PROC NESTED: Introduction

The NESTED procedure performs random effects analysis of variance for data from an experiment with a nested (hierarchical) structure. 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.

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. 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 is not sorted by the CLASS variables, then the results may not be valid.

PROC NESTED: Syntax

The following statements are available in PROC NESTED.

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

The PROC NESTED and CLASS statements are required.

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

- DATA=*SAS-data-set*

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

CLASS *variables* ***;***

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.

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

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.

BY *variables* ***;***

You can specify a BY statement with PROC NESTED 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 the order of the BY variables.

Note: When you use the NESTED procedure, your data must be sorted first by the BY variables and, within the BY variables, by the CLASS variables.

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 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 using the DATASETS procedure (in base SAS software).

PROC NESTED: 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.

PROC NESTED: 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 chi-square 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 is placed on the log, and the columns corresponding to the *F* tests in the analysis of variance are left blank.

OUTPUT from PROC NESTED

Nested Analysis of Variance 11:59 Tuesday, April 1, 2003

sex=F

The NESTED Procedure

Coefficients of Expected Mean Squares

Source	bndyr	cws	Error
bndyr	66.1408451	2.0000000	1.0000000
cws	0.0000000	2.0000000	1.0000000
Error	0.0000000	0.0000000	1.0000000

Nested Random Effects Analysis of Variance for Variable tarsus_T

Variance Source	DF	Sum of Squares	F Value	Pr > F	Error Term
Total	425	442283			
bndyr	5	29190			
cws	207	378502			
Error	213	34591			

Nested Random Effects Analysis of Variance for Variable tarsus_T

Variance Source	Mean Square	Variance Component	Percent of Total
Total	1040.666009	1056.076396	100.0000
bndyr	**5838.073650**	60.621552	5.7403
cws	**1828.512975**	833.058130	78.8824
Error	**162.396714**	162.396714	15.3774

tarsus_T Mean	933.09624413
Standard Error of tarsus_T Mean	4.22531237

F tests for the significance of the variance components were not computed since the nested design is not balanced.

F_{cws} = MS(cws)/MS(error)
F_{bndyr} = MS(bndyr)/MS(cws) → this has to be adjusted due to different sample sizes.

Check Box 10.6 - Two-level Nested ANOVA with unequal sample sizes.
in Sokal, R. R., and F. J. Rohlf. 1995. Biometry. The principles and practice of statistics in biological research, 3rd edition. W. H. Freeman and Company, New York. Pp: 294-299.

Chapter 3 Analyzing Data with Random Effects

.1 Introduction

Many studies incorporate blocking factors to provide replication over a selection of different conditions. Investigators are not specifically interested in individual blocks, but rather what the average across blocks reveals. For example, you might want to test chemical compounds at several laboratories to compare the compounds averaged across all the laboratories. Laboratories are selected to represent some broader possible set of laboratories. You might not be too interested in what happens at specific laboratories. Other studies employ experimental factors in which the levels of the factors are a sample of a much larger collection of possible levels. Interest would typically not be in the specific levels of the factors employed but instead on results averaged across the levels and on the degree of variability among the levels. If you work in industry, you probably have seen experiments that utilize a selection of batches of raw material, or a sample of the workers on an assembly line, or a subset of machines out of a much larger set of machines that are used in a production process. In these examples, interest is in what happens across the broader collection of laboratories

or batches or workers or machines rather than in what happens with a particular laboratory or batch or worker or machine that was actually employed in the experiment. The factor (laboratories, batches, workers, machines, or whatever) is called a *random* effect. Theoretically, the levels of the factor that are in the experiment are considered to be a random sample from a broader population of possible levels of the factor.

With balanced data, the presence of random factors does not present a major issue for the estimation of treatment means or differences between treatment means. You simply compute means or differences between means, averaged across the levels of the random factors in the experiment. However, the presence of random effects has a major impact on the use of appropriate statistical techniques for testing hypotheses and constructing standard errors of estimates. It is safe to say that improper attention to the presence of random effects is one of the most common and serious mistakes in statistical analysis of data. Random effects probably occur in one form or another in the majority of statistical studies. The RANDOM statement in the GLM procedure can help you determine correct methods for a large variety of applications.

3.2 Nested Classifications

Nested classifications of data have sampling units which are classified in a hierarchical or *nested* manner. Typically, these samples are taken in several stages:

1. selection of main units
2. selection of subunits from each main unit
3. selection of sub-subunits from the subunits, and so on.

Normally, the classification factors at each stage are considered random effects, but in some cases a classification factor may be considered fixed, especially one corresponding to the first stage of sampling.

Here is an example of a nested classification. Microbial counts are made on samples of ground beef in a study whose objective is to assess sources of variation in numbers of microbes. Twenty packages of ground beef (PACKAGE) are purchased and taken to a laboratory. Three samples (SAMPLE) are drawn from each package, and two replicate counts are made on each sample. Output 3.1 shows the raw data.

Output 3.1
Microbial Counts in Ground Beef

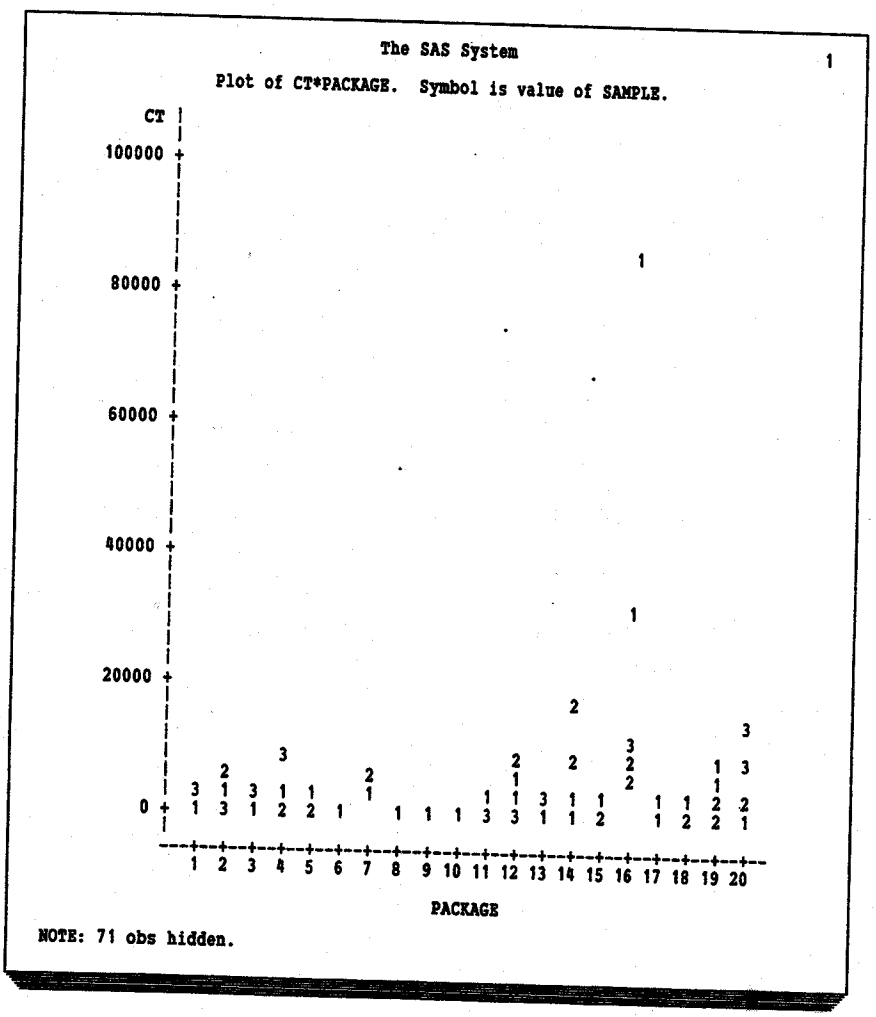
The SAS System 1

OBS	PACKAGE	CT11	CT12	CT21	CT22	CT31	CT32
1	1	527	821	107	299	1382	3524
2	2	2813	2322	3901	4422	383	479
3	3	703	652	745	995	2202	1298
4	4	1617	2629	103	96	2103	8814
5	5	4169	2907	4018	882	768	271
6	6	67	28	68	111	277	199
7	7	1612	1680	6619	4028	5625	6507
8	8	195	127	591	399	275	152
9	9	619	520	813	956	1219	923
10	10	436	555	58	58	236	188
11	11	1682	3235	2963	2249	457	2950
12	12	6050	3956	2782	7501	1952	1299

13	13	1330	758	132	93	1116	3186
14	14	1834	1200	18248	9496	252	433
15	15	2339	4057	106	146	430	442
16	16	31229	84451	6806	9156	12715	12011
17	17	1147	3437	132	175	719	1243
18	18	3440	3185	712	467	680	205
19	19	8196	4565	1459	1292	9707	8138
20	20	1090	1037	4188	1859	8464	14073

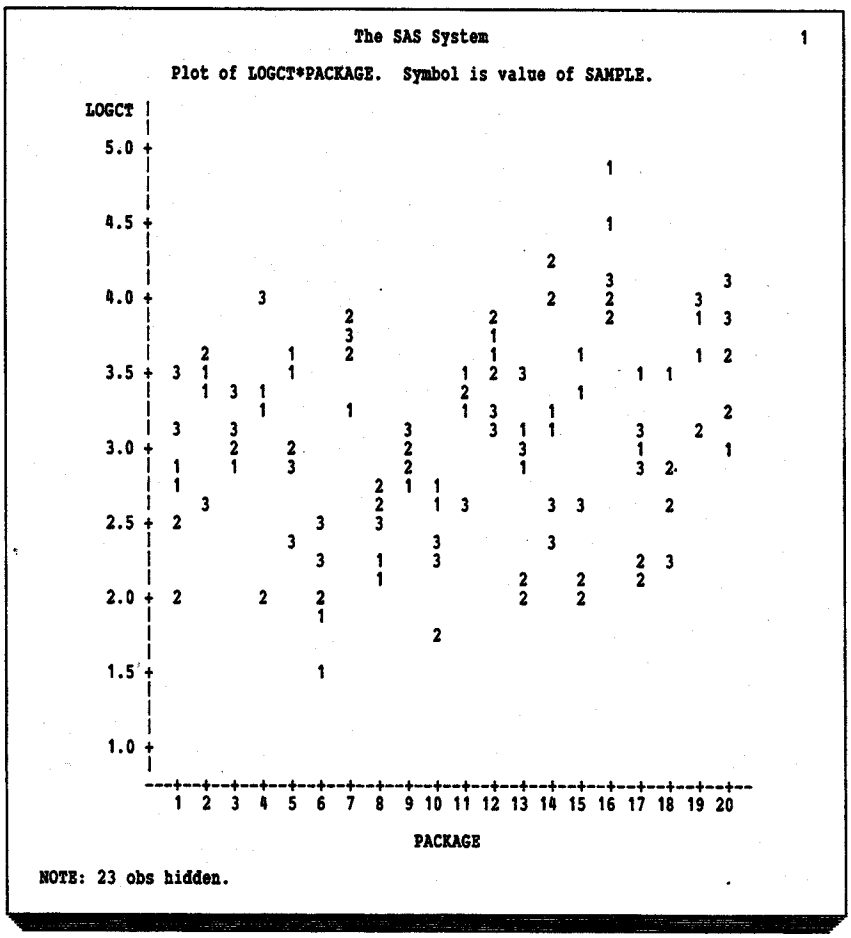
The data are plotted in Output 3.2, with points identified according to their SAMPLE number.

Output 3.2 Plot of Count versus Package Number



You can see the larger variation among larger counts. In order to stabilize the variance, the logarithm (base 10) of the counts (LOGCT) was computed and serves as the response variable to be analyzed. The plot of LOGCT, which appears in Output 3.3, indicates the transformation was successful in stabilizing the variance.

Output 3.3
Plot of Log Count versus Package Number



Logarithms are commonly computed for microbial data for the additional reason that interest is in differences in the order of magnitude rather than interval differences.

A model for the data is

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}$$

where

y_{ijk} is the $\log_{10}$ count for the kth replicate of the jth sample from the ith package.

μ is the overall mean of the sampled population.

α_i are random variables representing differences between packages, with variance σ_P^2, $i=1, \ldots, 20$.

β_{ij} are random variables representing differences between samples in the same package, with variance σ_S^2, $i=1, \ldots, 20$, $j=1,2,3$.

ε_{ijk} are random variables representing differences between replicate counts in the same sample, with variance σ^2, $i=1, \ldots, 20$, $j=1,2,3$, and $k=1,2$.

The random variables α_i, β_{ij}, and ε_{ijk} are assumed independent with means equal to 0. Then the variance (V) of the log counts can be expressed as

$$V(y_{ijk}) = \sigma_y^2$$
$$= \sigma_P^2 + \sigma_S^2 + \sigma^2 \quad .$$

Expressing the equation with words, the variance of the logarithms of microbial count is equal to the sum of the variances due to differences between packages, between samples in the same package, and between replicates in the same sample. These individual variances are therefore called *components of variance*. The first objective is to estimate the variance components, and there are several statistical techniques for doing so, including analysis of variance and maximum likelihood. In this chapter, analysis-of-variance methods are used.

3.2.1 Analysis of Variance for Nested Classification: Using PROC NESTED to Estimate Variance Components

An analysis-of-variance table for the ground beef microbial counts has the following form:

Source of Variation	DF
packages	19
samples in packages	40
replicates in samples	60

You can produce this table using the ANOVA or GLM procedures (see Chapter 2, "Analysis of Variance for Balanced Data"). The NESTED and VARCOMP procedures also produce this table. Which procedure is best to use depends on the objectives of the investigation.

PROC ANOVA and PROC GLM are general purpose procedures that can be used for a broad range of data classifications. In contrast, PROC NESTED is a specialized procedure that is useful only for nested classifications. It provides estimates of the components of variance using the analysis-of-variance method of estimation, which is the purpose here. Later sections discuss some of the features of PROC ANOVA and PROC GLM for nested classifications.

Because PROC NESTED is so specialized, it is very easy to use. You simply list the sources of variation in the proper order in a CLASS statement. This means the CLASS statement in PROC NESTED has a broader purpose than it does in PROC ANOVA and PROC GLM; it encompasses the purpose of the MODEL statement as well. But you must also have the data sorted appropriately, following the same order as the classification scheme. Here are the proper SAS statements:

```
proc sort; by package sample;
proc nested; class package sample;
var logct;
```

Results appear in Output 3.4.

Output 3.4
Nested Analysis of Variance of Log Count

```
                                The SAS System                              1
                    Coefficients of Expected Mean Squares
                Source       PACKAGE       SAMPLE        ERROR

                PACKAGE           6             2            1
                SAMPLE            0             2            1
                ERROR             0             0            1
```

```
                                                                            2
          Nested Random Effects Analysis of Variance for Variable LOGCT

                  Degrees
   Variance          of            Sum of                                 Error
   Source         Freedom         Squares    F Value       Pr > F          Term

   TOTAL              119       52.772098
   PACKAGE             19       30.529155      3.224       0.0009        SAMPLE
   SAMPLE              40       19.934312     12.952       0.0000        ERROR
   ERROR               60        2.308631

   Variance                         Variance                Percent
   Source          Mean Square     Component               of Total

   TOTAL              0.443463      0.453157               100.0000
   PACKAGE            1.606798      0.184740                40.7673
   SAMPLE             0.498358      0.229940                50.7418
   ERROR              0.038477      0.038477                 8.4909

                   Mean                              3.04945863
                   Standard error of mean            0.11571508
```

First look at the portion of output labeled Coefficients of Expected Mean Squares. This part of the output tells you the expressions for the expected values of the mean squares, that is, what is being estimated by the individual mean squares. Table 3.1 shows you how to interpret the coefficients of expected mean squares.

Table 3.1 *Coefficients of Expected Mean Squares*

Variance Source	Source of Variation	DF	Expected Mean Squares	This tells you:
PACKAGE	packages	19	$\sigma^2 + 2\sigma_S^2 + 6\sigma_P^2$	MS(PACKAGE) estimat[es] $\sigma^2 + 2\sigma_S^2 + 6\sigma_P^2$
SAMPLE	samples in packages	40	$\sigma^2 + 2\sigma_S^2$	MS(SAMPLE) estimates $\sigma^2 + 2\ \sigma_S^2$
ERROR	replicates in samples	60	σ^2	MS(ERROR) estimates σ^2

From the table of expected mean squares you get the estimates of variance components. These estimates are printed under the heading Variance Component.

- □ $\hat{\sigma}^2 = 0.0385 = \text{MS(ERROR)}$
- □ $\hat{\sigma}_S^2 = 0.2299 = [\text{MS(SAMPLE)} - \text{MS(ERROR)}]/2$
- □ $\hat{\sigma}_P^2 = 0.1847 = [\text{MS(PACKAGE)} - \text{MS(SAMPLE)}]/6$

The variance of a single microbial count is

$$\begin{aligned}\hat{\sigma}_y^2 &= \text{TOTAL Variance Estimate} \\ &= \hat{\sigma}^2 + \hat{\sigma}_S^2 + \hat{\sigma}_P^2 \\ &= 0.0385 + 0.2299 + 0.1847 \\ &= 0.4532 \quad .\end{aligned}$$

Note: The expression TOTAL Variance Estimate does not refer to MS(TOTAL)=0.4435, although the values are similar.

Under the heading Percent of Total you see that

- □ 8.49% of TOTAL variance is attributable to ERROR variance
- □ 50.74% of TOTAL variance is attributable to SAMPLE variance
- □ 40.77% of TOTAL variance is attributable to PACKAGE variance.