Chapter 17
The ANOVA Procedure

Chapter Table of Contents

Chapter 17
The ANOVA Procedure

Overview

The ANOVA procedure performs *analysis of variance* (ANOVA) for balanced data from a wide variety of experimental designs. In analysis of variance, a continuous response variable, known as a *dependent variable*, is measured under experimental conditions identified by classification variables, known as *independent variables*. The variation in the response is assumed to be due to effects in the classification, with random error accounting for the remaining variation.

The ANOVA procedure is one of several procedures available in SAS/STAT software for analysis of variance. The ANOVA procedure is designed to handle balanced data (that is, data with equal numbers of observations for every combination of the classification factors), whereas the GLM procedure can analyze both balanced and unbalanced data. Because PROC ANOVA takes into account the special structure of a balanced design, it is faster and uses less storage than PROC GLM for balanced data.

Use PROC ANOVA for the analysis of balanced data only, with the following exceptions: one-way analysis of variance, Latin square designs, certain partially balanced incomplete block designs, completely nested (hierarchical) designs, and designs with cell frequencies that are proportional to each other and are also proportional to the background population. These exceptions have designs in which the factors are all orthogonal to each other. For further discussion, refer to Searle (1971, p. 138). PROC ANOVA works for designs with block diagonal $\mathbf{X}'\mathbf{X}$ matrices where the elements of each block all have the same value. The procedure partially tests this requirement by checking for equal cell means. However, this test is imperfect: some designs that cannot be analyzed correctly may pass the test, and designs that can be analyzed correctly may not pass. If your design does not pass the test, PROC ANOVA produces a warning message to tell you that the design is unbalanced and that the ANOVA analyses may not be valid; if your design is not one of the special cases described here, then you should use PROC GLM instead. Complete validation of designs is not performed in PROC ANOVA since this would require the whole $\mathbf{X}'\mathbf{X}$ matrix; if you're unsure about the validity of PROC ANOVA for your design, you should use PROC GLM.

Caution: If you use PROC ANOVA for analysis of unbalanced data, you must assume responsibility for the validity of the results.

Getting Started

The following examples demonstrate how you can use the ANOVA procedure to perform analyses of variance for a one-way layout and a randomized complete block design.

One-Way Layout with Means Comparisons

A one-way analysis of variance considers one treatment factor with two or more treatment levels. The goal of the analysis is to test for differences among the means of the levels and to quantify these differences. If there are two treatment levels, this analysis is equivalent to a t test comparing two group means.

The assumptions of analysis of variance (Steel and Torrie 1980) are

- treatment effects are additive
- experimental errors
  - are random
  - are independently distributed
  - follow a normal distribution
  - have mean zero and constant variance

The following example studies the effect of bacteria on the nitrogen content of red clover plants. The treatment factor is bacteria strain, and it has six levels. Five of the six levels consist of five different *Rhizobium trifolii* bacteria cultures combined with a composite of five *Rhizobium meliloti* strains. The sixth level is a composite of the five *Rhizobium trifolii* strains with the composite of the *Rhizobium meliloti*. Red clover plants are inoculated with the treatments, and nitrogen content is later measured in milligrams. The data are derived from an experiment by Erdman (1946) and are analyzed in Chapters 7 and 8 of Steel and Torrie (1980). The following DATA step creates the SAS data set Clover:

```
title ’Nitrogen Content of Red Clover Plants’;
data Clover;
   input Strain $ Nitrogen @@;
   datalines;
3DOK1  19.4 3DOK1  32.6 3DOK1  27.0 3DOK1  32.1 3DOK1  33.0
3DOK5  17.7 3DOK5  24.8 3DOK5  27.9 3DOK5  25.2 3DOK5  24.3
3DOK4  17.0 3DOK4  19.4 3DOK4   9.1 3DOK4  11.9 3DOK4  15.8
3DOK7  20.7 3DOK7  21.0 3DOK7  20.5 3DOK7  18.8 3DOK7  18.6
3DOK13 14.3 3DOK13 14.4 3DOK13 11.8 3DOK13 11.6 3DOK13 14.2
COMPOS 17.3 COMPOS 19.4 COMPOS 19.1 COMPOS 16.9 COMPOS 20.8
;
```

The variable Strain contains the treatment levels, and the variable Nitrogen contains the response. The following statements produce the analysis.

```
proc anova;
   class Strain;
   model Nitrogen = Strain;
run;
```

The classification variable is specified in the CLASS statement. Note that, unlike the GLM procedure, PROC ANOVA does not allow continuous variables on the right-hand side of the model. Figure 17.1 and Figure 17.2 display the output produced by these statements.

```
                     Nitrogen Content of Red Clover Plants

                              The ANOVA Procedure

                            Class Level Information

         Class         Levels    Values

         Strain             6    3DOK1 3DOK13 3DOK4 3DOK5 3DOK7 COMPOS


                          Number of observations    30
```

Figure 17.1. Class Level Information

The “Class Level Information” table shown in Figure 17.1 lists the variables that appear in the CLASS statement, their levels, and the number of observations in the data set.

Figure 17.2 displays the ANOVA table, followed by some simple statistics and tests of effects.

```
                     Nitrogen Content of Red Clover Plants

                              The ANOVA Procedure

Dependent Variable: Nitrogen

                                       Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        5     847.046667     169.409333     14.37   <.0001

 Error                       24     282.928000      11.788667

 Corrected Total             29    1129.974667


             R-Square     Coeff Var      Root MSE    Nitrogen Mean

             0.749616      17.26515      3.433463         19.88667


 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Strain                       5    847.0466667    169.4093333     14.37   <.0001
```

Figure 17.2. ANOVA Table

The degrees of freedom (DF) column should be used to check the analysis results. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, $6 - 1 = 5$. The Corrected Total degrees of freedom are always the total number of observations minus one; in this case $30 - 1 = 29$. The sum of Model and Error degrees of freedom equal the Corrected Total.

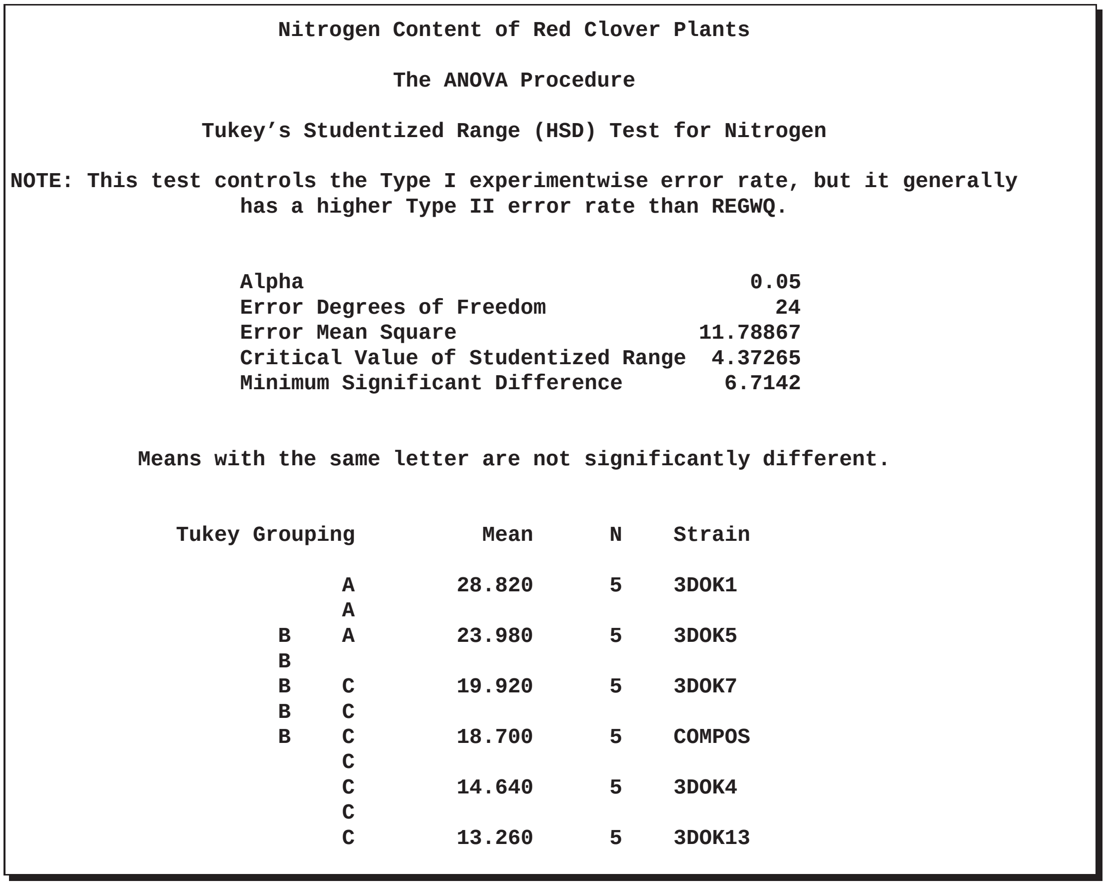
The overall F test is significant $(F = 14.37, p < 0.0001)$, indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable. The F test for Strain is significant, indicating that some contrast between the means for the different strains is different from zero. Notice that the Model and Strain F tests are identical, since Strain is the only term in the model.

The F test for Strain $(F = 14.37, p < 0.0001)$ suggests that there are differences among the bacterial strains, but it does not reveal any information about the nature of the differences. Mean comparison methods can be used to gather further information. The interactivity of PROC ANOVA enables you to do this without re-running the entire analysis. After you specify a model with a MODEL statement and execute the ANOVA procedure with a RUN statement, you can execute a variety of statements (such as MEANS, MANOVA, TEST, and REPEATED) without PROC ANOVA recalculating the model sum of squares.

The following command requests means of the Strain levels with Tukey's studentized range procedure.

```
   means Strain / tukey;
run;
```

Results of Tukey's procedure are shown in Figure 17.3.



```
                    Nitrogen Content of Red Clover Plants

                            The ANOVA Procedure

              Tukey’s Studentized Range (HSD) Test for Nitrogen

NOTE: This test controls the Type I experimentwise error rate, but it generally
                 has a higher Type II error rate than REGWQ.


                 Alpha                                     0.05
                 Error Degrees of Freedom                    24
                 Error Mean Square                     11.78867
                 Critical Value of Studentized Range  4.37265
                 Minimum Significant Difference        6.7142


         Means with the same letter are not significantly different.


            Tukey Grouping          Mean      N    Strain

                         A        28.820      5    3DOK1
                         A
                    B    A        23.980      5    3DOK5
                    B
                    B    C        19.920      5    3DOK7
                    B    C
                    B    C        18.700      5    COMPOS
                         C
                         C        14.640      5    3DOK4
                         C
                         C        13.260      5    3DOK13
```

Figure 17.3. Tukey’s Multiple Comparisons Procedure

The multiple comparisons results indicate, for example, that

- strain 3DOK1 fixes significantly more nitrogen than all but 3DOK5
- even though 3DOK5 is not significantly different from 3DOK1, it is also not significantly better than all the rest

Although the experiment has succeeded in separating the best strains from the worst, clearly distinguishing the very best strain requires more experimentation.

Randomized Complete Block with One Factor

This example illustrates the use of PROC ANOVA in analyzing a randomized complete block design. Researchers are interested in whether three treatments have different effects on the yield and worth of a particular crop. They believe that the experimental units are not homogeneous. So, a blocking factor is introduced that allows the experimental units to be homogeneous within each block. The three treatments are then randomly assigned within each block.

The data from this study are input into the SAS data set RCB:

```
title 'Randomized Complete Block';
data RCB;
   input Block Treatment $ Yield Worth @@;
   datalines;
1 A 32.6 112   1 B 36.4 130   1 C 29.5 106
2 A 42.7 139   2 B 47.1 143   2 C 32.9 112
3 A 35.3 124   3 B 40.1 134   3 C 33.6 116
;
```

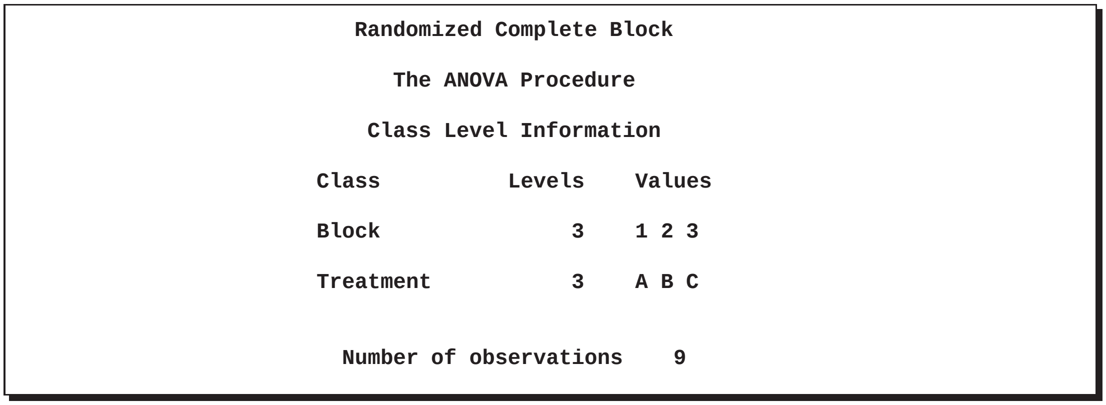
The variables Yield and Worth are continuous response variables, and the variables Block and Treatment are the classification variables. Because the data for the analysis are balanced, you can use PROC ANOVA to run the analysis.

The statements for the analysis are

```
proc anova;
   class Block Treatment;
   model Yield Worth=Block Treatment;
run;
```

The Block and Treatment effects appear in the CLASS statement. The MODEL statement requests an analysis for each of the two dependent variables, Yield and Worth.

Figure 17.4 shows the “Class Level Information” table.

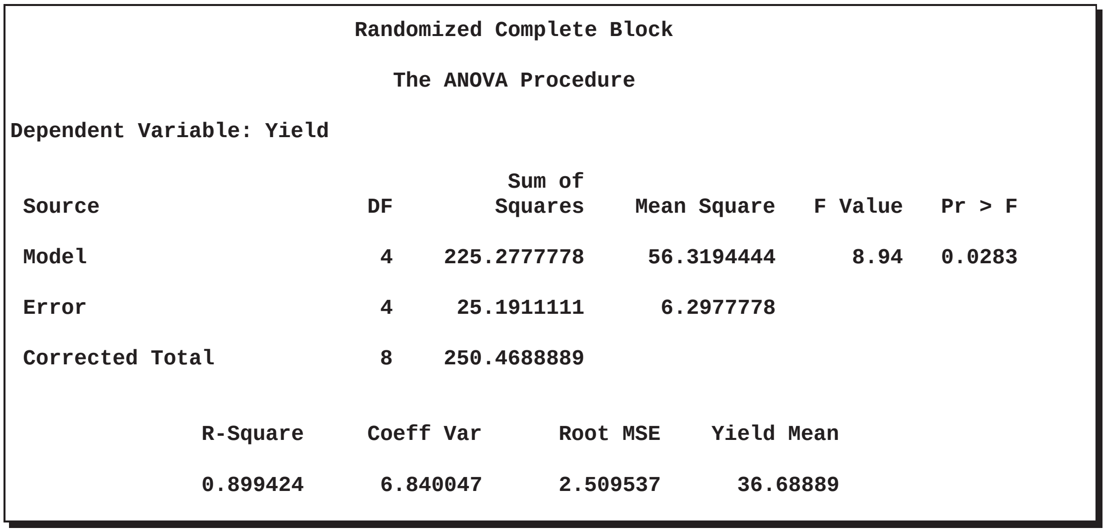


```
                    Randomized Complete Block

                       The ANOVA Procedure

                    Class Level Information

                Class          Levels    Values

                Block               3    1 2 3

                Treatment           3    A B C


                  Number of observations    9
```

Figure 17.4. Class Level Information

The “Class Level Information” table lists the number of levels and their values for all effects specified in the CLASS statement. The number of observations in the data set are also displayed. Use this information to make sure that the data have been read correctly.

The overall ANOVA table for Yield in Figure 17.5 appears first in the output because it is the first response variable listed on the left side in the MODEL statement.



```
                            Randomized Complete Block

                               The ANOVA Procedure

Dependent Variable: Yield

                                        Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        4    225.2777778     56.3194444      8.94   0.0283

 Error                        4     25.1911111      6.2977778

 Corrected Total              8    250.4688889


               R-Square     Coeff Var      Root MSE    Yield Mean

               0.899424      6.840047      2.509537      36.68889
```

Figure 17.5. Overall ANOVA Table for Yield

The overall F statistic is significant $(F = 8.94, p = 0.02583)$, indicating that the model as a whole accounts for a significant portion of the variation in Yield and that you may proceed to tests of effects.

The degrees of freedom (DF) are used to ensure correctness of the data and model. The Corrected Total degrees of freedom are one less than the total number of observations in the data set; in this case, $9 - 1 = 8$. The Model degrees of freedom for a randomized complete block are $(b - 1) + (t - 1)$, where $b =$number of block levels and $t =$number of treatment levels. In this case, $(3 - 1) + (3 - 1) = 4$.

Several simple statistics follow the ANOVA table. The R-Square indicates that the model accounts for nearly 90% of the variation in the variable Yield. The coefficient of variation (C.V.) is listed along with the Root MSE and the mean of the dependent variable. The Root MSE is an estimate of the standard deviation of the dependent variable. The C.V. is a unitless measure of variability.

The tests of the effects shown in Figure 17.6 are displayed after the simple statistics.

```
                           Randomized Complete Block

                              The ANOVA Procedure

Dependent Variable: Yield

 Source                      DF        Anova SS    Mean Square   F Value   Pr > F

 Block                        2      98.1755556     49.0877778      7.79   0.0417
 Treatment                    2     127.1022222     63.5511111     10.09   0.0274
```

Figure 17.6. Tests of Effects for Yield

For Yield, both the Block and Treatment effects are significant $(F = 7.79, p = 0.0417$ and $F = 10.09, p = 0.0274$, respectively) at the 95% level. From this you can conclude that blocking is useful for this variable and that some contrast between the treatment means is significantly different from zero.

Figure 17.7 shows the ANOVA table, simple statistics, and tests of effects for the variable Worth.

```
                           Randomized Complete Block

                              The ANOVA Procedure

Dependent Variable: Worth

                                             Sum of
 Source                      DF         Squares    Mean Square   F Value   Pr > F

 Model                        4     1247.333333     311.833333      8.28   0.0323

 Error                        4      150.666667      37.666667

 Corrected Total              8     1398.000000


               R-Square     Coeff Var      Root MSE    Worth Mean

               0.892227      4.949450      6.137318      124.0000


 Source                      DF        Anova SS    Mean Square   F Value   Pr > F

 Block                        2     354.6666667    177.3333333      4.71   0.0889
 Treatment                    2     892.6666667    446.3333333     11.85   0.0209
```

Figure 17.7. ANOVA Table for Worth

The overall F test is significant $(F = 8.28, p = 0.0323)$ at the 95% level for the variable Worth. The Block effect is not significant at the 0.05 level but is significant at the 0.10 confidence level $(F = 4.71, p = 0.0889)$. Generally, the usefulness of blocking should be determined before the analysis. However, since there are two dependent variables of interest, and Block is significant for one of them (Yield), blocking appears to be generally useful. For Worth, as with Yield, the effect of Treatment is significant $(F = 11.85, p = 0.0209)$.

Issuing the following command produces the Treatment means.

```
      means Treatment;
   run;
```

Figure 17.8 displays the treatment means and their standard deviations for both dependent variables.

```
                           Randomized Complete Block

                              The ANOVA Procedure

Level of             ------------Yield-----------    ------------Worth-----------
Treatment     N             Mean          Std Dev             Mean          Std Dev

A             3       36.8666667       5.22908532       125.000000       13.5277493
B             3       41.2000000       5.43415127       135.666667        6.6583281
C             3       32.0000000       2.19317122       111.333333        5.0332230
```

Figure 17.8. Means of Yield and Worth

Syntax

The following statements are available in PROC ANOVA.

PROC ANOVA < *options* > **;**
CLASS *variables* **;**
MODEL *dependents=effects* < */ options* > **;**
ABSORB *variables* **;**
BY *variables* **;**
FREQ *variable* **;**
MANOVA < *test-options* >< */ detail-options* > **;**
MEANS *effects* < */ options* > **;**
REPEATED *factor-specification* < */ options* > **;**
TEST < **H=***effects* > **E=***effect* **;**

The PROC ANOVA, CLASS, and MODEL statements are required, and they must precede the first RUN statement. The CLASS statement must precede the MODEL statement. If you use the ABSORB, FREQ, or BY statement, it must precede the first RUN statement. The MANOVA, MEANS, REPEATED, and TEST statements must follow the MODEL statement, and they can be specified in any order. These four statements can also appear after the first RUN statement.

The following table summarizes the function of each statement (other than the PROC statement) in the ANOVA procedure:

Table 17.1. Statements in the ANOVA Procedure

Statement	**Description**
ABSORB	absorbs classification effects in a model
BY	specifies variables to define subgroups for the analysis
CLASS	declares classification variables
FREQ	specifies a frequency variable
MANOVA	performs a multivariate analysis of variance
MEANS	computes and compares means
MODEL	defines the model to be fit
REPEATED	performs multivariate and univariate repeated measures analysis of variance
TEST	constructs tests using the sums of squares for effects and the error term you specify

PROC ANOVA Statement

PROC ANOVA $<$ *options* $>$ **;**

The PROC ANOVA statement starts the ANOVA procedure.

You can specify the following options in the PROC ANOVA statement:

DATA=*SAS-data-set*

names the SAS data set used by the ANOVA procedure. By default, PROC ANOVA uses the most recently created SAS data set.

MANOVA

requests the multivariate mode of eliminating observations with missing values. If any of the dependent variables have missing values, the procedure eliminates that observation from the analysis. The MANOVA option is useful if you use PROC ANOVA in interactive mode and plan to perform a multivariate analysis.

MULTIPASS

requests that PROC ANOVA reread the input data set, when necessary, instead of writing the values of dependent variables to a utility file. This option decreases disk space usage at the expense of increased execution times and is useful only in rare situations where disk space is at an absolute premium.

NAMELEN=n

specifies the length of effect names to be n characters long, where n is a value between 20 and 200 characters. The default length is 20 characters.

NOPRINT

suppresses the normal display of results. The NOPRINT option is useful when you want to create only the output data set with the procedure. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "Using the Output Delivery System," for more information.

ORDER=DATA | FORMATTED | FREQ | INTERNAL

specifies the sorting order for the levels of the classification variables (specified in the CLASS statement). This ordering determines which parameters in the model correspond to each level in the data. Note that the ORDER= option applies to the levels for all classification variables. The exception is ORDER=FORMATTED (the default) for numeric variables for which you have supplied no explicit format (that is, for which there is no corresponding FORMAT statement in the current PROC ANOVA run or in the DATA step that created the data set). In this case, the levels are ordered by their internal (numeric) value. Note that this represents a change from previous releases for how class levels are ordered. In releases previous to Version 8, numeric class levels with no explicit format were ordered by their BEST12. formatted values, and in order to revert to the previous ordering you can specify this format explicitly for the affected classification variables. The change was implemented because the former default behavior for ORDER=FORMATTED often resulted in levels not being ordered numerically and usually required the user to intervene with an explicit format or ORDER=INTERNAL to get the more natural ordering.

The following table shows how PROC ANOVA interprets values of the ORDER= option.

Value of ORDER=	Levels Sorted By
DATA	order of appearance in the input data set
FORMATTED	external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
FREQ	descending frequency count; levels with the most observations come first in the order
INTERNAL	unformatted value

OUTSTAT=*SAS-data-set*

names an output data set that contains sums of squares, degrees of freedom, F statistics, and probability levels for each effect in the model. If you use the CANONICAL option in the MANOVA statement and do not use an M= specification in the MANOVA statement, the data set also contains results of the canonical analysis. See the “Output Data Set” section on page 370 for more information.

ABSORB Statement

ABSORB *variables* **;**

Absorption is a computational technique that provides a large reduction in time and memory requirements for certain types of models. The *variables* are one or more variables in the input data set.

For a main effect variable that does not participate in interactions, you can absorb the effect by naming it in an ABSORB statement. This means that the effect can be adjusted out before the construction and solution of the rest of the model. This is particularly useful when the effect has a large number of levels.

Several variables can be specified, in which case each one is assumed to be nested in the preceding variable in the ABSORB statement.

Note: When you use the ABSORB statement, the data set (or each BY group, if a BY statement appears) must be sorted by the variables in the ABSORB statement. Including an absorbed variable in the CLASS list or in the MODEL statement may produce erroneous sums of squares. If the ABSORB statement is used, it must appear before the first RUN statement or it is ignored.

When you use an ABSORB statement and also use the INT option in the MODEL statement, the procedure ignores the option but produces the uncorrected total sum of squares (SS) instead of the corrected total SS.

See the “Absorption” section on page 1532 in Chapter 30, “The GLM Procedure,” for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC ANOVA to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the ANOVA procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in base SAS software).

Since sorting the data changes the order in which PROC ANOVA reads observations, the sorting order for the levels of the classification variables may be affected if you have also specified the ORDER=DATA option in the PROC ANOVA statement.

If the BY statement is used, it must appear before the first RUN statement or it is ignored. When you use a BY statement, the interactive features of PROC ANOVA are disabled.

When both a BY and an ABSORB statement are used, observations must be sorted first by the variables in the BY statement, and then by the variables in the ABSORB statement.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

CLASS Statement

CLASS *variables* **;**

The CLASS statement names the classification variables to be used in the model. Typical class variables are TREATMENT, SEX, RACE, GROUP, and REPLICATION. The CLASS statement is required, and it must appear before the MODEL statement.

Class levels are determined from up to the first 16 characters of the formatted values of the CLASS variables. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and the discussions for the FORMAT statement and SAS formats in *SAS Language Reference: Concepts*.

FREQ Statement

FREQ *variable* **;**

The FREQ statement names a variable that provides frequencies for each observation in the DATA= data set. Specifically, if n is the value of the FREQ variable for a given observation, then that observation is used n times.

The analysis produced using a FREQ statement reflects the expanded number of observations. For example, means and total degrees of freedom reflect the expanded number of observations. You can produce the same analysis (without the FREQ statement) by first creating a new data set that contains the expanded number of observations. For example, if the value of the FREQ variable is 5 for the first observation, the first 5 observations in the new data set would be identical. Each observation in the old data set would be replicated n_i times in the new data set, where n_i is the value of the FREQ variable for that observation.

If the value of the FREQ variable is missing or is less than 1, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.

If the FREQ statement is used, it must appear before the first RUN statement or it is ignored.

MANOVA Statement

MANOVA $<$ *test-options* $><$ */ detail-options* $>$ **;**

If the MODEL statement includes more than one dependent variable, you can perform multivariate analysis of variance with the MANOVA statement. The *test-options* define which effects to test, while the *detail-options* specify how to execute the tests and what results to display.

When a MANOVA statement appears before the first RUN statement, PROC ANOVA enters a multivariate mode with respect to the handling of missing values; in addition to observations with missing independent variables, observations with *any* missing dependent variables are excluded from the analysis. If you want to use this mode of handling missing values but do not need any multivariate analyses, specify the MANOVA option in the PROC ANOVA statement.

Test-Options

You can specify the following options in the MANOVA statement as *test-options* in order to define which multivariate tests to perform.

H=*effects* | **INTERCEPT** | **_ALL_**

specifies effects in the preceding model to use as hypothesis matrices. for multivariate tests For each SSCP matrix $\mathbf{H}$ associated with an effect, the H= specification computes an analysis based on the characteristic roots of $\mathbf{E}^{-1}\mathbf{H}$, where $\mathbf{E}$ is the matrix associated with the error effect. The characteristic roots and vectors are displayed, along with the Hotelling-Lawley trace, Pillai's trace, Wilks' criterion, and Roy's maximum root criterion with approximate F statistics. Use the keyword INTERCEPT to produce tests for the intercept. To produce tests for all effects listed in the MODEL statement, use the keyword _ALL_ in place of a list of effects. For background and further details, see the "Multivariate Analysis of Variance" section on page 1558 in Chapter 30, "The GLM Procedure."

E=*effect*

specifies the error effect. If you omit the E= specification, the ANOVA procedure uses the error SSCP (residual) matrix from the analysis.

M=*equation,. . .,equation* | **(***row-of-matrix,. . .,row-of-matrix***)**

specifies a transformation matrix for the dependent variables listed in the MODEL statement. The equations in the M= specification are of the form

$$\begin{aligned} c_1 \times \textit{dependent-variable} \quad & \pm \quad c_2 \times \textit{dependent-variable} \\ \cdots \quad & \pm \quad c_n \times \textit{dependent-variable} \end{aligned}$$

where the c_i values are coefficients for the various *dependent-variables*. If the value of a given c_i is 1, it may be omitted; in other words $1 \times Y$ is the same as Y. Equations should involve two or more dependent variables. For sample syntax, see the "Examples" section on page 354.

Alternatively, you can input the transformation matrix directly by entering the elements of the matrix with commas separating the rows, and parentheses surrounding the matrix. When this alternate form of input is used, the number of elements in each row must equal the number of dependent variables. Although these combinations actually represent the columns of the $\mathbf{M}$ matrix, they are displayed by rows.

When you include an M= specification, the analysis requested in the MANOVA statement is carried out for the variables defined by the equations in the specification, not the original dependent variables. If you omit the M= option, the analysis is performed for the original dependent variables in the MODEL statement.

If an M= specification is included without either the MNAMES= or the PREFIX= option, the variables are labeled MVAR1, MVAR2, and so forth by default. For further information, see the section "Multivariate Analysis of Variance" on page 1558 in Chapter 30, "The GLM Procedure."

MNAMES=*names*

provides names for the variables defined by the equations in the M= specification. Names in the list correspond to the M= equations or the rows of the $\mathbf{M}$ matrix (as it is entered).

PREFIX=*name*

is an alternative means of identifying the transformed variables defined by the M= specification. For example, if you specify PREFIX=DIFF, the transformed variables are labeled DIFF1, DIFF2, and so forth.

Detail-Options

You can specify the following options in the MANOVA statement after a slash as *detail-options*:

CANONICAL

produces a canonical analysis of the $\mathbf{H}$ and $\mathbf{E}$ matrices (transformed by the $\mathbf{M}$ matrix, if specified) instead of the default display of characteristic roots and vectors.

ORTH

requests that the transformation matrix in the M= specification of the MANOVA statement be orthonormalized by rows before the analysis.

PRINTE

displays the error SSCP matrix $\mathbf{E}$. If the $\mathbf{E}$ matrix is the error SSCP (residual) matrix from the analysis, the partial correlations of the dependent variables given the independent variables are also produced.

For example, the statement

```
manova / printe;
```

displays the error SSCP matrix and the partial correlation matrix computed from the error SSCP matrix.

PRINTH

displays the hypothesis SSCP matrix $\mathbf{H}$ associated with each effect specified by the H= specification.

SUMMARY

produces analysis-of-variance tables for each dependent variable. When no $\mathbf{M}$ matrix is specified, a table is produced for each original dependent variable from the MODEL statement; with an $\mathbf{M}$ matrix other than the identity, a table is produced for each transformed variable defined by the $\mathbf{M}$ matrix.

Examples

The following statements give several examples of using a MANOVA statement.

```
proc anova;
   class A B;
   model Y1-Y5=A B(A);
   manova h=A e=B(A) / printh printe;
   manova h=B(A) / printe;
   manova h=A e=B(A) m=Y1-Y2,Y2-Y3,Y3-Y4,Y4-Y5
          prefix=diff;
   manova h=A e=B(A) m=(1 -1  0  0  0,
                        0  1 -1  0  0,
                        0  0  1 -1  0,
                        0  0  0  1 -1) prefix=diff;
run;
```

The first MANOVA statement specifies A as the hypothesis effect and B(A) as the error effect. As a result of the PRINTH option, the procedure displays the hypothesis SSCP matrix associated with the A effect; and, as a result of the PRINTE option, the procedure displays the error SSCP matrix associated with the B(A) effect.

The second MANOVA statement specifies B(A) as the hypothesis effect. Since no error effect is specified, PROC ANOVA uses the error SSCP matrix from the analysis as the **E** matrix. The PRINTE option displays this **E** matrix. Since the **E** matrix is the error SSCP matrix from the analysis, the partial correlation matrix computed from this matrix is also produced.

The third MANOVA statement requests the same analysis as the first MANOVA statement, but the analysis is carried out for variables transformed to be successive differences between the original dependent variables. The PREFIX=DIFF specification labels the transformed variables as DIFF1, DIFF2, DIFF3, and DIFF4.

Finally, the fourth MANOVA statement has the identical effect as the third, but it uses an alternative form of the M= specification. Instead of specifying a set of equations, the fourth MANOVA statement specifies rows of a matrix of coefficients for the five dependent variables.

As a second example of the use of the M= specification, consider the following:

```
proc anova;
   class group;
   model dose1-dose4=group / nouni;
   manova h = group
          m = -3*dose1 -   dose2 +   dose3 + 3*dose4,
                 dose1 -   dose2 -   dose3 +   dose4,
                -dose1 + 3*dose2 - 3*dose3 +   dose4
          mnames = Linear Quadratic Cubic
          / printe;
run;
```

The M= specification gives a transformation of the dependent variables dose1 through dose4 into orthogonal polynomial components, and the MNAMES= option labels the transformed variables as LINEAR, QUADRATIC, and CUBIC, respectively. Since the PRINTE option is specified and the default residual matrix is used as an error term, the partial correlation matrix of the orthogonal polynomial components is also produced.

For further information, see the “Multivariate Analysis of Variance” section on page 1558 in Chapter 30, “The GLM Procedure.”

MEANS Statement

MEANS *effects* < **/** *options* > **;**

PROC ANOVA can compute means of the dependent variables for any effect that appears on the right-hand side in the MODEL statement.

You can use any number of MEANS statements, provided that they appear after the MODEL statement. For example, suppose A and B each have two levels. Then, if you use the following statements

```
proc anova;
   class A B;
   model Y=A B A*B;
   means A B / tukey;
   means A*B;
run;
```

means, standard deviations, and Tukey's multiple comparison tests are produced for each level of the main effects A and B, and just the means and standard deviations for each of the four combinations of levels for A*B. Since multiple comparisons options apply only to main effects, the single MEANS statement

```
means A B A*B / tukey;
```

produces the same results.

Options are provided to perform multiple comparison tests for only main effects in the model. PROC ANOVA does not perform multiple comparison tests for interaction terms in the model; for multiple comparisons of interaction terms, see the LSMEANS statement in Chapter 30, "The GLM Procedure." The following table summarizes categories of options available in the MEANS statement.

Table 17.2. Options Available in the MEANS Statement

Task	**Available options**
Perform multiple comparison tests	BON DUNCAN DUNNETT DUNNETTL DUNNETTU GABRIEL GT2 LSD REGWQ SCHEFFE SIDAK

Table 17.2. (continued)

Task	Available options
Perform multiple comparison tests	SMM SNK T TUKEY WALLER
Specify additional details for multiple comparison tests	ALPHA= CLDIFF CLM E= KRATIO= LINES NOSORT
Test for homogeneity of variances	HOVTEST
Compensate for heterogeneous variances	WELCH

Descriptions of these options follow. For a further discussion of these options, see the section "Multiple Comparisons" on page 1540 in Chapter 30, "The GLM Procedure."

ALPHA=*p*

specifies the level of significance for comparisons among the means. By default, ALPHA=0.05. You can specify any value greater than 0 and less than 1.

BON

performs Bonferroni t tests of differences between means for all main effect means in the MEANS statement. See the CLDIFF and LINES options, which follow, for a discussion of how the procedure displays results.

CLDIFF

presents results of the BON, GABRIEL, SCHEFFE, SIDAK, SMM, GT2, T, LSD, and TUKEY options as confidence intervals for all pairwise differences between means, and the results of the DUNNETT, DUNNETTU, and DUNNETTL options as confidence intervals for differences with the control. The CLDIFF option is the default for unequal cell sizes unless the DUNCAN, REGWQ, SNK, or WALLER option is specified.

CLM

presents results of the BON, GABRIEL, SCHEFFE, SIDAK, SMM, T, and LSD options as intervals for the mean of each level of the variables specified in the MEANS statement. For all options except GABRIEL, the intervals are confidence intervals for the true means. For the GABRIEL option, they are *comparison intervals* for comparing means pairwise: in this case, if the intervals corresponding to two means overlap, the difference between them is insignificant according to Gabriel's method.

DUNCAN

performs Duncan's multiple range test on all main effect means given in the MEANS statement. See the LINES option for a discussion of how the procedure displays results.

DUNNETT < **(***formatted-control-values***)** >

performs Dunnett's two-tailed t test, testing if any treatments are significantly different from a single control for all main effects means in the MEANS statement.

To specify which level of the effect is the control, enclose the formatted value in quotes in parentheses after the keyword. If more than one effect is specified in the MEANS statement, you can use a list of control values within the parentheses. By default, the first level of the effect is used as the control. For example,

```
means a / dunnett('CONTROL');
```

where CONTROL is the formatted control value of A. As another example,

```
means a b c / dunnett('CNTLA' 'CNTLB' 'CNTLC');
```

where CNTLA, CNTLB, and CNTLC are the formatted control values for A, B, and C, respectively.

DUNNETTL < **(***formatted-control-value***)** >

performs Dunnett's one-tailed t test, testing if any treatment is significantly less than the control. Control level information is specified as described previously for the DUNNETT option.

DUNNETTU < **(***formatted-control-value***)** >

performs Dunnett's one-tailed t test, testing if any treatment is significantly greater than the control. Control level information is specified as described previously for the DUNNETT option.

E=*effect*

specifies the error mean square used in the multiple comparisons. By default, PROC ANOVA uses the residual Mean Square (MS). The effect specified with the E= option must be a term in the model; otherwise, the procedure uses the residual MS.

GABRIEL

performs Gabriel's multiple-comparison procedure on all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

GT2

see the SMM option.

HOVTEST
HOVTEST=BARTLETT
HOVTEST=BF
HOVTEST=LEVENE <**(TYPE=ABS | SQUARE)**>
HOVTEST=OBRIEN <**(W=***number* **)**>

requests a homogeneity of variance test for the groups defined by the MEANS effect. You can optionally specify a particular test; if you do not specify a test, Levene's test (Levene 1960) with TYPE=SQUARE is computed. Note that this option is ignored unless your MODEL statement specifies a simple one-way model.

The HOVTEST=BARTLETT option specifies Bartlett's test (Bartlett 1937), a modification of the normal-theory likelihood ratio test.

The HOVTEST=BF option specifies Brown and Forsythe's variation of Levene's test (Brown and Forsythe 1974).

The HOVTEST=LEVENE option specifies Levene's test (Levene 1960), which is widely considered to be the standard homogeneity of variance test. You can use the TYPE= option in parentheses to specify whether to use the absolute residuals (TYPE=ABS) or the squared residuals (TYPE=SQUARE) in Levene's test. The default is TYPE=SQUARE.

The HOVTEST=OBRIEN option specifies O'Brien's test (O'Brien 1979), which is basically a modification of HOVTEST=LEVENE(TYPE=SQUARE). You can use the W= option in parentheses to tune the variable to match the suspected kurtosis of the underlying distribution. By default, W=0.5, as suggested by O'Brien (1979, 1981).

See the section "Homogeneity of Variance in One-Way Models" on page 1553 in Chapter 30, "The GLM Procedure," for more details on these methods. Example 30.10 on page 1623 in the same chapter illustrates the use of the HOVTEST and WELCH options in the MEANS statement in testing for equal group variances.

KRATIO=*value*

specifies the Type 1/Type 2 error seriousness ratio for the Waller-Duncan test. Reasonable values for KRATIO are 50, 100, and 500, which roughly correspond for the two-level case to ALPHA levels of 0.1, 0.05, and 0.01. By default, the procedure uses the default value of 100.

LINES

presents results of the BON, DUNCAN, GABRIEL, REGWQ, SCHEFFE, SIDAK, SMM, GT2, SNK, T, LSD, TUKEY, and WALLER options by listing the means in descending order and indicating nonsignificant subsets by line segments beside the corresponding means. The LINES option is appropriate for equal cell sizes, for which it is the default. The LINES option is also the default if the DUNCAN, REGWQ, SNK, or WALLER option is specified, or if there are only two cells of unequal size. If the cell sizes are unequal, the harmonic mean of the cell sizes is used, which may lead to somewhat liberal tests if the cell sizes are highly disparate. The LINES option cannot be used in combination with the DUNNETT, DUNNETTL, or DUNNETTU option. In addition, the procedure has a restriction that no more than 24 overlapping groups of means can exist. If a mean belongs to more than 24 groups, the procedure issues an error message. You can either reduce the number of levels of the variable or use a multiple comparison test that allows the CLDIFF option rather than the LINES option.

LSD

see the T option.

NOSORT

prevents the means from being sorted into descending order when the CLDIFF or CLM option is specified.

REGWQ

performs the Ryan-Einot-Gabriel-Welsch multiple range test on all main effect means in the MEANS statement. See the LINES option for a discussion of how the procedure displays results.

SCHEFFE

performs Scheffé's multiple-comparison procedure on all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SIDAK

performs pairwise t tests on differences between means with levels adjusted according to Sidak's inequality for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SMM
GT2

performs pairwise comparisons based on the studentized maximum modulus and Sidak's uncorrelated-t inequality, yielding Hochberg's GT2 method when sample sizes are unequal, for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

SNK

performs the Student-Newman-Keuls multiple range test on all main effect means in the MEANS statement. See the LINES option for a discussion of how the procedure displays results.

T
LSD

performs pairwise t tests, equivalent to Fisher's least-significant-difference test in the case of equal cell sizes, for all main effect means in the MEANS statement. See the CLDIFF and LINES options for discussions of how the procedure displays results.

TUKEY

performs Tukey's studentized range test (HSD) on all main effect means in the MEANS statement. (When the group sizes are different, this is the Tukey-Kramer test.) See the CLDIFF and LINES options for discussions of how the procedure displays results.

WALLER

performs the Waller-Duncan k-ratio t test on all main effect means in the MEANS statement. See the KRATIO= option for information on controlling details of the test, and see the LINES option for a discussion of how the procedure displays results.

WELCH

requests Welch's (1951) variance-weighted one-way ANOVA. This alternative to the usual analysis of variance for a one-way model is robust to the assumption of equal within-group variances. This option is ignored unless your MODEL statement specifies a simple one-way model.

Note that using the WELCH option merely produces one additional table consisting of Welch's ANOVA. It does not affect all of the other tests displayed by the ANOVA procedure, which still require the assumption of equal variance for exact validity.

See the "Homogeneity of Variance in One-Way Models" section on page 1553 in Chapter 30, "The GLM Procedure," for more details on Welch's ANOVA. Example 30.10 on page 1623 in the same chapter illustrates the use of the HOVTEST and WELCH options in the MEANS statement in testing for equal group variances.

MODEL Statement

MODEL *dependents=effects* < */ options* > **;**

The MODEL statement names the dependent variables and independent effects. The syntax of effects is described in the section "Specification of Effects" on page 366. If no independent effects are specified, only an intercept term is fit. This tests the hypothesis that the mean of the dependent variable is zero. All variables in effects that you specify in the MODEL statement must appear in the CLASS statement because PROC ANOVA does not allow for continuous effects.

You can specify the following options in the MODEL statement; they must be separated from the list of independent effects by a slash.

INTERCEPT
INT

displays the hypothesis tests associated with the intercept as an effect in the model. By default, the procedure includes the intercept in the model but does not display associated tests of hypotheses. Except for producing the uncorrected total SS instead of the corrected total SS, the INT option is ignored when you use an ABSORB statement.

NOUNI

suppresses the display of univariate statistics. You typically use the NOUNI option with a multivariate or repeated measures analysis of variance when you do not need the standard univariate output. The NOUNI option in a MODEL statement does not affect the univariate output produced by the REPEATED statement.

REPEATED Statement

REPEATED *factor-specification* < **/** *options* > **;**

When values of the dependent variables in the MODEL statement represent repeated measurements on the same experimental unit, the REPEATED statement enables you to test hypotheses about the measurement factors (often called *within-subject factors*), as well as the interactions of within-subject factors with independent variables in the MODEL statement (often called *between-subject factors*). The REPEATED statement provides multivariate and univariate tests as well as hypothesis tests for a variety of single-degree-of-freedom contrasts. There is no limit to the number of within-subject factors that can be specified. For more details, see the “Repeated Measures Analysis of Variance” section on page 1560 in Chapter 30, “The GLM Procedure.”

The REPEATED statement is typically used for handling repeated measures designs with one repeated response variable. Usually, the variables on the left-hand side of the equation in the MODEL statement represent one repeated response variable. This does not mean that only one factor can be listed in the REPEATED statement. For example, one repeated response variable (hemoglobin count) might be measured 12 times (implying variables Y1 to Y12 on the left-hand side of the equal sign in the MODEL statement), with the associated within-subject factors treatment and time (implying two factors listed in the REPEATED statement). See the “Examples” section on page 365 for an example of how PROC ANOVA handles this case. Designs with two or more repeated response variables can, however, be handled with the IDENTITY transformation; see Example 30.9 on page 1618 in Chapter 30, “The GLM Procedure,” for an example of analyzing a doubly-multivariate repeated measures design.

When a REPEATED statement appears, the ANOVA procedure enters a multivariate mode of handling missing values. If any values for variables corresponding to each combination of the within-subject factors are missing, the observation is excluded from the analysis.

The simplest form of the REPEATED statement requires only a *factor-name*. With two repeated factors, you must specify the *factor-name* and number of levels (*levels*) for each factor. Optionally, you can specify the actual values for the levels (*level-values*), a *transformation* that defines single-degree-of freedom contrasts, and *options* for additional analyses and output. When more than one within-subject factor is specified, *factor-names* (and associated level and transformation information) must be separated by a comma in the REPEATED statement. These terms are described in the following section, “Syntax Details.”

Syntax Details

You can specify the following terms in the REPEATED statement.

factor-specification

The *factor-specification* for the REPEATED statement can include any number of individual factor specifications, separated by commas, of the following form:

factor-name levels < **(***level-values***)** > < *transformation* >

where

factor-name names a factor to be associated with the dependent variables. The name should not be the same as any variable name that already exists in the data set being analyzed and should conform to the usual conventions of SAS variable names.

When specifying more than one factor, list the dependent variables in the MODEL statement so that the within-subject factors defined in the REPEATED statement are nested; that is, the first factor defined in the REPEATED statement should be the one with values that change least frequently.

levels specifies the number of levels associated with the factor being defined. When there is only one within-subject factor, the number of levels is equal to the number of dependent variables. In this case, *levels* is optional. When more than one within-subject factor is defined, however, *levels* is required, and the product of the number of levels of all the factors must equal the number of dependent variables in the MODEL statement.

(level-values) specifies values that correspond to levels of a repeated-measures factor. These values are used to label output; they are also used as spacings for constructing orthogonal polynomial contrasts if you specify a POLYNOMIAL transformation. The number of level values specified must correspond to the number of levels for that factor in the REPEATED statement. Enclose the *level-values* in parentheses.

The following *transformation* keywords define single-degree-of-freedom contrasts for factors specified in the REPEATED statement. Since the number of contrasts generated is always one less than the number of levels of the factor, you have some control over which contrast is omitted from the analysis by which transformation you select. The only exception is the IDENTITY transformation; this transformation is not composed of contrasts, and it has the same degrees of freedom as the factor has levels. By default, the procedure uses the CONTRAST transformation.

CONTRAST < **(***ordinal-reference-level* **)** > generates contrasts between levels of the factor and a reference level. By default, the procedure uses the last level; you can optionally specify a reference level in parentheses after the keyword CONTRAST. The reference level corresponds to the ordinal value of the level rather than the level value specified. For example, to generate contrasts between the first level of a factor and the other levels, use

```
contrast(1)
```

HELMERT generates contrasts between each level of the factor and the mean of subsequent levels.

IDENTITY generates an identity transformation corresponding to the associated factor. This transformation is *not* composed of contrasts; it has n degrees of freedom for an n-level factor, instead of $n - 1$. This can be used for doubly-multivariate repeated measures.

MEAN < **(***ordinal-reference-level* **)** > generates contrasts between levels of the factor and the mean of all other levels of the factor. Specifying a reference level eliminates the contrast between that level and the mean. Without a reference level, the contrast involving the last level is omitted. See the CONTRAST transformation for an example.

POLYNOMIAL generates orthogonal polynomial contrasts. Level values, if provided, are used as spacings in the construction of the polynomials; otherwise, equal spacing is assumed.

PROFILE generates contrasts between adjacent levels of the factor.

For examples of the transformation matrices generated by these contrast transformations, see the section "Repeated Measures Analysis of Variance" on page 1560 in Chapter 30, "The GLM Procedure."

You can specify the following options in the REPEATED statement after a slash:

CANONICAL

performs a canonical analysis of the $\mathbf{H}$ and $\mathbf{E}$ matrices corresponding to the transformed variables specified in the REPEATED statement.

NOM

displays only the results of the univariate analyses.

NOU

displays only the results of the multivariate analyses.

PRINTE

displays the $\mathbf{E}$ matrix for each combination of within-subject factors, as well as partial correlation matrices for both the original dependent variables and the variables defined by the transformations specified in the REPEATED statement. In addition, the PRINTE option provides sphericity tests for each set of transformed variables. If the requested transformations are not orthogonal, the PRINTE option also provides a sphericity test for a set of orthogonal contrasts.

PRINTH

displays the $\mathbf{H}$ (SSCP) matrix associated with each multivariate test.

PRINTM

displays the transformation matrices that define the contrasts in the analysis. PROC ANOVA always displays the $\mathbf{M}$ matrix so that the transformed variables are defined by the rows, not the columns, of the displayed $\mathbf{M}$ matrix. In other words, PROC ANOVA actually displays $\mathbf{M}'$.

PRINTRV

produces the characteristic roots and vectors for each multivariate test.

SUMMARY

produces analysis-of-variance tables for each contrast defined by the within-subjects factors. Along with tests for the effects of the independent variables specified in the MODEL statement, a term labeled MEAN tests the hypothesis that the overall mean of the contrast is zero.

Examples

When specifying more than one factor, list the dependent variables in the MODEL statement so that the within-subject factors defined in the REPEATED statement are nested; that is, the first factor defined in the REPEATED statement should be the one with values that change least frequently. For example, assume that three treatments are administered at each of four times, for a total of twelve dependent variables on each experimental unit. If the variables are listed in the MODEL statement as Y1 through Y12, then the following REPEATED statement

```
repeated trt 3, time 4;
```

implies the following structure:

	Dependent Variables											
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
Value of trt	1	1	1	1	2	2	2	2	3	3	3	3
Value of time	1	2	3	4	1	2	3	4	1	2	3	4

The REPEATED statement always produces a table like the preceding one. For more information on repeated measures analysis and on using the REPEATED statement, see the section "Repeated Measures Analysis of Variance" on page 1560 in Chapter 30, "The GLM Procedure."

TEST Statement

TEST < **H=** *effects* > **E=** *effect* **;**

Although an F value is computed for all SS in the analysis using the residual MS as an error term, you can request additional F tests using other effects as error terms. You need a TEST statement when a nonstandard error structure (as in a split plot) exists.

Caution: The ANOVA procedure does not check any of the assumptions underlying the F statistic. When you specify a TEST statement, you assume sole responsibility for the validity of the F statistic produced. To help validate a test, you may want to use the GLM procedure with the RANDOM statement and inspect the expected mean squares. In the GLM procedure, you can also use the TEST option in the RANDOM statement.

You can use as many TEST statements as you want, provided that they appear after the MODEL statement.

You can specify the following terms in the TEST statement.

H=*effects* specifies which effects in the preceding model are to be used as hypothesis (numerator) effects.

E=*effect* specifies one, and only one, effect to use as the error (denominator) term. The E= specification is required.

The following example uses two TEST statements and is appropriate for analyzing a split-plot design.

```
proc anova;
   class a b c;
   model y=a|b(a)|c;
   test h=a e=b(a);
   test h=c a*c e=b*c(a);
run;
```

Details

Specification of Effects

In SAS analysis-of-variance procedures, the variables that identify levels of the classifications are called *classification variables*, and they are declared in the CLASS statement. Classification variables are also called *categorical*, *qualitative*, *discrete*, or *nominal variables*. The values of a class variable are called *levels*. Class variables can be either numeric or character. This is in contrast to the *response* (or *dependent*) *variables*, which are continuous. Response variables must be numeric.

The analysis-of-variance model specifies *effects*, which are combinations of classification variables used to explain the variability of the dependent variables in the following manner:

- Main effects are specified by writing the variables by themselves in the CLASS statement: A B C. Main effects used as independent variables test the hypothesis that the mean of the dependent variable is the same for each level of the factor in question, ignoring the other independent variables in the model.
- Crossed effects (interactions) are specified by joining the class variables with asterisks in the MODEL statement: A*B A*C A*B*C. Interaction terms in a model test the hypothesis that the effect of a factor does not depend on the levels of the other factors in the interaction.
- Nested effects are specified by following a main effect or crossed effect with a class variable or list of class variables enclosed in parentheses in the MODEL statement. The main effect or crossed effect is nested within the effects listed in parentheses: B(A) C*D(A B). Nested effects test hypotheses similar to interactions, but the levels of the nested variables are not the same for every combination within which they are nested.

The general form of an effect can be illustrated using the class variables A, B, C, D, E, and F:

$$\mathrm{A} * \mathrm{B} * \mathrm{C}(\mathrm{D\ E\ F})$$

The crossed list should come first, followed by the nested list in parentheses. Note that no asterisks appear within the nested list or immediately before the left parenthesis.

Main Effects Models

For a three-factor main effects model with A, B, and C as the factors and Y as the dependent variable, the necessary statements are

```
proc anova;
   class A B C;
   model Y=A B C;
run;
```

Models with Crossed Factors

To specify interactions in a factorial model, join effects with asterisks as described previously. For example, these statements specify a complete factorial model, which includes all the interactions:

```
proc anova;
   class A B C;
   model Y=A B C A*B A*C B*C A*B*C;
run;
```

Bar Notation

You can shorten the specifications of a full factorial model by using bar notation. For example, the preceding statements can also be written

```
proc anova;
   class A B C;
   model Y=A|B|C;
run;
```

When the bar (|) is used, the expression on the right side of the equal sign is expanded from left to right using the equivalents of rules 2–4 given in Searle (1971, p. 390). The variables on the right- and left-hand sides of the bar become effects, and the cross of them becomes an effect. Multiple bars are permitted. For instance, A | B | C is evaluated as follows:

$$\begin{aligned} \mathrm{A} \mid \mathrm{B} \mid \mathrm{C} \quad &\rightarrow \quad \{\, \mathrm{A} \mid \mathrm{B} \,\} \mid \mathrm{C} \\ &\rightarrow \quad \{\, \mathrm{A}\ \ \mathrm{B}\ \ \mathrm{A{*}B} \,\} \mid \mathrm{C} \\ &\rightarrow \quad \mathrm{A}\ \ \mathrm{B}\ \ \mathrm{A{*}B}\ \ \mathrm{A{*}C}\ \ \mathrm{B{*}C}\ \ \mathrm{A{*}B{*}C} \end{aligned}$$

You can also specify the maximum number of variables involved in any effect that results from bar evaluation by specifying that maximum number, preceded by an @ sign, at the end of the bar effect. For example, the specification A | B | C@2 results in only those effects that contain two or fewer variables; in this case, A B A*B C A*C and B*C.

The following table gives more examples of using the bar and at operators.

<table>
<tr><td>A | C(B)</td><td>is equivalent to</td><td>A C(B) A*C(B)</td></tr>
<tr><td>A(B) | C(B)</td><td>is equivalent to</td><td>A(B) C(B) A*C(B)</td></tr>
<tr><td>A(B) | B(D E)</td><td>is equivalent to</td><td>A(B) B(D E)</td></tr>
<tr><td>A | B(A) | C</td><td>is equivalent to</td><td>A B(A) C A*C B*C(A)</td></tr>
<tr><td>A | B(A) | C@2</td><td>is equivalent to</td><td>A B(A) C A*C</td></tr>
<tr><td>A | B | C | D@2</td><td>is equivalent to</td><td>A B A*B C A*C B*C D A*D B*D C*D</td></tr>
</table>

Consult the “Specification of Effects” section on page 1517 in Chapter 30, “The GLM Procedure,” for further details on bar notation.

Nested Models

Write the effect that is nested within another effect first, followed by the other effect in parentheses. For example, if A and B are main effects and C is nested within A and B (that is, the levels of C that are observed are not the same for each combination of A and B), the statements for PROC ANOVA are

```
proc anova;
   class A B C;
   model y=A B C(A B);
run;
```

The identity of a level is viewed within the context of the level of the containing effects. For example, if City is nested within State, then the identity of City is viewed within the context of State.

The distinguishing feature of a nested specification is that nested effects never appear as main effects. Another way of viewing nested effects is that they are effects that pool the main effect with the interaction of the nesting variable. See the “Automatic Pooling” section, which follows.

Models Involving Nested, Crossed, and Main Effects

Asterisks and parentheses can be combined in the MODEL statement for models involving nested and crossed effects:

```
proc anova;
   class A B C;
   model Y=A B(A) C(A) B*C(A);
run;
```

Automatic Pooling

In line with the general philosophy of the GLM procedure, there is no difference between the statements

```
   model Y=A B(A);
```

and

```
   model Y=A A*B;
```

The effect B becomes a nested effect by virtue of the fact that it does not occur as a main effect. If B is not written as a main effect in addition to participating in A*B, then the sum of squares that is associated with B is pooled into A*B.

This feature allows the automatic pooling of sums of squares. If an effect is omitted from the model, it is automatically pooled with all the higher-level effects containing the class variables in the omitted effect (or within-error). This feature is most useful in split-plot designs.

Using PROC ANOVA Interactively

PROC ANOVA can be used interactively. After you specify a model in a MODEL statement and run PROC ANOVA with a RUN statement, a variety of statements (such as MEANS, MANOVA, TEST, and REPEATED) can be executed without PROC ANOVA recalculating the model sum of squares.

The "Syntax" section (page 348) describes which statements can be used interactively. You can execute these interactive statements individually or in groups by following the single statement or group of statements with a RUN statement. Note that the MODEL statement cannot be repeated; the ANOVA procedure allows only one MODEL statement.

If you use PROC ANOVA interactively, you can end the procedure with a DATA step, another PROC step, an ENDSAS statement, or a QUIT statement. The syntax of the QUIT statement is

```
   quit;
```

When you use PROC ANOVA interactively, additional RUN statements do not end the procedure but tell PROC ANOVA to execute additional statements.

When a WHERE statement is used with PROC ANOVA, it should appear before the first RUN statement. The WHERE statement enables you to select only certain observations for analysis without using a subsetting DATA step. For example, the statement `where group ne 5` omits observations with GROUP=5 from the analysis. Refer to *SAS Language Reference: Dictionary* for details on this statement.

When a BY statement is used with PROC ANOVA, interactive processing is not possible; that is, once the first RUN statement is encountered, processing proceeds for each BY group in the data set, and no further statements are accepted by the procedure.

Interactivity is also disabled when there are different patterns of missing values among the dependent variables. For details, see the section “Missing Values,” which follows.

Missing Values

For an analysis involving one dependent variable, PROC ANOVA uses an observation if values are nonmissing for that dependent variable and for all the variables used in independent effects.

For an analysis involving multiple dependent variables without the MANOVA or REPEATED statement, or without the MANOVA option in the PROC ANOVA statement, a missing value in one dependent variable does not eliminate the observation from the analysis of other nonmissing dependent variables. For an analysis with the MANOVA or REPEATED statement, or with the MANOVA option in the PROC ANOVA statement, the ANOVA procedure requires values for all dependent variables to be nonmissing for an observation before the observation can be used in the analysis.

During processing, PROC ANOVA groups the dependent variables by their pattern of missing values across observations so that sums and cross products can be collected in the most efficient manner.

If your data have different patterns of missing values among the dependent variables, interactivity is disabled. This could occur when some of the variables in your data set have missing values and

- you do not use the MANOVA option in the PROC ANOVA statement
- you do not use a MANOVA or REPEATED statement before the first RUN statement

Output Data Set

The OUTSTAT= option in the PROC ANOVA statement produces an output data set that contains the following:

- the BY variables, if any
- _TYPE_, a new character variable. This variable has the value ‘ANOVA’ for observations corresponding to sums of squares; it has the value ‘CANCORR’, ‘STRUCTUR’, or ‘SCORE’ if a canonical analysis is performed through the MANOVA statement and no M= matrix is specified.
- _SOURCE_, a new character variable. For each observation in the data set, _SOURCE_ contains the name of the model effect from which the corresponding statistics are generated.

- _NAME_, a new character variable. The variable _NAME_ contains the name of one of the dependent variables in the model or, in the case of canonical statistics, the name of one of the canonical variables (CAN1, CAN2, and so on).
- four new numeric variables, SS, DF, F, and PROB, containing sums of squares, degrees of freedom, F values, and probabilities, respectively, for each model or contrast sum of squares generated in the analysis. For observations resulting from canonical analyses, these variables have missing values.
- if there is more than one dependent variable, then variables with the same names as the dependent variables represent
  - for _TYPE_='ANOVA', the crossproducts of the hypothesis matrices
  - for _TYPE_='CANCORR', canonical correlations for each variable
  - for _TYPE_='STRUCTUR', coefficients of the total structure matrix
  - for _TYPE_='SCORE', raw canonical score coefficients

The output data set can be used to perform special hypothesis tests (for example, with the IML procedure in SAS/IML software), to reformat output, to produce canonical variates (through the SCORE procedure), or to rotate structure matrices (through the FACTOR procedure).

Computational Method

Let $\mathbf{X}$ represent the $n \times p$ design matrix. The columns of $\mathbf{X}$ contain only 0s and 1s. Let $\mathbf{Y}$ represent the $n \times 1$ vector of dependent variables.

In the GLM procedure, $\mathbf{X'X}$, $\mathbf{X'Y}$, and $\mathbf{Y'Y}$ are formed in main storage. However, in the ANOVA procedure, only the diagonals of $\mathbf{X'X}$ are computed, along with $\mathbf{X'Y}$ and $\mathbf{Y'Y}$. Thus, PROC ANOVA saves a considerable amount of storage as well as time. The memory requirements for PROC ANOVA are asymptotically linear functions of n^2 and nr, where n is the number of dependent variables and r the number of independent parameters.

The elements of $\mathbf{X'Y}$ are cell totals, and the diagonal elements of $\mathbf{X'X}$ are cell frequencies. Since PROC ANOVA automatically pools omitted effects into the next higher-level effect containing the names of the omitted effect (or within-error), a slight modification to the rules given by Searle (1971, p. 389) is used.

1. PROC ANOVA computes the sum of squares for each effect as if it is a main effect. In other words, for each effect, PROC ANOVA squares each cell total and divides by its cell frequency. The procedure then adds these quantities together and subtracts the correction factor for the mean (total squared over N).
2. For each effect involving two class names, PROC ANOVA subtracts the SS for any main effect with a name that is contained in the two-factor effect.
3. For each effect involving three class names, PROC ANOVA subtracts the SS for all main effects and two-factor effects with names that are contained in the three-factor effect. If effects involving four or more class names are present, the procedure continues this process.

Displayed Output

PROC ANOVA first displays a table that includes the following:

- the name of each variable in the CLASS statement
- the number of different values or Levels of the Class variables
- the Values of the Class variables
- the Number of observations in the data set and the number of observations excluded from the analysis because of missing values, if any

PROC ANOVA then displays an analysis-of-variance table for each dependent variable in the MODEL statement. This table breaks down

- the Total Sum of Squares for the dependent variable into the portion attributed to the Model and the portion attributed to Error
- the Mean Square term, which is the Sum of Squares divided by the degrees of freedom (DF)

The analysis-of-variance table also lists the following:

- the Mean Square for Error (MSE), which is an estimate of σ^2, the variance of the true errors
- the F Value, which is the ratio produced by dividing the Mean Square for the Model by the Mean Square for Error. It tests how well the model as a whole (adjusted for the mean) accounts for the dependent variable's behavior. This F test is a test of the null hypothesis that all parameters except the intercept are zero.
- the significance probability associated with the F statistic, labeled "Pr > F"
- R-Square, R^2, which measures how much variation in the dependent variable can be accounted for by the model. The R^2 statistic, which can range from 0 to 1, is the ratio of the sum of squares for the model divided by the sum of squares for the corrected total. In general, the larger the R^2 value, the better the model fits the data.
- C.V., the coefficient of variation, which is often used to describe the amount of variation in the population. The C.V. is 100 times the standard deviation of the dependent variable divided by the Mean. The coefficient of variation is often a preferred measure because it is unitless.
- Root MSE, which estimates the standard deviation of the dependent variable. Root MSE is computed as the square root of Mean Square for Error, the mean square of the error term.
- the Mean of the dependent variable

For each effect (or source of variation) in the model, PROC ANOVA then displays the following:

- DF, degrees of freedom
- Anova SS, the sum of squares, and the associated Mean Square
- the F Value for testing the hypothesis that the group means for that effect are equal
- Pr > F, the significance probability value associated with the F Value

When you specify a TEST statement, PROC ANOVA displays the results of the requested tests. When you specify a MANOVA statement and the model includes more than one dependent variable, PROC ANOVA produces these additional statistics:

- the characteristic roots and vectors of $\mathbf{E}^{-1}\mathbf{H}$ for each $\mathbf{H}$ matrix
- the Hotelling-Lawley trace
- Pillai’s trace
- Wilks’ criterion
- Roy’s maximum root criterion

See Example 30.6 on page 1600 in Chapter 30, “The GLM Procedure,” for an example of the MANOVA results. These MANOVA tests are discussed in Chapter 3, “Introduction to Regression Procedures.”

ODS Table Names

PROC ANOVA assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, “Using the Output Delivery System.”

Table 17.3. ODS Tables Produced in PROC ANOVA

ODS Table Name	Description	Statement / Option
AltErrTests	Anova tests with error other than MSE	TEST /E=
Bartlett	Bartlett’s homogeneity of variance test	MEANS / HOVTEST=BARTLETT
CLDiffs	Multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and not LINES)
CLDiffsInfo	Information for multiple comparisons of pairwise differences	MEANS / CLDIFF or DUNNETT or (Unequal cells and not LINES)
CLMeans	Multiple comparisons of means with confidence/comparison interval	MEANS / CLM with (BON or GABRIEL or SCHEFFE or SIDAK or SMM or T or LSD)
CLMeansInfo	Information for multiple comparisons of means with confidence/comparison interval	MEANS / CLM

Table 17.3. (continued)

ODS Table Name	Description	Statement / Option
CanAnalysis	Canonical analysis	(MANOVA or REPEATED) / CANONICAL
CanCoefficients	Canonical coefficients	(MANOVA or REPEATED) / CANONICAL
CanStructure	Canonical structure	(MANOVA or REPEATED) / CANONICAL
CharStruct	Characteristic roots and vectors	(MANOVA / not CANONICAL) or (REPEATED / PRINTRV)
ClassLevels	Classification variable levels	CLASS statement
DependentInfo	Simultaneously analyzed dependent variables	default when there are multiple dependent variables with different patterns of missing values
Epsilons	Greenhouse-Geisser and Huynh-Feldt epsilons	REPEATED statement
ErrorSSCP	Error SSCP matrix	(MANOVA or REPEATED) / PRINTE
FitStatistics	R-Square, C.V., Root MSE, and dependent mean	default
HOVFTest	Homogeneity of variance ANOVA	MEANS / HOVTEST
HypothesisSSCP	Hypothesis SSCP matrix	(MANOVA or REPEATED) / PRINTE
MANOVATransform	Multivariate transformation matrix	MANOVA / M=
MCLines	Multiple comparisons LINES output	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not(CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MCLinesInfo	Information for multiple comparison LINES output	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not (CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MCLinesRange	Ranges for multiple range MC tests	MEANS / LINES or ((DUNCAN or WALLER or SNK or REGWQ) and not (CLDIFF or CLM)) or (Equal cells and not CLDIFF)
MTests	Multivariate tests	MANOVA statement
Means	Group means	MEANS statement
ModelANOVA	ANOVA for model terms	default
NObs	Number of observations	default
OverallANOVA	Over-all ANOVA	default
PartialCorr	Partial correlation matrix	(MANOVA or REPEATED) / PRINTE
RepTransform	Repeated transformation matrix	REPEATED (CONTRAST or HELMERT or MEAN or POLYNOMIAL or PROFILE)

Table 17.3. (continued)

ODS Table Name	Description	Statement / Option
RepeatedLevelInfo	Correspondence between dependents and repeated measures levels	REPEATED statement
Sphericity	Sphericity tests	REPEATED / PRINTE
Tests	Summary ANOVA for specified MANOVA H= effects	MANOVA / H= SUMMARY
Welch	Welch's ANOVA	MEANS / WELCH

Examples

Example 17.1. Randomized Complete Block With Factorial Treatment Structure

This example uses statements for the analysis of a randomized block with two treatment factors occuring in a factorial structure. The data, from Neter, Wasserman, and Kutner (1990, p. 941), are from an experiment examining the effects of codeine and acupuncture on post-operative dental pain in male subjects. Both treatment factors have two levels. The codeine levels are a codeine capsule or a sugar capsule. The acupuncture levels are two inactive acupuncture points or two active acupuncture points. There are four distinct treatment combinations due to the factorial treatment structure. The 32 subjects are assigned to eight blocks of four subjects each based on an assessment of pain tolerance.

The data for the analysis are balanced, so PROC ANOVA is used. The data are as follows:

```
title 'Randomized Complete Block With Two Factors';
data PainRelief;
   input PainLevel Codeine Acupuncture Relief @@;
   datalines;
1 1 1 0.0  1 2 1 0.5  1 1 2 0.6  1 2 2 1.2
2 1 1 0.3  2 2 1 0.6  2 1 2 0.7  2 2 2 1.3
3 1 1 0.4  3 2 1 0.8  3 1 2 0.8  3 2 2 1.6
4 1 1 0.4  4 2 1 0.7  4 1 2 0.9  4 2 2 1.5
5 1 1 0.6  5 2 1 1.0  5 1 2 1.5  5 2 2 1.9
6 1 1 0.9  6 2 1 1.4  6 1 2 1.6  6 2 2 2.3
7 1 1 1.0  7 2 1 1.8  7 1 2 1.7  7 2 2 2.1
8 1 1 1.2  8 2 1 1.7  8 1 2 1.6  8 2 2 2.4
;
```

The variable PainLevel is the blocking variable, and Codeine and Acupuncture represent the levels of the two treatment factors. The variable Relief is the pain relief score (the higher the score, the more relief the patient has).

The following code invokes PROC ANOVA. The blocking variable and treatment factors appear in the CLASS statement. The bar between the treatment factors Codeine and Acupuncture adds their main effects as well as their interaction Codeine*Acupuncture to the model.

```
proc anova;
   class PainLevel Codeine Acupuncture;
   model Relief = PainLevel Codeine|Acupuncture;
run;
```

The results from the analysis are shown in Output 17.1.1 and Output 17.1.2.

Output 17.1.1. Class Level Information and ANOVA Table

```
                   Randomized Complete Block With Two Factors

                              The ANOVA Procedure

                            Class Level Information

                  Class            Levels    Values

                  PainLevel             8    1 2 3 4 5 6 7 8

                  Codeine               2    1 2

                  Acupuncture           2    1 2


                          Number of observations    32
```

```
                   Randomized Complete Block With Two Factors

                              The ANOVA Procedure

Dependent Variable: Relief

                                       Sum of
 Source                      DF       Squares    Mean Square   F Value   Pr > F

 Model                       10    11.33500000     1.13350000     78.37   <.0001

 Error                       21     0.30375000     0.01446429

 Corrected Total             31    11.63875000


              R-Square     Coeff Var      Root MSE    Relief Mean

              0.973902      10.40152      0.120268       1.156250
```

The Class Level Information and ANOVA table are shown in Output 17.1.1. The class level information summarizes the structure of the design. It is good to check these consistently in search of errors in the data step. The overall F test is significant, indicating that the model accounts for a significant amount of variation in the dependent variable.

Output 17.1.2. Tests of Effects

```
                   Randomized Complete Block With Two Factors

                              The ANOVA Procedure

Dependent Variable: Relief

 Source                     DF       Anova SS    Mean Square   F Value   Pr > F

 PainLevel                   7     5.59875000     0.79982143     55.30   <.0001
 Codeine                     1     2.31125000     2.31125000    159.79   <.0001
 Acupuncture                 1     3.38000000     3.38000000    233.68   <.0001
 Codeine*Acupuncture         1     0.04500000     0.04500000      3.11   0.0923
```

Output 17.1.2 shows tests of the effects. The blocking effect is significant; hence, it is useful. The interaction between codeine and acupuncture is significant at the 90% level but not at the 95% level. The significance level of this test should be determined before the analysis. The main effects of both treatment factors are highly significant.

Example 17.2. Alternative Multiple Comparison Procedures

The following is a continuation of the first example in the the "One-Way Layout with Means Comparisons" section on page 340. You are studying the effect of bacteria on the nitrogen content of red clover plants, and the analysis of variance shows a highly significant effect. The following statements create the data set and compute the analysis of variance as well as Tukey's multiple comparisons test for pairwise differences between bacteria strains; the results are shown in Figure 17.1, Figure 17.2, and Figure 17.3

```
title 'Nitrogen Content of Red Clover Plants';
data Clover;
   input Strain $ Nitrogen @@;
   datalines;
3DOK1  19.4 3DOK1  32.6 3DOK1  27.0 3DOK1  32.1 3DOK1  33.0
3DOK5  17.7 3DOK5  24.8 3DOK5  27.9 3DOK5  25.2 3DOK5  24.3
3DOK4  17.0 3DOK4  19.4 3DOK4   9.1 3DOK4  11.9 3DOK4  15.8
3DOK7  20.7 3DOK7  21.0 3DOK7  20.5 3DOK7  18.8 3DOK7  18.6
3DOK13 14.3 3DOK13 14.4 3DOK13 11.8 3DOK13 11.6 3DOK13 14.2
COMPOS 17.3 COMPOS 19.4 COMPOS 19.1 COMPOS 16.9 COMPOS 20.8
;

proc anova;
   class Strain;
   model Nitrogen = Strain;
   means Strain / tukey;
run;
```

The interactivity of PROC ANOVA enables you to submit further MEANS statements without re-running the entire analysis. For example, the following command requests means of the Strain levels with Duncan's multiple range test and the Waller-Duncan k-ratio t test.

```
      means Strain / duncan waller;
   run;
```

Results of the Waller-Duncan k-ratio t test are shown in Output 17.2.1.

Output 17.2.1. Waller-Duncan K-ratio t Test

```
                     Nitrogen Content of Red Clover Plants

                              The ANOVA Procedure

                   Waller-Duncan K-ratio t Test for Nitrogen

NOTE: This test minimizes the Bayes risk under additive loss and certain other
                                  assumptions.


                    Kratio                              100
                    Error Degrees of Freedom             24
                    Error Mean Square              11.78867
                    F Value                           14.37
                    Critical Value of t             1.91873
                    Minimum Significant Difference   4.1665


          Means with the same letter are not significantly different.


            Waller Grouping          Mean      N    Strain

                          A        28.820      5    3DOK1

                          B        23.980      5    3DOK5
                          B
                     C    B        19.920      5    3DOK7
                     C
                     C    D        18.700      5    COMPOS
                          D
                     E    D        14.640      5    3DOK4
                     E
                     E             13.260      5    3DOK13
```

The Waller-Duncan k-ratio t test is a multiple range test. Unlike Tukey's test, this test does not operate on the principle of controlling Type I error. Instead, it compares the Type I and Type II error rates based on Bayesian principles (Steel and Torrie 1980).

The Waller Grouping column in Output 17.2.1 shows which means are significantly different. From this test, you can conclude the following:

- The mean nitrogen content for strain 3DOK1 is higher than the means for all other strains.
- The mean nitrogen content for strain 3DOK5 is higher than the means for COMPOS, 3DOK4, and 3DOK13.
- The mean nitrogen content for strain 3DOK7 is higher than the means for 3DOK4 and 3DOK13.
- The mean nitrogen content for strain COMPOS is higher than the mean for 3DOK13.

- Differences between all other means are not significant based on this sample size.

Output 17.2.2 shows the results of Duncan's multiple range test. Duncan's test is a result-guided test that compares the treatment means while controlling the comparison-wise error rate. You should use this test for planned comparisons only (Steel and Torrie 1980). The results and conclusions for this example are the same as for the Waller-Duncan k-ratio t test. This is not always the case.

Output 17.2.2. Duncan's Multiple Range Test

```
                     Nitrogen Content of Red Clover Plants

                              The ANOVA Procedure

                   Duncan's Multiple Range Test for Nitrogen

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


                       Alpha                             0.05
                       Error Degrees of Freedom            24
                       Error Mean Square             11.78867


   Number of Means          2          3          4          5          6
   Critical Range       4.482      4.707      4.852      4.954      5.031


        Means with the same letter are not significantly different.


          Duncan Grouping          Mean      N    Strain

                        A        28.820      5    3DOK1

                        B        23.980      5    3DOK5
                        B
                   C    B        19.920      5    3DOK7
                   C
                   C    D        18.700      5    COMPOS
                        D
                   E    D        14.640      5    3DOK4
                   E
                   E             13.260      5    3DOK13
```

Tukey and Least Significant Difference (LSD) tests are requested with the following MEANS statement. The CLDIFF option requests confidence intervals for both tests.

```
   means Strain / lsd tukey cldiff;
run;
```

The LSD tests for this example are shown in Output 17.2.3, and they give the same results as the previous two multiple comparison tests. Again, this is not always the case.

Output 17.2.3. T Tests (LSD)

```
                     Nitrogen Content of Red Clover Plants

                              The ANOVA Procedure

                           t Tests (LSD) for Nitrogen

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


                     Alpha                              0.05
                     Error Degrees of Freedom             24
                     Error Mean Square              11.78867
                     Critical Value of t             2.06390
                     Least Significant Difference     4.4818


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


                               Difference
              Strain             Between      95% Confidence
            Comparison            Means           Limits

            3DOK1  - 3DOK5         4.840       0.358    9.322  ***
            3DOK1  - 3DOK7         8.900       4.418   13.382  ***
            3DOK1  - COMPOS       10.120       5.638   14.602  ***
            3DOK1  - 3DOK4        14.180       9.698   18.662  ***
            3DOK1  - 3DOK13       15.560      11.078   20.042  ***
            3DOK5  - 3DOK1        -4.840      -9.322   -0.358  ***
            3DOK5  - 3DOK7         4.060      -0.422    8.542
            3DOK5  - COMPOS        5.280       0.798    9.762  ***
            3DOK5  - 3DOK4         9.340       4.858   13.822  ***
            3DOK5  - 3DOK13       10.720       6.238   15.202  ***
            3DOK7  - 3DOK1        -8.900     -13.382   -4.418  ***
            3DOK7  - 3DOK5        -4.060      -8.542    0.422
            3DOK7  - COMPOS        1.220      -3.262    5.702
            3DOK7  - 3DOK4         5.280       0.798    9.762  ***
            3DOK7  - 3DOK13        6.660       2.178   11.142  ***
            COMPOS - 3DOK1       -10.120     -14.602   -5.638  ***
            COMPOS - 3DOK5        -5.280      -9.762   -0.798  ***
            COMPOS - 3DOK7        -1.220      -5.702    3.262
            COMPOS - 3DOK4         4.060      -0.422    8.542
            COMPOS - 3DOK13        5.440       0.958    9.922  ***
            3DOK4  - 3DOK1       -14.180     -18.662   -9.698  ***
            3DOK4  - 3DOK5        -9.340     -13.822   -4.858  ***
            3DOK4  - 3DOK7        -5.280      -9.762   -0.798  ***
            3DOK4  - COMPOS       -4.060      -8.542    0.422
            3DOK4  - 3DOK13        1.380      -3.102    5.862
            3DOK13 - 3DOK1       -15.560     -20.042  -11.078  ***
            3DOK13 - 3DOK5       -10.720     -15.202   -6.238  ***
            3DOK13 - 3DOK7        -6.660     -11.142   -2.178  ***
            3DOK13 - COMPOS       -5.440      -9.922   -0.958  ***
            3DOK13 - 3DOK4        -1.380      -5.862    3.102
```

If you only perform the LSD tests when the overall model F-test is significant, then this is called Fisher’s protected LSD test. Note that the LSD tests should be used for planned comparisons.

The TUKEY tests shown in Output 17.2.4 find fewer significant differences than the other three tests. This is not unexpected, as the TUKEY test controls the Type I experimentwise error rate. For a complete discussion of multiple comparison methods, see the “Multiple Comparisons” section on page 1540 in Chapter 30, “The GLM Procedure.”

Output 17.2.4. Tukey’s Studentized Range Test

```
                Nitrogen Content of Red Clover Plants

                         The ANOVA Procedure

           Tukey’s Studentized Range (HSD) Test for Nitrogen

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


                Alpha                                   0.05
                Error Degrees of Freedom                  24
                Error Mean Square                   11.78867
                Critical Value of Studentized Range  4.37265
                Minimum Significant Difference        6.7142


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


                             Difference
            Strain              Between     Simultaneous 95%
          Comparison              Means    Confidence Limits

        3DOK1  - 3DOK5            4.840      -1.874   11.554
        3DOK1  - 3DOK7            8.900       2.186   15.614  ***
        3DOK1  - COMPOS          10.120       3.406   16.834  ***
        3DOK1  - 3DOK4           14.180       7.466   20.894  ***
        3DOK1  - 3DOK13          15.560       8.846   22.274  ***
        3DOK5  - 3DOK1           -4.840     -11.554    1.874
        3DOK5  - 3DOK7            4.060      -2.654   10.774
        3DOK5  - COMPOS           5.280      -1.434   11.994
        3DOK5  - 3DOK4            9.340       2.626   16.054  ***
        3DOK5  - 3DOK13          10.720       4.006   17.434  ***
        3DOK7  - 3DOK1           -8.900     -15.614   -2.186  ***
        3DOK7  - 3DOK5           -4.060     -10.774    2.654
        3DOK7  - COMPOS           1.220      -5.494    7.934
        3DOK7  - 3DOK4            5.280      -1.434   11.994
        3DOK7  - 3DOK13           6.660      -0.054   13.374
        COMPOS - 3DOK1          -10.120     -16.834   -3.406  ***
        COMPOS - 3DOK5           -5.280     -11.994    1.434
        COMPOS - 3DOK7           -1.220      -7.934    5.494
        COMPOS - 3DOK4            4.060      -2.654   10.774
        COMPOS - 3DOK13           5.440      -1.274   12.154
        3DOK4  - 3DOK1          -14.180     -20.894   -7.466  ***
        3DOK4  - 3DOK5           -9.340     -16.054   -2.626  ***
        3DOK4  - 3DOK7           -5.280     -11.994    1.434
        3DOK4  - COMPOS          -4.060     -10.774    2.654
        3DOK4  - 3DOK13           1.380      -5.334    8.094
        3DOK13 - 3DOK1          -15.560     -22.274   -8.846  ***
        3DOK13 - 3DOK5          -10.720     -17.434   -4.006  ***
        3DOK13 - 3DOK7           -6.660     -13.374    0.054
        3DOK13 - COMPOS          -5.440     -12.154    1.274
        3DOK13 - 3DOK4           -1.380      -8.094    5.334
```

Example 17.3. Split Plot

In some experiments, treatments can be applied only to groups of experimental observations rather than separately to each observation. When there are two nested groupings of the observations on the basis of treatment application, this is known as a *split plot design*. For example, in integrated circuit fabrication it is of interest to see how different manufacturing methods affect the characteristics of individual chips. However, much of the manufacturing process is applied to a relatively large wafer of material, from which many chips are made. Additionally, a chip's position within a wafer may also affect chip performance. These two groupings of chips—by wafer and by position-within-wafer—might form the *whole plots* and the *subplots*, respectively, of a split plot design for integrated circuits.

The following statements produce an analysis for a split-plot design. The CLASS statement includes the variables Block, A, and B, where B defines subplots within BLOCK*A whole plots. The MODEL statement includes the independent effects Block, A, Block*A, B, and A*B. The TEST statement asks for an F test of the A effect, using the Block*A effect as the error term. The following statements produce Output 17.3.1 and Output 17.3.2:

```
title 'Split Plot Design';
data Split;
   input Block 1 A 2 B 3 Response;
   datalines;
142 40.0
141 39.5
112 37.9
111 35.4
121 36.7
122 38.2
132 36.4
131 34.8
221 42.7
222 41.6
212 40.3
211 41.6
241 44.5
242 47.6
231 43.6
232 42.8
;

proc anova;
   class Block A B;
   model Response = Block A Block*A B A*B;
   test h=A e=Block*A;
run;
```

Output 17.3.1. Class Level Information and ANOVA Table

```
                              Split Plot Design

                             The ANOVA Procedure

                           Class Level Information

                         Class         Levels    Values

                         Block              2    1 2

                         A                  4    1 2 3 4

                         B                  2    1 2


                           Number of observations    16
```

```
                              Split Plot Design

                             The ANOVA Procedure

Dependent Variable: Response

                                            Sum of
 Source                      DF          Squares    Mean Square   F Value   Pr > F

 Model                       11      182.0200000     16.5472727      7.85   0.0306

 Error                        4        8.4300000      2.1075000

 Corrected Total             15      190.4500000


             R-Square     Coeff Var      Root MSE    Response Mean

             0.955736      3.609007      1.451723         40.22500
```

First, notice that the overall F test for the model is significant.

Output 17.3.2. Tests of Effects

```
                              Split Plot Design

                             The ANOVA Procedure

Dependent Variable: Response

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Block                        1    131.1025000    131.1025000     62.21   0.0014
 A                            3     40.1900000     13.3966667      6.36   0.0530
 Block*A                      3      6.9275000      2.3091667      1.10   0.4476
 B                            1      2.2500000      2.2500000      1.07   0.3599
 A*B                          3      1.5500000      0.5166667      0.25   0.8612


      Tests of Hypotheses Using the Anova MS for Block*A as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 A                            3    40.19000000    13.39666667      5.80   0.0914
```

The effect of Block is significant. The effect of A is not significant: look at the F test produced by the TEST statement, not at the F test produced by default. Neither the B nor A*B effects are significant. The test for Block*A is irrelevant, as this is simply the main-plot error.

Example 17.4. Latin Square Split Plot

The data for this example is taken from Smith (1951). A Latin square design is used to evaluate six different sugar beet varieties arranged in a six-row (Rep) by six-column (Column) square. The data are collected over two harvests. The variable Harvest then becomes a split plot on the original Latin square design for whole plots. The following statements produce Output 17.4.1 and Output 17.4.2:

```
title 'Sugar Beet Varieties';
title3 'Latin Square Split-Plot Design';
data Beets;
   do Harvest=1 to 2;
      do Rep=1 to 6;
         do Column=1 to 6;
            input Variety Y @;
            output;
            end;
         end;
      end;
   datalines;
3 19.1 6 18.3 5 19.6 1 18.6 2 18.2 4 18.5
6 18.1 2 19.5 4 17.6 3 18.7 1 18.7 5 19.9
1 18.1 5 20.2 6 18.5 4 20.1 3 18.6 2 19.2
2 19.1 3 18.8 1 18.7 5 20.2 4 18.6 6 18.5
4 17.5 1 18.1 2 18.7 6 18.2 5 20.4 3 18.5
5 17.7 4 17.8 3 17.4 2 17.0 6 17.6 1 17.6
3 16.2 6 17.0 5 18.1 1 16.6 2 17.7 4 16.3
```

```
6 16.0 2 15.3 4 16.0 3 17.1 1 16.5 5 17.6
1 16.5 5 18.1 6 16.7 4 16.2 3 16.7 2 17.3
2 17.5 3 16.0 1 16.4 5 18.0 4 16.6 6 16.1
4 15.7 1 16.1 2 16.7 6 16.3 5 17.8 3 16.2
5 18.3 4 16.6 3 16.4 2 17.6 6 17.1 1 16.5
;

proc anova;
   class Column Rep Variety Harvest;
   model Y=Rep Column Variety Rep*Column*Variety
           Harvest Harvest*Rep
           Harvest*Variety;
   test h=Rep Column Variety e=Rep*Column*Variety;
   test h=Harvest            e=Harvest*Rep;
run;
```

Output 17.4.1. Class Level Information and ANOVA Table

```
        Sugar Beet Varieties

   Latin Square Split-Plot Design

        The ANOVA Procedure

      Class Level Information

Class         Levels    Values

Column             6    1 2 3 4 5 6

Rep                6    1 2 3 4 5 6

Variety            6    1 2 3 4 5 6

Harvest            2    1 2


    Number of observations    72
```

```
                              Sugar Beet Varieties

                         Latin Square Split-Plot Design

                              The ANOVA Procedure

Dependent Variable: Y

                                         Sum of
 Source                      DF         Squares    Mean Square   F Value   Pr > F

 Model                       46      98.9147222      2.1503200      7.22   <.0001

 Error                       25       7.4484722      0.2979389

 Corrected Total             71     106.3631944


               R-Square     Coeff Var      Root MSE        Y Mean

               0.929971      3.085524      0.545838      17.69028


 Source                      DF        Anova SS    Mean Square   F Value   Pr > F

 Rep                          5      4.32069444     0.86413889      2.90   0.0337
 Column                       5      1.57402778     0.31480556      1.06   0.4075
 Variety                      5     20.61902778     4.12380556     13.84   <.0001
 Column*Rep*Variety          20      3.25444444     0.16272222      0.55   0.9144
 Harvest                      1     60.68347222    60.68347222    203.68   <.0001
 Rep*Harvest                  5      7.71736111     1.54347222      5.18   0.0021
 Variety*Harvest              5      0.74569444     0.14913889      0.50   0.7729
```

First, note from Output 17.4.1 that the overall model is significant.

Output 17.4.2. Tests of Effects

```
                              Sugar Beet Varieties

                         Latin Square Split-Plot Design

                              The ANOVA Procedure

Dependent Variable: Y

 Tests of Hypotheses Using the Anova MS for Column*Rep*Variety as an Error Term

 Source                     DF       Anova SS    Mean Square   F Value   Pr > F

 Rep                         5     4.32069444     0.86413889      5.31   0.0029
 Column                      5     1.57402778     0.31480556      1.93   0.1333
 Variety                     5    20.61902778     4.12380556     25.34   <.0001


    Tests of Hypotheses Using the Anova MS for Rep*Harvest as an Error Term

 Source                     DF       Anova SS    Mean Square   F Value   Pr > F

 Harvest                     1    60.68347222    60.68347222     39.32   0.0015
```

Output 17.4.2 shows that the effects for Rep and Harvest are significant, while the Column effect is not. The average Ys for the six different Varietys are significantly different. For these four tests, look at the output produced by the two TEST statements, not at the usual ANOVA procedure output. The Variety*Harvest interaction is not significant. All other effects in the default output should either be tested using the results from the TEST statements or are irrelevant as they are only error terms for portions of the model.

Example 17.5. Strip-Split Plot

In this example, four different fertilizer treatments are laid out in vertical strips, which are then split into subplots with different levels of calcium. Soil type is stripped across the split-plot experiment, and the entire experiment is then replicated three times. The dependent variable is the yield of winter barley. The data come from the notes of G. Cox and A. Rotti.

The input data are the 96 values of Y, arranged so that the calcium value (Calcium) changes most rapidly, then the fertilizer value (Fertilizer), then the Soil value, and, finally, the Rep value. Values are shown for Calcium (0 and 1); Fertilizer (0, 1, 2, 3); Soil (1, 2, 3); and Rep (1, 2, 3, 4). The following example produces Output 17.5.1, Output 17.5.2, and Output 17.5.3.

```
title 'Strip-split Plot';
data Barley;
   do Rep=1 to 4;
      do Soil=1 to 3;                   /* 1=d 2=h 3=p */
         do Fertilizer=0 to 3;
            do Calcium=0,1;
               input Yield @;
               output;
            end;
         end;
      end;
   end;
   datalines;
4.91 4.63 4.76 5.04 5.38 6.21 5.60 5.08
4.94 3.98 4.64 5.26 5.28 5.01 5.45 5.62
5.20 4.45 5.05 5.03 5.01 4.63 5.80 5.90
6.00 5.39 4.95 5.39 6.18 5.94 6.58 6.25
5.86 5.41 5.54 5.41 5.28 6.67 6.65 5.94
5.45 5.12 4.73 4.62 5.06 5.75 6.39 5.62
4.96 5.63 5.47 5.31 6.18 6.31 5.95 6.14
5.71 5.37 6.21 5.83 6.28 6.55 6.39 5.57
4.60 4.90 4.88 4.73 5.89 6.20 5.68 5.72
5.79 5.33 5.13 5.18 5.86 5.98 5.55 4.32
5.61 5.15 4.82 5.06 5.67 5.54 5.19 4.46
5.13 4.90 4.88 5.18 5.45 5.80 5.12 4.42
;

proc anova;
   class Rep Soil Calcium Fertilizer;
   model Yield =
           Rep
           Fertilizer Fertilizer*Rep
           Calcium Calcium*Fertilizer
                   Calcium*Rep(Fertilizer)
           Soil Soil*Rep
           Soil*Fertilizer Soil*Rep*Fertilizer
           Soil*Calcium Soil*Fertilizer*Calcium
           Soil*Calcium*Rep(Fertilizer);
   test h=Fertilizer         e=Fertilizer*Rep;
   test h=Calcium
          Calcium*Fertilizer e=Calcium*Rep(Fertilizer);
   test h=Soil               e=Soil*Rep;
   test h=Soil*Fertilizer    e=Soil*Rep*Fertilizer;
   test h=Soil*Calcium
          Soil*Fertilizer*Calcium
                             e=Soil*Calcium*Rep(Fertilizer);
   means Fertilizer Calcium Soil Calcium*Fertilizer;
run;
```

Output 17.5.1. Class Level Information and ANOVA Table

```
                                Strip-split Plot

                              The ANOVA Procedure

                            Class Level Information

                       Class           Levels    Values

                       Rep                  4    1 2 3 4

                       Soil                 3    1 2 3

                       Calcium              2    0 1

                       Fertilizer           4    0 1 2 3


                          Number of observations    96
```

```
                                Strip-split Plot

                              The ANOVA Procedure

Dependent Variable: Yield

                                       Sum of
 Source                     DF        Squares    Mean Square   F Value   Pr > F

 Model                      95    31.89149583     0.33569996        .      .

 Error                       0     0.00000000      .

 Corrected Total            95    31.89149583


               R-Square     Coeff Var      Root MSE    Yield Mean

               1.000000           .               .      5.427292


 Source                     DF       Anova SS    Mean Square   F Value   Pr > F

 Rep                         3     6.27974583     2.09324861        .      .
 Fertilizer                  3     7.22127083     2.40709028        .      .
 Rep*Fertilizer              9     6.08211250     0.67579028        .      .
 Calcium                     1     0.27735000     0.27735000        .      .
 Calcium*Fertilizer          3     1.96395833     0.65465278        .      .
 Rep*Calcium(Fertili)       12     1.76705833     0.14725486        .      .
 Soil                        2     1.92658958     0.96329479        .      .
 Rep*Soil                    6     1.66761042     0.27793507        .      .
 Soil*Fertilizer             6     0.68828542     0.11471424        .      .
 Rep*Soil*Fertilizer        18     1.58698125     0.08816563        .      .
 Soil*Calcium                2     0.04493125     0.02246562        .      .
 Soil*Calcium*Fertili        6     0.18936042     0.03156007        .      .
 Rep*Soil*Calc(Ferti)       24     2.19624167     0.09151007        .      .
```

As the model is completely specified by the MODEL statement, the entire top portion of output (Output 17.5.1) should be ignored. Look at the following output produced by the various TEST statements.

Output 17.5.2. Tests of Effects

```
                                Strip-split Plot

                              The ANOVA Procedure

Dependent Variable: Yield

   Tests of Hypotheses Using the Anova MS for Rep*Fertilizer as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Fertilizer                   3     7.22127083     2.40709028      3.56   0.0604


                   Tests of Hypotheses Using the Anova MS for
                     Rep*Calcium(Fertili) as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Calcium                      1     0.27735000     0.27735000      1.88   0.1950
 Calcium*Fertilizer           3     1.96395833     0.65465278      4.45   0.0255


      Tests of Hypotheses Using the Anova MS for Rep*Soil as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Soil                         2     1.92658958     0.96329479      3.47   0.0999


                   Tests of Hypotheses Using the Anova MS for
                      Rep*Soil*Fertilizer as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Soil*Fertilizer              6     0.68828542     0.11471424      1.30   0.3063


                   Tests of Hypotheses Using the Anova MS for
                     Rep*Soil*Calc(Ferti) as an Error Term

 Source                      DF       Anova SS    Mean Square   F Value   Pr > F

 Soil*Calcium                 2     0.04493125     0.02246562      0.25   0.7843
 Soil*Calcium*Fertili         6     0.18936042     0.03156007      0.34   0.9059
```

The only significant effect is the Calcium*Fertilizer interaction.

Output 17.5.3. Results of MEANS statement

```
                              Strip-split Plot

                            The ANOVA Procedure

              Level of              ------------Yield------------
              Fertilizer      N             Mean          Std Dev

              0              24       5.18416667       0.48266395
              1              24       5.12916667       0.38337082
              2              24       5.75458333       0.53293265
              3              24       5.64125000       0.63926801


               Level of             ------------Yield------------
               Calcium       N              Mean          Std Dev

               0            48       5.48104167       0.54186141
               1            48       5.37354167       0.61565219


               Level of             ------------Yield------------
               Soil          N              Mean          Std Dev

               1            32       5.54312500       0.55806369
               2            32       5.51093750       0.62176315
               3            32       5.22781250       0.51825224


        Level of     Level of              ------------Yield------------
        Calcium      Fertilizer      N             Mean          Std Dev

        0            0              12       5.34666667       0.45029956
        0            1              12       5.08833333       0.44986530
        0            2              12       5.62666667       0.44707806
        0            3              12       5.86250000       0.52886027
        1            0              12       5.02166667       0.47615569
        1            1              12       5.17000000       0.31826233
        1            2              12       5.88250000       0.59856077
        1            3              12       5.42000000       0.68409197
```

The final portion of output shows the results of the MEANS statement. This portion shows means for various effects and combinations of effects, as requested. Because no multiple comparison procedures are requested, none are performed. You can examine the Calcium*Fertilizer means to understand the interaction better.

In this example, you could reduce memory requirements by omitting the Soil*Calcium*Rep(Fertilizer) effect from the model in the MODEL statement. This effect then becomes the ERROR effect, and you can omit the last TEST statement (in the code shown earlier). The test for the Soil*Calcium effect is then given in the Analysis of Variance table in the top portion of output. However, for all other tests, you should look at the results from the TEST statement. In large models, this method may lead to significant reductions in memory requirements.

References

Bartlett, M.S. (1937), “Properties of Sufficiency and Statistical Tests,” *Proceedings of the Royal Society of London, Series A* 160, 268–282.

Brown, M.B. and Forsythe, A.B. (1974), “Robust Tests for Equality of Variances,” *Journal of the American Statistical Association,* 69, 364–367.

Erdman, L.W. (1946), “Studies to Determine if Antibiosis Occurs among Rhizobia,” *Journal of the American Society of Agronomy,* 38, 251–258.

Fisher, R.A. (1942), *The Design of Experiments,* Third Edition, Edinburgh: Oliver & Boyd.

Freund, R.J., Littell, R.C., and Spector, P.C. (1986), *SAS System for Linear Models, 1986 Edition*, Cary, NC: SAS Institute Inc.

Graybill, F.A. (1961), *An Introduction to Linear Statistical Models*, Volume I, New York: McGraw-Hill Book Co.

Henderson, C.R. (1953), “Estimation of Variance and Covariance Components,” *Biometrics*, 9, 226–252.

Levene, H. (1960), “Robust Tests for the Equality of Variance,” in *Contributions to Probability and Statistics,* ed. I. Olkin, Palo Alto, CA: Stanford University Press, 278–292.

Neter, J., Wasserman, W., and Kutner, M.H. (1990), *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, Homewood, IL: Richard D. Irwin, Inc.

O’Brien, R.G. (1979), “A General ANOVA Method for Robust Tests of Additive Models for Variances,” *Journal of the American Statistical Association*, 74, 877–880.

O’Brien, R.G. (1981), “A Simple Test for Variance Effects in Experimental Designs,” *Psychological Bulletin,* 89(3), 570–574.

Remington, R.D. and Schork, M.A. (1970), *Statistics with Applications to the Biological and Health Sciences*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons, Inc.

Schlotzhauer, S.D. and Littell, R.C. (1987), *SAS System for Elementary Statistical Analysis*, Cary, NC: SAS Institute Inc.

Searle, S.R. (1971), *Linear Models*, New York: John Wiley & Sons, Inc.

Smith, W.G. (1951), Dissertation Notes on Canadian Sugar Factories, Ltd., Alberta, Canada: Taber.

Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*, Sixth Edition, Ames, IA: Iowa State University Press.

Steel, R.G.D. and Torrie, J.H. (1980), *Principles and Procedures of Statistics*, New York: McGraw-Hill Book Co.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.