis System is available in the computer labs:

- under Windows double click on Statistical Apps icon then on SAS icon
- also available on Hawk executes in batch under Unix.

Basic keystrokes

F1 function key used for help
F5 function key used to move to program area

Enter code

F3 function key used to run code in
F5 area execute that code
F6 function key used to move to log area

Check for problems/errors

F7 function key used to move to output area

Inspect output

F4 function key used to recall previously executed code

LMB click with left mouse button
RMB click with right mouse button to pop-up menus in F5/F6/F7 areas.

Menu

Main menu works on currently selected area determined by F5/F6/F7. Separate menus for each area may appear on screen, click with LMB on entries. It may not appear on screen, use RMB to pop it up then click with LMB.

SAS code: fda01a.sas
F1 for help, F5 go to program area or click on window. If visible enter SAS ® code all at once or in segments.

```
* ...;
options pagesize=53 linesize=76 pageno=1;
data keybrd;
input time;
* ...;
centered=time-47.20;
label time="..." centered="Completion Time - 47.20";
datalines;
42.86
37.56
...
41.63
38.12
run;
```

SAS OUTPUT
keybrd data set

```
time       centered
42.86      -4.34
37.56      -9.64
 ...         ...
41.63      -5.57
38.12      -9.08
```

Data set = 2-dimensional table
- rows are called observations
- columns are called variables

Names are assigned through INPUT command
- numeric type by default through assignment command
- type determined by SAS at most 8 characters (in SAS version 8, longer than 8 characters in later versions). 1st character must be alphabetic.

Labels may be assigned, used in output if assigned. Names used otherwise
Formats may be assigned, multiple values per line

SAS code for defining formats

```
data keybrd;
input time @@;
centered=...;
qrtrtime=time/4;
label ...;
format qrtrtime 4.1;
/* 4 positions: two integer digits, the decimal point,
one decimal digit;
datalines;
42.86 37.56
...
41.63 38.12
run;
```

SAS code for character data

```
data keybrd;
input time keybdtyp $;
label ...;
datalines;
42.86 A
37.56 A
...
41.63 B
38.12 B
run;
```

Data are read in free format separated by 1 or more blanks. Data can be numbers or characters with no embedded blanks.

Fixed format alternative

```
input time 1-7 keybrtyp $ 9;
```

Steps to execute SAS code

1. F5 type in code – open program editor window
2. F3 executes current code
3. Code is then stored in a buffer
4. check for errors in the log window.
   Suppose there are some errors
   a. F5 program area is empty
   b. F4 recall last code segment
   c. be careful to hit only once, fix errors.
5. F3 rerun code
6. check for errors
7. clear window area with EDIT/CLEAR TEXT
8. In our example output area is empty, because code creates no output. The output window will contain all output for all runs until cleared.

Use data in a separate file

Clear program area enter data step code

```
data keybrd;
infile "a:\keybrd.dat";
/* use double quote mark ("),or apostrophe ('),
but not backquote (`)*/
input time @@;
centered=...;
label ...;
run;
```

Save it under a different name with FILE/SAVE AS/..., call it a:\fda01a.sas.

Adding more code

Go to program area, clear it if not empty, read in code from disk file using FILE/OPEN/...
Read in a:\fda01a.sas, assuming only data step coded, add code to end, save it to disk using FILE/SAVE AS/...
Use name a:\fda01a.sas
default name = last one written not necessarily last one read
FILE/SAVE would also work in this case default file name set by prior FILE/OPEN
if no default writes to temporary file possibly deleted after session ends.

File: fd01a.sas

```
* ...;
options nonumber nodate;
data keybrd;
...
run;
proc sort;
by time;
* increasing order;
* by descending time;
* for decreasing order;
```

```
title "Testing for Population Mean Completion Time of 47.20";
* page heading literal;
proc print label;
* column headings are labels;
* variable names otherwise;
var centered time;
* optional statement;
* default: all variables in order created;
```

Proc print output

Testing for Population Mean Completion Time of 47.20

Obs	Completion Time - 47.20	Completion Time
1	-15.73	31.47
2	-11.71	35.49
3	-11.01	36.19
4	-10.84	36.36
5	-9.64	37.56
6	-9.57	37.63
7	-9.30	37.90
8	-9.08	38.12
9	-8.43	38.77
10	-5.57	41.63
11	-5.56	41.64
12	-5.52	41.68
13	-4.41	42.79
14	-4.34	42.86
15	-3.89	43.31
16	-3.42	43.78
17	-0.74	46.46
18	0.54	47.74
19	1.74	48.94
20	3.50	50.70
21	7.79	54.99
22	8.50	55.70
23	10.85	58.05
24	12.73	59.93

Note:

1. The page number will be produced at the end of the title line, starting with page number 1, since the option PAGENO=1 is set at the start of all code given in these lecture notes, but that page number is not reproduced in these notes. The option NONUMBER may be used to not produce page numbers.
2. A line containing the date will be produced just after the title line in the output generated by the code given in these lecture notes, but this date line is not reproduced in these notes. It can be removed from the output by including the NODATE option in the options statement at the beginning of the program. This option is not included in sample code segments since it is usually more appropriate to produce a date.

Simple univariate statistics

```
proc means n nmiss mean std stderr t prt maxdec=3;
   * n: number of observations;
   * nmiss: number missing obs.;
   * mean: average of obs.;
   * std: standard deviation;
   * stderr: standard error = std/(n)1/2 ;
   * t: t for testing if population mean is zero = mean/stderr;
   * can only test for zero, testing centered for zero,same as
     testing time for 47.20;
   * prt: p-value for t test;
   * maxdec=3: round to 3 digits;
var centered;
   * optional statement;
   * default: all numeric variables in order created;
```

proc means output

Testing for Population Mean Completion Time of **47.20**

The MEANS Procedure

Analysis Variable : centered Completion Time - **47.20**

N	Miss	Mean	Std Dev	Std Error	t Value	Pr > t|
24	**0**	**-3.463**	**7.651**	**1.562**	**-2.22**	**0.0368**

Interpretation of output

Prob>|t|: p-value for t test for zero population mean; a two-sided p-value divide by 2 for associated one-sided p-value.
Significant two-sided p-value: 0.0368<0.05
Significant one-sided p-value: 0.0368/2=0.0184<0.05

More extensive univariate statistics:
t test requires large sample size or approximate normality otherwise or in any case use Wilcoxon signed rank test that requires symmetry.

```
proc univariate normal plot;
* normal: test for normality;
* plot: request three plots,
stem and leaf plot,
box (and whiskers) plot,
normal (probability) plot;
var centered;
* same as for proc means;
```

proc univariate output

Testing for Population Mean Completion Time of **47.20**

The UNIVARIATE Procedure

Variable: centered (Completion Time - **47.20**)

Moments

N	**24**	Sum Weights	**24**
Mean	**-3.4629167**	Sum Observations	**-83.11**
Std Deviation	**7.65092463**	Variance	**58.5366476**
Skewness	**0.70412121**	Kurtosis	**-0.3077277**
Uncorrected SS	**1634.1459**	Corrected SS	**1346.3429**
Coeff Variation	**-220.93875**	Std Error Mean	**1.56173845**

Basic Statistical Measures

Location		Variability	
Mean	**-3.46292**	Std Deviation	**7.65092**
Median	**-4.96500**	Variance	**58.53665**
Mode	.	Range	**28.46000**
		Interquartile Range	**10.57500**

Tests for Location: Mu0=**0**

Test	-Statistic-		-----p Value------	
Student's t	t	-2.21735	Pr > |t|	0.0368
Sign	M	-5	Pr >= |M|	0.0639
Signed Rank	S	-72	Pr >= |S|	0.0366

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.932331	Pr < W	0.1099
Kolmogorov-Smirnov	D	0.164429	Pr > D	0.0912
Cramer-von Mises	W-Sq	0.10828	Pr > W-Sq	0.0851
Anderson-Darling	A-Sq	0.650551	Pr > A-Sq	0.0825

Quantiles (Definition 5)

Quantile	Estimate
100% Max	12.730
99%	12.730
95%	10.850
90%	8.500

Output interpretation: p-values

Pr>|T| - t test
Pr>=|M| - sign test
Pr>=|S| - Wilcoxon signed rank test
Pr<W - Shapiro-Wilk test for normality.

proc univariate output (cont.d)

Testing for Population Mean Completion Time of 47.20

The UNIVARIATE Procedure
Variable: centered (Completion Time - 47.20)

```
                    Quantiles (Definition 5)
                     Quantile      Estimate

                     75% Q3           1.140
                     50% Median      -4.965
                     25% Q1          -9.435
                     10%            -11.010
                     5%             -11.710
                     1%             -15.730
                     0% Min         -15.730


                      Extreme Observations
                  -----Lowest----        ----Highest----
                   Value      Obs         Value      Obs

                  -15.73        1          3.50       20
                  -11.71        2          7.79       21
                  -11.01        3          8.50       22
                  -10.84        4         10.85       23
                   -9.64        5         12.73       24
```

proc univariate output (cont.d) – plot option

```
  Stem Leaf                     #             Boxplot
     1 13                       2                |
     0 88                       2                |
     0 124                      3             +-----+
    -0 44431                    5             |  +  |
    -0 998666                   6             *-----*
    -1 21100                    5                |
    -1 6                        1                |
       ----+----+----+----+
   Multiply Stem.Leaf by 10**+1
```

Box (and whiskers) plot

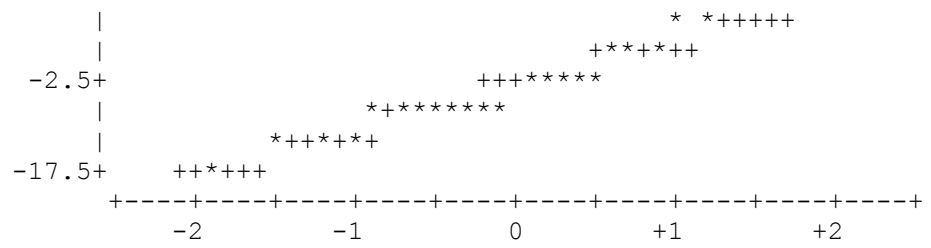
+	Mean
box	25th to 75th quartile, center is median
IQR	Inter-quartile range, equal to the box length. It is a measure of spread similar to the standard deviation
lines	Whiskers
0	Mild outlier
*	Extreme outlier.

proc univariate output (cont.d) – plot option

```
Testing for Population Mean Completion Time of 47.20

                The UNIVARIATE Procedure
     Variable:  centered  (Completion Time - 47.20)

                     Normal Probability Plot
12.5+                                        *  +*++++
```



Normal plot
+ Straight line
* Normal plot
The closer *'s are to +'s, the more reasonable it is to assume that data came from a normal population.

```
fda01a.sas
* Example for one sample testing problem;
options pagesize=53 linesize=76 pageno=1;
data keybrd;
input time;
centered=time-47.20;
label time="Completion Time"
centered="Completion Time - 47.20";
datalines;
42.86
37.56
31.47
46.46
36.36
58.05
35.49
43.31
36.19
42.79
41.64
38.77
54.99
37.63
55.70
43.78
41.68
37.90
50.70
47.74
48.94
59.93
41.63
38.12
run;
proc sort;
by time;
title "Testing for Population Mean Completion Time of 47.20";
proc print label;
var centered time;
proc means n nmiss mean std stderr t prt maxdec=3;
var centered;
```

```
proc univariate normal plot;
var centered;
proc rank normal=blom out=new;
var centered;
ranks normscr;
proc plot formchar="|----|+|---";
plot centered*normscr="*" / vaxis=-20 to 20 by 5;
label normscr="Normal Score";
run;
```