SAS/STAT® 9.22 User's Guide
The NESTED Procedure
(Book Excerpt)

Chapter 59

The NESTED Procedure

Contents

Overview: NESTED Procedure

The NESTED procedure performs random-effects analysis of variance for data from an experiment with a nested (hierarchical) structure.[1] A random-effects model for data from a completely nested design with two factors has the form

$$y_{ijr} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijr}$$

where

---

[1]PROC NESTED is modeled after the General Purpose Nested Analysis of Variance program of the Dairy Cattle Research Branch of the United States Department of Agriculture. That program was originally written by M. R. Swanson, Statistical Reporting Service, United States Department of Agriculture.

y_{ijr} is the value of the dependent variable observed at the rth replication with the first factor at its ith level and the second factor at its jth level.

μ is the overall (fixed) mean of the sampling population.

$\alpha_i, \beta_{ij}, \epsilon_{ijr}$ are mutually uncorrelated random effects with zero means and respective variances $\sigma_\alpha^2, \sigma_\beta^2$, and σ_ϵ^2 (the variance components).

This model is appropriate for an experiment with a multistage nested sampling design. An example of this is given in Example 59.1, where four turnip plants are randomly chosen (the first factor), then three leaves are randomly chosen from each plant (the second factor nested within the first), and then two samples are taken from each leaf (the different replications at fixed levels of the two factors).

Note that PROC NESTED is appropriate for models with only classification effects; it does not handle models that contain continuous covariates. For random effects models with covariates, use either the GLM or MIXED procedure.

Contrasted with Other SAS Procedures

The NESTED procedure performs a computationally efficient analysis of variance for data with a nested design, estimating the different components of variance and also testing for their significance if the design is balanced (see the section "Unbalanced Data" on page 4883). Although other procedures (such as GLM and MIXED) provide similar analyzes, PROC NESTED is both easier to use and computationally more efficient for this special type of design. This is especially true when the design involves a large number of factors, levels, or observations.

For example, to specify a four-factor completely nested design in the GLM procedure, you use the following form:

```
class a b c d;
model y=a b(a) c(a b) d(a b c);
```

However, to specify the same design in PROC NESTED, you simply use the following form:

```
class a b c d;
var y;
```

In addition, other procedures require TEST statements to perform appropriate tests, whereas the NESTED procedure produces the appropriate tests automatically. However, PROC NESTED makes one assumption about the input data that the other procedures do not: **PROC NESTED assumes that the input data set is sorted by the classification (CLASS) variables defining the effects.** If you use PROC NESTED on data that are not sorted by the CLASS variables, then the results might not be valid.

Getting Started: NESTED Procedure

Reliability of Automobile Models

A study is performed to compare the reliability of several models of automobiles. Three different automobile models (Model) from each of four U.S. automobile manufacturers (Make) are tested. Three different cars of each make and model are subjected to a reliability test and given a score between 1 and 100 (Score), where higher scores indicate greater reliability.

The following statements create the SAS data set auto.

```
title1 'Reliability of Automobile Models';
data auto;
   input Make $ Model Score @@;
   datalines;
a 1 62   a 2 77   a 3 59
a 1 67   a 2 73   a 3 64
a 1 60   a 2 79   a 3 60
b 1 72   b 2 58   b 3 80
b 1 75   b 2 63   b 3 84
b 1 69   b 2 57   b 3 89
c 1 94   c 2 76   c 3 81
c 1 90   c 2 75   c 3 85
c 1 88   c 2 78   c 3 85
d 1 69   d 2 73   d 3 90
d 1 72   d 2 88   d 3 87
d 1 76   d 2 87   d 3 92
;
```

The Make variable contains the make of the automobile, represented here by 'a', 'b', 'c', or 'd', while the Model variable represents the automobile model with a '1', '2', or '3'. The Score variable contains the reliability scores given to the three sampled cars from each Make-Model group. Since the automobile models are nested within their makes, the NESTED procedure is used to analyze these data. The NESTED procedure requires the data to be sorted by Make and, within Make, by Model, so the following statements execute a PROC SORT before completing the analysis.

```
proc sort data=auto;
   by Make Model;
run;

title1 'Reliability of Automobile Models';
proc nested data=auto;
   classes Make Model;
   var Score;
run;
```

The Model variable appears after the Make variable in the CLASS statement because it is nested within Make. The VAR statement specifies the response variable. The output is displayed in Figure 59.1.

Figure 59.1 Output from PROC NESTED

```
                         Reliability of Automobile Models

                              The NESTED Procedure

                          Coefficients of Expected Mean Squares

                        Source       Make       Model       Error

                        Make          9           3           1
                        Model         0           3           1
                        Error         0           0           1


                 Nested Random Effects Analysis of Variance for Variable Score
```

Variance Source	DF	Sum of Squares	F Value	Pr > F	Error Term	Mean Square	Variance Component	Percent of Total
Total	35	4177.888889				119.368254	131.876543	100.0000
Make	3	1709.000000	2.15	0.1719	Model	569.666667	33.867284	25.6811
Model	8	2118.888889	18.16	<.0001	Error	264.861111	83.425926	63.2606
Error	24	350.000000				14.583333	14.583333	11.0583

```
                            Score Mean                      75.94444444
                            Standard Error of Score Mean     3.97794848
```

Figure 59.1 first displays the coefficients of the variance components that make up each of the expected mean squares, and then it displays the ANOVA table. The results do not indicate significant variation between the different automobile makes ($F = 2.15$, $p = 0.1719$). However, they do suggest that there is significant variation between the different models within the makes ($F = 18.16$, $p < 0.0001$). This is evident in the fact that the make of car accounts for only 25.7% of the total variation in the data, while the car model accounts for 63.3% (as shown in the Percent of Total column). The estimated variance components are shown in the Variance Component column.

Syntax: NESTED Procedure

The following statements are available in PROC NESTED:

PROC NESTED < *options* > ;
    CLASS *variables* < / *option* > ;
    VAR *variables* ;
    BY *variables* ;

The PROC NESTED and CLASS statements are required. The BY, CLASS, and VAR statements are described after the PROC NESTED statement.

PROC NESTED Statement

> **PROC NESTED** < *options* > **;**

The PROC NESTED statement has the following options:

AOV

> displays only the analysis of variance statistics when there is more than one dependent variable. The "analysis of covariation" statistics are suppressed (see the section "Analysis of Covariation" on page 4884).

DATA=_SAS-data-set_

> names the SAS data set to be used by PROC NESTED. By default, the procedure uses the most recently created SAS data set.

BY Statement

> **BY** *variables* **;**

You can specify a BY statement with PROC NESTED to obtain separate analyses on observations in groups that are 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. If you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.

- Specify the NOTSORTED or DESCENDING option in the BY statement for the NESTED 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

> **CLASS** *variables* < / *option* > ;

You must include a CLASS statement with PROC NESTED specifying the classification variables for the analysis.

Values of a variable in the CLASS statement denote the levels of an effect. The name of that variable is also the name of the corresponding effect. The second effect is assumed to be nested within the first effect, the third effect is assumed to be nested within the second effect, and so on.

By default, class levels are determined from the entire formatted values of the CLASS variables. Note that this represents a slight change from previous releases in the way in which class levels are determined. Prior to SAS 9, class levels were determined using no more than the first eight characters of the formatted values, except for numeric variables with no explicit format, for which class levels were determined from the raw numeric values. If you want to revert to this previous behavior, you can use the TRUNCATE option in the CLASS statement. In any case, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *Base SAS Procedures Guide*, and the discussions for the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

NOTE: The data set must be sorted by the classification variables in the order in which they are given in the CLASS statement. Use PROC SORT to sort the data if they are not already sorted.

You can specify the following option in the CLASS statement after a slash (/):

TRUNCATE

> specifies that class levels should be determined by using only up to the first 16 characters of the formatted values of CLASS variables. When formatted values are longer than 16 characters, you can use this option to revert to the levels as determined in releases prior to SAS 9.

VAR Statement

> **VAR** *variables* ;

The VAR statement lists the dependent variables for the analysis. The dependent variables must be numeric variables. If you do not specify a VAR statement, PROC NESTED performs an analysis of variance for all numeric variables in the data set, except those already specified in the CLASS statement.

Details: NESTED Procedure

Missing Values

An observation with missing values for any of the variables used by PROC NESTED is omitted from the analysis. Blank values of CLASS character variables are treated as missing values.

Unbalanced Data

A completely nested design is defined to be unbalanced if the groups corresponding to the levels of some classification variable are not all of the same size. The NESTED procedure can compute unbiased estimates for the variance components in an unbalanced design, but because the sums of squares on which these estimates are based no longer have χ^2 distributions under a Gaussian model for the data, F tests for the significance of the variance components cannot be computed. PROC NESTED checks to see that the design is balanced. If it is not, a warning to that effect appears in the log, and the columns corresponding to the F tests in the analysis of variance are left blank.

General Random-Effects Model

A random-effects model for data from a completely nested design with n factors has the general form

$$y_{i_1 i_2 \cdots i_n r} = \mu + \alpha_{i_1} + \beta_{i_1 i_2} + \cdots + \epsilon_{i_1 i_2 \cdots i_n r}$$

where

$y_{i_1 i_2 \cdots i_n r}$	is the value of the dependent variable observed at the rth replication with factor j at level i_j, for $j = 1, \ldots, n$.
μ	is the overall (fixed) mean of the sampled population.
$\alpha_{i_1}, \beta_{i_1 i_2}, \ldots, \epsilon_{i_1 i_2 \cdots i_n r}$	are mutually uncorrelated random effects with zero means and respective variances $\sigma_\alpha^2, \sigma_\beta^2, \ldots, \sigma_\epsilon^2$.

Analysis of Covariation

When you specify more than one dependent variable, the NESTED procedure produces a descriptive analysis of the covariance between each pair of dependent variables in addition to a separate analysis of variance for each variable. The analysis of covariation is computed under the basic random-effects model for each pair of dependent variables:

$$
\begin{aligned}
y_{i_1 i_2 \cdots i_n r} &= \mu + \alpha_{i_1} + \beta_{i_1 i_2} + \cdots + \epsilon_{i_1 i_2 \cdots i_n r} \\
y'_{i_1 i_2 \cdots i_n r} &= \mu' + \alpha'_{i_1} + \beta'_{i_1 i_2} + \cdots + \epsilon'_{i_1 i_2 \cdots i_n r}
\end{aligned}
$$

where the notation is the same as that used in the preceding general random-effects model.

There is an additional assumption that all the random effects in the two models are mutually uncorrelated except for corresponding effects, for which

$$
\begin{aligned}
\mathrm{Corr}(\alpha_{i_1}, \alpha'_{i_1}) &= \rho_\alpha \\
\mathrm{Corr}(\beta_{i_1 i_2}, \beta'_{i_1 i_2}) &= \rho_\beta \\
&\vdots \\
\mathrm{Corr}(\epsilon_{i_1 i_2 \cdots i_n r}, \epsilon'_{i_1 i_2 \cdots i_n r}) &= \rho_\epsilon
\end{aligned}
$$

Error Terms in *F* Tests

Random-effects ANOVAs are distinguished from fixed-effects ANOVAs by which error mean squares are used as the denominator for F tests. Under a fixed-effects model, there is only one true error term in the model, and the corresponding mean square is used as the denominator for all tests. This is how the usual analysis is computed in PROC ANOVA, for example. However, in a random-effects model for a nested experiment, mean squares are compared sequentially. The correct denominator in the test for the first factor is the mean square due to the second factor; the correct denominator in the test for the second factor is the mean square due to the third factor; and so on. Only the mean square due to the last factor, the one at the bottom of the nesting order, should be compared to the error mean square.

Computational Method

The building blocks of the analysis are the sums of squares for the dependent variables for each classification variable within the factors that precede it in the model, corrected for the factors that follow it. For example, for a two-factor nested design, PROC NESTED computes the following sums of squares:

$$\text{Total SS} \qquad \sum_{ijr} (y_{ijr} - y_{...})^2$$

$$\text{SS for Factor 1} \qquad \sum_{i} n_{i \cdot} \left(\frac{y_{i \cdot \cdot}}{n_{i \cdot}} - \frac{y_{...}}{n_{..}} \right)^2$$

$$\text{SS for Factor 2 within Factor 1} \qquad \sum_{ij} n_{ij} \left(\frac{y_{ij \cdot}}{n_{ij}} - \frac{y_{i \cdot \cdot}}{n_{i \cdot}} \right)^2$$

$$\text{Error SS} \qquad \sum_{ijr} \left(y_{ijr} - \frac{y_{ij \cdot}}{n_{ij}} \right)^2$$

where y_{ijr} is the rth replication, n_{ij} is the number of replications at level i of the first factor and level j of the second, and a dot as a subscript indicates summation over the corresponding index. If there is more than one dependent variable, PROC NESTED also computes the corresponding sums of crossproducts for each pair. The expected value of the sum of squares for a given classification factor is a linear combination of the variance components corresponding to this factor and to the factors that are nested within it. For each factor, the coefficients of this linear combination are computed. (The efficiency of PROC NESTED is partly due to the fact that these various sums can be accumulated with just one pass through the data, assuming that the data have been sorted by the classification variables.) Finally, estimates of the variance components are derived as the solution to the set of linear equations that arise from equating the mean squares to their expected values.

Displayed Output

PROC NESTED displays the following items for each dependent variable:

- Coefficients of Expected Mean Squares, which are the coefficients of the $n + 1$ variance components making up the expected mean square. Denoting the element in the ith row and jth column of this matrix by C_{ij}, the expected value of the mean square due to the ith classification factor is

$$C_{i1}\sigma_1^2 + \cdots + C_{in}\sigma_n^2 + C_{i,n+1}\sigma_\epsilon^2$$

C_{ij} is always zero for $i > j$, and if the design is balanced, C_{ij} is equal to the common size of all classification groups of the jth factor for $i \leq j$. Finally, the mean square for error is always an unbiased estimate of σ_ϵ^2. In other words, $C_{n+1,n+1} = 1$.

For every dependent variable, PROC NESTED displays an analysis of variance table. Each table contains the following:

- each Variance Source in the model (the different components of variance) and the total variance

- degrees of freedom (DF) for the corresponding sum of squares

- Sum of Squares for each classification factor. The sum of squares for a given classification factor is the sum of squares in the dependent variable within the factors that precede it in the model, corrected for the factors that follow it. (See the section "Computational Method" on page 4885.)

- *F* Value for a factor, which is the ratio of its mean square to the appropriate error mean square. The next column, labeled PR > F, gives the significance levels that result from testing the hypothesis that each variance component equals zero.

- the appropriate Error Term for an *F* test, which is the mean square due to the next classification factor in the nesting order. (See the section "Error Terms in *F* Tests" on page 4884.)

- Mean Square due to a factor, which is the corresponding sum of squares divided by the degrees of freedom

- estimates of the Variance Components. These are computed by equating the mean squares to their expected values and solving for the variance terms. (See the section "Computational Method" on page 4885.)

- Percent of Total, the proportion of variance due to each source. For the ith factor, the value is

$$100 \times \frac{\text{source variance component}}{\text{total variance component}}$$

- Mean, the overall average of the dependent variable. This gives an unbiased estimate of the mean of the population. Its variance is estimated by a certain linear combination of the estimated variance components, which is identical to the mean square due to the first factor in the model divided by the total number of observations when the design is balanced.

If there is more than one dependent variable, then the NESTED procedure displays an "analysis of covariation" table for each pair of dependent variables (unless the AOV option is specified in the PROC NESTED statement). See the section "Analysis of Covariation" on page 4884 for details. For each source of variation, this table includes the following:

- Degrees of Freedom

- Sum of Products

- Mean Products

- Covariance Component, the estimate of the covariance component

Items in the analysis of covariation table are computed analogously to their counterparts in the analysis of variance table. The analysis of covariation table also includes the following:

- Variance Component Correlation for a given factor. This is an estimate of the correlation between corresponding effects due to this factor. This correlation is the ratio of the covariance component for this factor to the square root of the product of the variance components for the factor for the two different dependent variables. (See the section "Analysis of Covariation" on page 4884.)

- Mean Square Correlation for a given classification factor. This is the ratio of the Mean Products for this factor to the square root of the product of the Mean Squares for the factor for the two different dependent variables.

ODS Table Names

PROC NESTED 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 about ODS, see Chapter 20, "Using the Output Delivery System."

Table 59.1   ODS Tables Produced by PROC NESTED

ODS Table Name	Description	Statement
ANCOVA	Analysis of covariance	default with more than one dependent variable
ANOVA	Analysis of variance	default
EMSCoef	Coefficients of expected mean squares	default
Statistics	Overall statistics for fit	default

Example: NESTED Procedure

Example 59.1: Variability of Calcium Concentration in Turnip Greens

In the following example from Snedecor and Cochran (1976), an experiment is conducted to study the variability of calcium concentration in turnip greens. Four plants are selected at random; then three leaves are randomly selected from each plant. Two 100-mg samples are taken from each leaf. The amount of calcium is determined by microchemical methods.

Because the data are read in sorted order, it is not necessary to use PROC SORT on the CLASS variables. Leaf is nested in Plant; Sample is nested in Leaf and is left for the residual term. All the effects are random effects. The following statements read the data and invoke PROC NESTED. These statements produce Output 59.1.1.

```
title 'Calcium Concentration in Turnip Leaves--Nested Random Model';
title2 'Snedecor and Cochran, ''Statistical Methods'', 1976, p. 286';
data Turnip;
   do Plant=1 to 4;
      do Leaf=1 to 3;
         do Sample=1 to 2;
            input Calcium @@;
            output;
         end;
      end;
   end;
   datalines;
3.28 3.09 3.52 3.48 2.88 2.80 2.46 2.44
1.87 1.92 2.19 2.19 2.77 2.66 3.74 3.44
2.55 2.55 3.78 3.87 4.07 4.12 3.31 3.31
;

proc nested data=Turnip;
   classes plant leaf;
   var calcium;
run;
```

Output 59.1.1 Analysis of Calcium Concentration in Turnip Greens Using PROC NESTED

```
               Calcium Concentration in Turnip Leaves--Nested Random Model
                Snedecor and Cochran, 'Statistical Methods', 1976, p. 286

                               The NESTED Procedure

                          Coefficients of Expected Mean Squares

                        Source        Plant        Leaf        Error

                        Plant           6            2            1
                        Leaf            0            2            1
                        Error           0            0            1


                  Nested Random Effects Analysis of Variance for Variable Calcium

  Variance              Sum of                        Error                    Variance    Percent
  Source        DF      Squares   F Value  Pr > F     Term     Mean Square    Component   of Total

  Total         23     10.270396                                 0.446539      0.532938   100.0000
  Plant          3      7.560346    7.67   0.0097     Leaf       2.520115      0.365223    68.5302
  Leaf           8      2.630200   49.41   <.0001     Error      0.328775      0.161060    30.2212
  Error         12      0.079850                                 0.006654      0.006654     1.2486


                        Calcium Mean                              3.01208333
                        Standard Error of Calcium Mean            0.32404445
```

The results indicate that there is significant (nonzero) variation from plant to plant (Pr > F is 0.0097) and from leaf to leaf within a plant (Pr > F is less than 0.0001). Notice that the variance component for Plant uses the Leaf mean square as an error term in the model rather than the error mean square.

References

Snedecor, G. W. and Cochran, W. G. (1976), *Statistical Methods*, Sixth Edition, Ames: Iowa State University Press.

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

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

§sas. | THE POWER TO KNOW®