SAS Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery
Chapter 12: Analysis of Variance

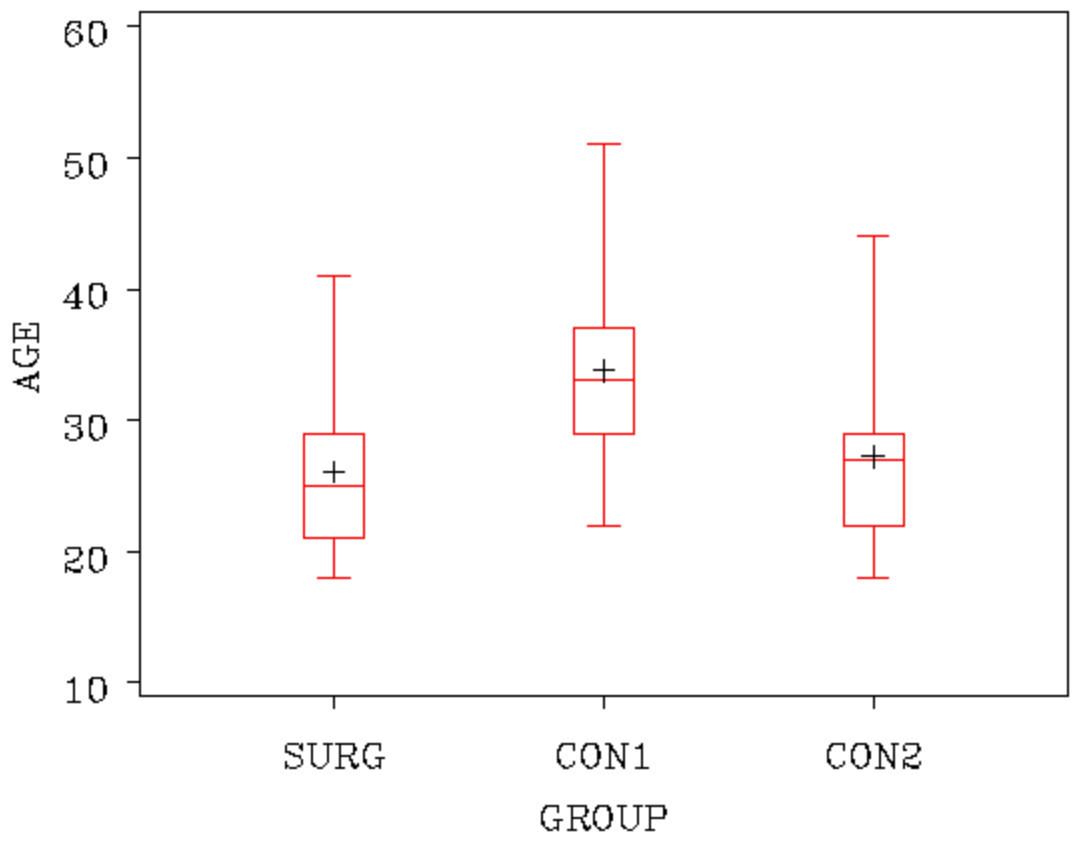
Program Note 12.1- One-Way ANOVA and Multiple Comparisons

PROC ANOVA can be used to analyze the age data shown in Table 12.1. **PROC ANOVA** uses a MODEL statement to identify the dependent and independent variables. The variable to the left of the equal sign is the dependent variable and the variable to the right of the equal sign is the independent variable. The independent variable is also identified in the CLASS statement. The MEANS statement tells SAS that we wish to see the mean of the dependent variable for each level of the variable shown in the MEANS statement. The words after the / symbol indicate which types of multiple comparisons we wish to use in the analysis. For the Dunnett procedure, we must specify which level of the independent variable is to be used in the comparisons with the other levels.

SAS commands:

```
DATA AGES;
  INPUT GROUP $ AGE @@;
DATALINES;
SURG 32  SURG 28  SURG 22  SURG 25  SURG 20  SURG 20  SURG 28
SURG 28  SURG 20  SURG 29  SURG 22  SURG 37  SURG 18  SURG 29
SURG 22  SURG 32  SURG 21  SURG 34  SURG 19  SURG 23  SURG 23
SURG 26  SURG 41  SURG 20  SURG 33
CON1 32  CON1 26  CON1 31  CON1 39  CON1 34  CON1 33  CON1 29
CON1 41  CON1 35  CON1 33  CON1 33  CON1 43  CON1 25  CON1 39
CON1 36  CON1 37  CON1 28  CON1 34  CON1 27  CON1 45  CON1 22
CON1 29  CON1 51  CON1 28  CON1 35
CON2 31  CON2 35  CON2 26  CON2 28  CON2 22  CON2 29  CON2 27
CON2 21  CON2 22  CON2 27  CON2 24  CON2 44  CON2 21  CON2 25
CON2 27  CON2 18  CON2 27  CON2 36
;
GOPTIONS DEVICE= GIF VPOS= 24 HPOS= 75 VSIZE= 5 HSIZE= 6 FTEXT=COMPLEX;
ODS HTML;
ODS GRAPHICS ON;
PROC BOXPLOT DATA=AGES;
  PLOT AGE*GROUP;
RUN;
QUIT;
ODS GRAPHICS OFF;
ODS HTML CLOSE;
```

SAS output:



The results from **PROC BOXPLOT** produce three box plots for the surgery and two control groups. Below are the SAS commands used to run the ANOVA procedure.

SAS commands:

```
PROC ANOVA DATA=AGES;
  CLASS GROUP;
  MODEL AGE=GROUP;
  MEANS GROUP / TUKEY LSD DUNNETT ('SURG');
RUN;
```

SAS output:

```
                          The SAS System
                       The ANOVA Procedure

                    Class Level Information

          Class         Levels    Values

          GROUP              3    CON1 CON2 SURG
```

Number of Observations Read 68
Number of Observations Used 68

The ANOVA Procedure

Dependent Variable: AGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	842.740065	421.370033	10.29	0.0001
Error	65	2660.951111	40.937709		
Corrected Total	67	3503.691176			

R-Square	Coeff Var	Root MSE	AGE Mean
0.240529	21.89640	6.398258	29.22059

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GROUP	2	842.7400654	421.3700327	10.29	0.0001

The ANOVA Procedure
t Tests (LSD) for AGE

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	40.93771
Critical Value of t	1.99714

Comparisons significant at the 0.05 level are indicated by ***.

GROUP Comparison	Difference Between Means	95% Confidence Limits		
CON1 - CON2	6.578	2.628	10.528	***
CON1 - SURG	7.720	4.106	11.334	***
CON2 - CON1	-6.578	-10.528	-2.628	***
CON2 - SURG	1.142	-2.808	5.092	
SURG - CON1	-7.720	-11.334	-4.106	***
SURG - CON2	-1.142	-5.092	2.808	

The ANOVA Procedure
Tukey's Studentized Range (HSD) Test for AGE

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	65
Error Mean Square	40.93771
Critical Value of Studentized Range	3.39207

Comparisons significant at the 0.05 level are indicated by ***.

	Difference	Simultaneous

```
                        GROUP          Between      95% Confidence
                    Comparison           Means          Limits

                    CON1 - CON2          6.578       1.834  11.322  ***
                    CON1 - SURG          7.720       3.379  12.061  ***
                    CON2 - CON1         -6.578     -11.322  -1.834  ***
                    CON2 - SURG          1.142      -3.602   5.886
                    SURG - CON1         -7.720     -12.061  -3.379  ***
                    SURG - CON2         -1.142      -5.886   3.602


                                  The ANOVA Procedure
                               Dunnett's t Tests for AGE
NOTE: This test controls the Type I experimentwise error for comparisons of all
treatments against a control.

                          Alpha                            0.05
                          Error Degrees of Freedom           65
                          Error Mean Square             40.93771
                          Critical Value of Dunnett's t  2.26597

              Comparisons significant at the 0.05 level are indicated by ***.

                                    Difference      Simultaneous
                        GROUP          Between      95% Confidence
                    Comparison           Means          Limits

                    CON1 - SURG          7.720       3.619  11.821  ***
                    CON2 - SURG          1.142      -3.339   5.624
```

Program Note 12.2- ANOVA Continued

1. Randomized Block with *k* Replicates per Cell

PROC ANOVA can also be used with the two-way ANOVA (or other more general ANOVAs as well). The model shown here does not include any interaction terms. The MODEL statement again has the dependent variable to the left of the equal sign and the independent variables to the right of the equal sign. The MEANS statement indicates that we wish to see the mean of the dependent variable shown for the levels of both of the independent variables.

SAS commands:

```
DATA WEIGHT;
  INPUT PROG $  SITE $  WTCHANGE  @@;
CARDS;
DIET OFFICE 6  DIET OFFICE 2  DIET OFFICE 10  DIET OFFICE -1
DIET OFFICE 8
DIET FACTORY 3  DIET FACTORY 15  DIET FACTORY 4  DIET FACTORY 8
DIET FACTORY 6
EXERCISE OFFICE 3  EXERCISE OFFICE 4  EXERCISE OFFICE -2
EXERCISE OFFICE 6  EXERCISE OFFICE -2
EXERCISE FACTORY -4  EXERCISE FACTORY 6  EXERCISE FACTORY 8
```

```
EXERCISE FACTORY -2  EXERCISE FACTORY 3
BOTH OFFICE 8  BOTH OFFICE 12  BOTH OFFICE 7  BOTH OFFICE 10
BOTH OFFICE 5
BOTH FACTORY 15  BOTH FACTORY 8  BOTH FACTORY 10
BOTH FACTORY 16  BOTH FACTORY 3
;
PROC ANOVA;
  CLASS PROG SITE;
  MODEL WTCHANGE = PROG SITE;
  MEANS PROG SITE;
RUN;
```

SAS output:

```
                                The SAS System
                              The ANOVA Procedure

                           Class Level Information

                    Class         Levels    Values
                    PROG               3    BOTH DIET EXERCISE
                    SITE               2    FACTORY OFFICE

                      Number of Observations Read          30
                      Number of Observations Used          30


                              The ANOVA Procedure
Dependent Variable: WTCHANGE
                                   Sum of
Source                     DF       Squares     Mean Square    F Value    Pr > F
Model                       3    292.5000000      97.5000000       5.33    0.0054
Error                      26    475.6666667      18.2948718
Corrected Total            29    768.1666667

                  R-Square     Coeff Var      Root MSE    WTCHANGE Mean
                  0.380777      73.32429      4.277250         5.833333


Source                     DF        Anova SS     Mean Square    F Value    Pr > F
PROG                        2    274.8666667     137.4333333       7.51    0.0027
SITE                        1     17.6333333      17.6333333       0.96    0.3353


                              The ANOVA Procedure

                      Level of            ----------WTCHANGE----------
                      PROG          N             Mean          Std Dev

                      BOTH         10       9.40000000       4.11501316
                      DIET         10       6.10000000       4.50801755
                      EXERCISE     10       2.00000000       4.18993503


                      Level of            ----------WTCHANGE----------
                      SITE          N             Mean          Std Dev
```

```
FACTORY      15      6.60000000      5.85296018
OFFICE       15      5.06666667      4.39913411
```

2. Balanced Two-Way ANOVA with Interaction

As you might have guessed, **PROC ANOVA** can also be used here to analyze the data shown in Table 12.7. Now that we are familiar with the simple, but very tedious, way of entering the data that we have used throughout, we shall complicate the input section a little by using a DO statement. The DO statement indicates how many times the statements included between the DO statement and its closing END statement are to be performed. In the following, we have one DO statement nested within another DO statement. When I equals 1, the value of METHOD is LECTURE. We then encounter the second DO statement and it tells SAS to set BOOK's value equal to 1. Then SAS reads all the values of the variable INCREASE on the next line of INPUT. The END statement for the second DO is encountered, and that sends us back to the top of the second DO loop. BOOK is now set equal to 2 and we read the values of INCREASE on the next line of INPUT. The END statement is again encountered, and we return to the top of the second DO statement and set BOOK equal to 3 and then read the next 6 values of INCREASE. We again encounter the END statement, but now there is no need to return to the top of the second DO statement because we have processed this DO all 3 times that were called for. Thus we move to the next statement which is the END statement for the first DO statement. We now set I equal to 2 and read the value of METHOD. Its value is now DISCUSS and we go through the second DO statement 3 more times, reading the values of INCREASE for each of the 3 BOOKs.

SAS commands:

```
DATA SCORES;
 ARRAY REP REP1 - REP6;
  INPUT METHOD $ BOOK REP1 - REP6;
    DO OVER REP;
      INCREASE = REP;
      OUTPUT;
    END;
CARDS;
LECTURE 1 30 43 12 18 22 16
LECTURE 2 21 26 10 14 17 16
LECTURE 3 42 30 18 10 21 18
DISCUSS 1 36 34 15 18 40 45
DISCUSS 2 33 31 28 15 29 26
DISCUSS 3 41 46 19 23 38 48
;
PROC ANOVA;
  CLASS METHOD BOOK;
  MODEL INCREASE = METHOD BOOK METHOD*BOOK;
  MEANS METHOD BOOK;
RUN;
```

SAS output:

```
                                The SAS System
                              The ANOVA Procedure

                           Class Level Information

                      Class         Levels    Values
                      METHOD             2    DISCUSS LECTURE
                      BOOK               3    1 2 3

                      Number of Observations Read          36
                      Number of Observations Used          36


                              The ANOVA Procedure

Dependent Variable: INCREASE
                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        5     1288.472222      257.694444       2.49    0.0529
Error                       30     3099.833333      103.327778
Corrected Total             35     4388.305556

                    R-Square     Coeff Var      Root MSE    INCREASE Mean
                    0.293615      38.56069      10.16503         26.36111


Source                      DF        Anova SS     Mean Square    F Value    Pr > F
METHOD                       1     910.0277778     910.0277778       8.81    0.0058
BOOK                         2     342.7222222     171.3611111       1.66    0.2074
METHOD*BOOK                  2      35.7222222      17.8611111       0.17    0.8421


                              The ANOVA Procedure

                       Level of            -----------INCREASE----------
                       METHOD         N            Mean          Std Dev
                       DISCUSS       18      31.3888889       10.5782178
                       LECTURE       18      21.3333333        9.6283894


                       Level of            -----------INCREASE----------
                       BOOK           N            Mean          Std Dev
                       1             12      27.4166667       11.9198714
                       2             12      22.1666667        7.6137834
                       3             12      29.5000000       12.9509564
```

In the MODEL statement, we have the term METHOD*BOOK which means to include the interaction of the two independent variables in the analysis as well as the two main effects. It was not necessary to write the terms METHOD and BOOK in the MODEL statement, as the use of the METHOD*BOOK term also tells SAS to include the main effect terms.

Program Note 12.3- General Linear Models Procedure

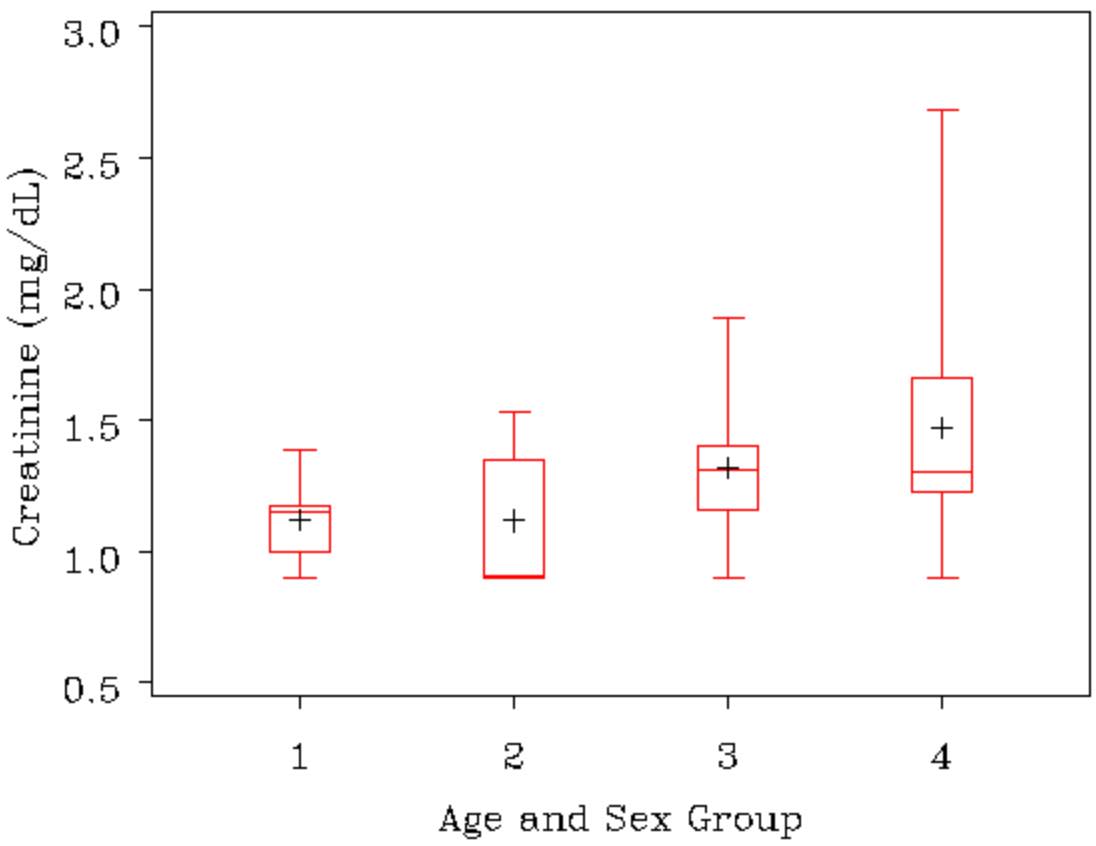
In this section we will show the commands for **PROC GLM** the SAS procedure for General Linear Models. A good reference book is the **SAS System for Linear Models 3rd edition** by Ramon C. Littell, Rudolf J. Freund, and Philip C. Spector. Beginning with the data from Example 12.9, we use the data shown in Table 12.10 which shows a cross-classification of creatinine measurements by sex and age. Notice in the SAS commands below, we use the DIG40 dataset to create age categories where individuals under 56 years of age are in AGECAT '1' and individuals 56 and older are in AGECAT '2'. For the purpose of creating box plots, we create the variable GROUP which consists of males (SEX '1') and females (SEX '2') further categorized by AGECAT.

SAS commands:

```
PROC IMPORT DATAFILE='C:\DIG40.XLS' OUT=DIG40DATA REPLACE;
RUN;

DATA DIG40;
  SET DIG40DATA;
IF AGE <  56 THEN AGECAT = 1;
IF AGE >= 56 THEN AGECAT = 2;
IF SEX = 2 AND AGE  < 56 THEN GROUP = '1';
IF SEX = 2 AND AGE >= 56 THEN GROUP = '2';
IF SEX = 1 AND AGE  < 56 THEN GROUP = '3';
IF SEX = 1 AND AGE >= 56 THEN GROUP = '4';
PROC SORT DATA=DIG40;
  BY GROUP;
RUN;
GOPTIONS DEVICE= GIF VPOS= 24 HPOS= 75 VSIZE= 5 HSIZE= 6 FTEXT=COMPLEX;
ODS HTML;
ODS GRAPHICS ON;
PROC BOXPLOT DATA=DIG40;
  PLOT CREAT*GROUP/ HAXIS=AXIS1 VAXIS=AXIS2;
  AXIS1 LABEL=('Age and Sex Group');
  AXIS2 LABEL=('Creatinine (mg/dL)');
RUN;
QUIT;
ODS GRAPHICS OFF;
ODS HTML CLOSE;
```

SAS output:



In the box plot presented above, the plus sign represents the group mean. Next we use **PROC GLM** to produce ANOVA by general linear models. Notice that the dependent variable CREAT follows the MODEL statement, and after the equal sign are the independent variables with the interaction term SEX*AGECAT. The output below refers to Table 12.11 in the text where we provide the Adjusted Type III Sum of Squares (Type III SS) along with the corresponding p-values (Pr > F) for the main effects and interaction term. In the final column in Table 12.11, we show the Type I Sum of Squares (Type I SS).

```
PROC GLM DATA=DIG40;
  CLASS SEX AGECAT;
  MODEL CREAT = SEX AGECAT SEX*AGECAT;
RUN;
QUIT;
```

SAS output:

```
                                The SAS System
                              The GLM Procedure

                          Class Level Information

                Class         Levels    Values
```

sex	2	1 2
AGECAT	2	1 2

Number of Observations Read	40
Number of Observations Used	40

The SAS System
The GLM Procedure

Dependent Variable: creat creat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.85927773	0.28642591	2.18	0.1070
Error	36	4.72465617	0.13124045		
Corrected Total	39	5.58393390			

R-Square	Coeff Var	Root MSE	creat Mean
0.153884	26.81405	0.362271	1.351050

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	0.71240430	0.71240430	5.43	0.0255
AGECAT	1	0.10416077	0.10416077	0.79	0.3789
sex*AGECAT	1	0.04271267	0.04271267	0.33	0.5719

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	0.55194454	0.55194454	4.21	0.0476
AGECAT	1	0.04074514	0.04074514	0.31	0.5808
sex*AGECAT	1	0.04271267	0.04271267	0.33	0.5719