BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 40 DATE: August 6, 1992

SUBJECT: Finding the Expected Mean Squares and the Proper Error Terms with SAS

When you analyze data using ANOVA, one challenging task is to determine the expected mean square of a source and its error term. This task becomes more difficult for mixed effect models or unbalanced designs where simple F-tests may not be valid. This pamphlet discusses the use of the RANDOM statement in PROC GLM of SAS to help solve this problem, and the controversy over expected mean squares for mixed effect models.

A mixed effect model is one that involves both random and fixed factors. We will use the data taken from Milliken and Johnson (1984, p. 285) as our example.

A company wanted to evaluate the productivity of three machines when operated by the company's own personnel. Six employees were randomly selected to operate each machine at three different trials, each trial is assumed to be independent of the others. A score was assigned to reflect the quality of production. The data are reproduced in Table 1 of appendix 1.

This is a two-way mixed factor experiment where MACHINE is a fixed factor and PERSON is a random factor. The data can be analyzed with the following PROC GLM step:

```
PROC GLM    DATA = EXAMPLE;
    CLASS   MACHINE  PERSON;
    MODEL   SCORE = MACHINE | PERSON / SS3;
    RANDOM  PERSON MACHINE*PERSON / TEST;
RUN;
```

EXAMPLE is a SAS data set created previously in the program. The CLASS statement specifies the classification variables; the MODEL statement states the model to be fitted; and the RANDOM statement lists the random sources.

An interaction effect is random if at least one of the factors involved is random (Milliken and Johnson 1984, p. 275). Therefore, PERSON and MACHINE*PERSON are random and are listed in the RANDOM statement. The RANDOM statement requests that the expected mean square, E(MS), for all the effects listed in the MODEL statement be output. It has an option (new in version 6.03):

TEST to test the effects in the model with the proper error term.

The SAS output on the score data using the above PROC GLM step is shown in appendix 2. The SAS output pages are concatenated with the page number shown as a single digit on the upper right hand corner of the output.

Page 1 of the SAS output gives a summary of the design structure. Page 2 gives the ANOVA table with the TYPE III SS and mean squares. The F-value for the sources are all computed with the default error term as the denominator, which is correct only for the MACHINE*PERSON effect.

Page 3 is generated by the RANDOM statement. It gives the E(MS)'s of all the effects in the model. By comparing these equations, we can find the proper error term for each effect. The rule for determining the error term of an effect is:

> The E(MS) of the error term must be identical to the E(MS) of the effect of interest, except for the variance component due to that effect.

Applying this rule, we can see that the E(MS) of MACHINE contains the same terms as E(MS) of MACHINE* PERSON and an extra term Q(MACHINE), the variance component due to MACHINE. Therefore MACHINE*PERSON is the error term for the effect MACHINE. Similarly MACHINE*PERSON is the error term for PERSON; the default error is the error term for MACHINE*PERSON. Note that in SAS, the variance component of a fixed effect is denoted by Q, and the variance component of a random effect is denoted by var. The F-test results for all the effects using the proper error term are given in page 4 of the SAS output, and is summarized in the following ANOVA table:

Source	df	SS	MS	Error Term	F	p-value
MACHINE, M	2	1755.26	877.63	M x P	20.576	0.0003
PERSON, P	5	1241.90	248.38	M x P	5.823	0.0089
M x P	10	426.53	42.65	Error	46.130	0.0001
Error	36	33.29	0.92	--	--	--
Total	53	3456.98				

In this simple design, the proper error term can be found easily by comparing the E(MS)'s of the various effects. When the design becomes complicated or is unbalanced, finding the right error term would no longer be an easy task. The RANDOM statement, however, can still be used to find the error terms in such cases.

Table 2 in appendix 1 shows an incomplete data set obtained by randomly deleting several data points from the full data set in Table 1. We can analyze the unbalanced data with the same PROC GLM step as before. The SAS output is given in appendix 3. The test results are summarized in the ANOVA tables shown on the next page. Notice that the E(MS)'s are different for the balanced and unbalanced cases. In the unbalanced case, none of the sources can be used directly to test the effects M and P. Hence pseudo F-tests must be used (Bergerud 1989). The proper denominators for the pseudo F-tests can be found using the TEST option.

Source	df	E(MS)	Error Term (from SAS)
MACHINE, M	2	$\sigma_E^2 + 2.137\sigma_{MP}^2 + \phi_M$	0.9226 MS(MxP) + 0.0774 MS(E)
PERSON, P	5	$\sigma_E^2 + 2.241\sigma_{MP}^2 + 6.7224\sigma_P^2$	0.9674 MS(MxP) + 0.0326 MS(E)
M x P	10	$\sigma_E^2 + 2.316\ \sigma_{MP}^2$	MS(E)
Error	36	σ_E^2	--

Source	df	SS	MS	denominator df	F	p-value
MACHINE, M	2	1238.20	619.10	10.04	16.57	0.0007
PERSON, P	5	1011.05	202.21	10.01	5.17	0.0133
M x P	10	404.32	40.43	26	46.34	0.0001
Error	36	22.69	0.87	--	--	--
Total	53	3084.43				

We can check the appropriateness of an error term by writing in full its E(MS) equation. For example, the error term for MACHINE is:

0.9226 MS(MxP) + 0.0774 MS(E) which has expected value:

$$0.9226[\sigma_E^2 + 2.316\ \sigma_{MP}^2] + 0.0774\ \sigma_E^2$$
$$= \sigma_E^2 + 2.137\sigma_{MP}^2$$

which is identical to the E(MS) of MACHINE except for the term ϕ_M. Since the error term is a linear combination of two mean squares, the denominator df must be adjusted accordingly, as described in Bergerud (1989).

You may notice that in the SAS output for the balanced data, the E(MS) for the random effect PERSON includes the variance component due to the interaction effect MACHINE*PERSON. However, if you follow the E(MS) rules suggested by Kirk (1982), Scheffé (1959), or Schultz (1955), you will obtain an E(MS) without the interaction component. This discrepancy is evident in many places. For example, Hartley and Searle (1969), Milliken and Johnson(1984), and Searle (1971) include the interaction component; Graybill (1961), Wilk and Kempthorne (1955), and Snedecor and Cochran (1967) do not include it; Mood and Graybill (1963) do not discuss the topic.

Cornfield and Tukey (1956) pointed out that the two approaches have the same assumptions that "observed values are linear combinations of certain fixed and random variables, but differ in the nature of the restrictions that are imposed upon these variables". Some argue that if an interaction effect contains a random factor, then the interaction effect should be treated as a random variable with no constraints imposed on them. In such a case, the variance component due to the interaction should be included in the main effect mean squares (result 1). On the other hand, if we impose the restriction that interaction effects sum to zero over the levels of the fixed factor, then the interaction component of variance will drop out of the random effect mean squares (result 2). For unbalanced data, however, both approaches would include the interaction component in the E(MS), as shown in the example.

Result 2 includes the interaction component for unbalanced data but excludes it for balanced data. Hartley and Searle (1969) refer this as a discontinuity between the analysis of balanced and unbalanced data. The Biometrics section has been using result 2 in the past. I prefer result 1 and include the interaction component because this approach is consistent in both balanced and unbalanced cases.

In any case, if you keep this E(MS) controversy in mind and proceed with care, the `RANDOM` statement in SAS is a useful tool for determining the expected mean squares and error terms.

References

Bergerud, W. (1989). ANOVA: Approximate or Pseudo f-tests. *Biom. Infor. Pamp., 19.*

Cornfield, J. and Tukey, J. W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist., 27*, 907-949.

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

Hartley, H. O. and S. R. Searle (1969). A discontinuity in mixed model analyses, *Biometrics, 25*, 573-576.

Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Science*. Brooks/Cole, Belmont, California.

Milliken, G. and D. Johnson (1984). *Analysis of Messy Data. Vol. 1*. Wadsworth Inc., Belmont, Cal.

Mood, A. M. and F. A, Graybill (1963). *Introduction to the Theory of Statistics*. 2nd Ed. McGraw-Hill, New York.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Schultz, E. F., Jr. (1955). Rules of thumb for determining expectations of mean squares in analysis of variance. *Biometrics, 11*, 123-135.

Searle, S. R. (1971). *Linear Models*. Wiley, New York.

Snedecor, G. W. and W. G. Cochran. (1967). *Statistical Methods*. 6th Ed. Iowa State University Press, Ames, Iowa.

Wilk, M. B. and O. Kempthorne. (1955). Fixed, mixed and random models. *J. Am. Stat. Assoc., 50*, 1144-1167.

CONTACT: Vera Sit
356-0435

APPENDIX 1: Productivity Scores for Machine-Person Example Taken from Milliken and Johnson (1984, p. 285)

Table 1: Balanced Case

Machine	Person	Trial 1	Trial 2	Trial 3
1	1	52.0	52.8	53.1
1	2	51.8	52.8	53.1
1	3	60.0	60.2	58.4
1	4	51.1	52.3	50.3
1	5	50.9	51.8	51.4
1	6	46.4	44.8	49.2
2	1	62.1	62.6	64.0
2	2	59.7	60.0	59.0
2	3	68.6	65.8	69.7
2	4	63.2	62.8	62.2
2	5	64.8	65.0	65.4
2	6	43.7	44.2	43.0
3	1	67.5	67.2	66.9
3	2	61.5	61.7	62.3
3	3	70.8	70.6	71.0
3	4	64.1	66.2	64.0
3	5	72.1	72.0	71.1
3	6	62.0	61.4	60.5

Table 2: Unbalanced Case

Machine	Person	Trial 1	Trial 2	Trial 3
1	1	52.0	.	.
1	2	51.8	52.8	.
1	3	60.0	.	.
1	4	51.1	52.3	.
1	5	50.9	51.8	51.4
1	6	46.4	44.8	49.2
2	1	.	.	64.0
2	2	59.7	60.0	59.0
2	3	68.6	65.8	.
2	4	63.2	62.8	62.2
2	5	64.8	65.0	.
2	6	43.7	44.2	43.0
3	1	67.5	67.2	66.9
3	2	61.5	61.7	62.3
3	3	70.8	70.6	71.0
3	4	64.1	66.2	64.0
3	5	72.1	72.0	71.1
3	6	62.0	61.4	60.5

APPENDIX 2: SAS output for the balanced case

```
                        The SAS System                          1
                General Linear Models Procedure
                    Class Level Information
                   Class    Levels    Values
                   MACHINE       3    1 2 3
                   PERSON        6    1 2 3 4 5 6
               Number of observations in data set = 54
```

```
                        The SAS System                          2
                     General Linear Models Procedure
Dependent Variable: SCORE
                             Sum of          Mean
Source              DF       Squares        Square  F Value    Pr > F
Model               17     3423.68833     201.39343   217.81    0.0001
Error               36       33.28667       0.92463
Corrected Total     53     3456.97500

R-Square           C.V.      Root MSE         SCORE Mean
0.990371       1.612031       0.96158            59.6500

Source              DF    Type III SS   Mean Square  F Value    Pr > F
MACHINE              2     1755.26333     877.63167   949.17    0.0001
PERSON               5     1241.89500     248.37900   268.63    0.0001
MACHINE*PERSON      10      426.53000      42.65300    46.13    0.0001
```

```
                      The SAS System                            3
                General Linear Models Procedure
Source             Type III Expected Mean Square
MACHINE            Var(Error) + 3 Var(MACHINE*PERSON) + Q(MACHINE)
PERSON             Var(Error) + 3 Var(MACHINE*PERSON) + 9 Var(PERSON)
MACHINE*PERSON     Var(Error) + 3 Var(MACHINE*PERSON)
```

```
                      The SAS System                            4
                General Linear Models Procedure
     Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: SCORE

Source: MACHINE
Error: MS(MACHINE*PERSON)
                      Denominator    Denominator
  DF    Type III MS            DF             MS      F Value   Pr > F
   2   877.63166667            10         42.653       20.576   0.0003

Source: PERSON
Error: MS(MACHINE*PERSON)
                      Denominator    Denominator
  DF    Type III MS            DF             MS      F Value   Pr > F
   5        248.379            10         42.653        5.823   0.0089

Source: MACHINE*PERSON
Error: MS(Error)
                      Denominator    Denominator
  DF    Type III MS            DF             MS      F Value   Pr > F
  10         42.653            36   0.9246296296       46.130   0.0001
```

APPENDIX 3: SAS output for the unbalanced case

```
                                The SAS System                              1
                      General Linear Models Procedure
                          Class Level Information
                       Class     Levels    Values
                       MACHINE        3    1 2 3
                       PERSON         6    1 2 3 4 5 6
                  Number of observations in data set = 54
NOTE: Due to missing values, only 44 observations can be used in this
      analysis.
```

```
                                The SAS System                              2
                      General Linear Models Procedure
Dependent Variable: SCORE
                                   Sum of            Mean
Source                  DF        Squares          Square   F Value     Pr > F
Model                   17    3061.743333      180.102549    206.41     0.0001
Error                   26      22.686667        0.872564
Corrected Total         43    3084.430000
                  R-Square           C.V.        Root MSE           SCORE Mean
                  0.992645       1.560754        0.934111             59.85000
Source                  DF    Type III SS     Mean Square   F Value     Pr >
FMACHINE                 2    1238.197626      619.098813    709.52     0.0001
PERSON                   5    1011.053834      202.210767    231.74     0.0001
MACHINE*PERSON          10     404.315028       40.431503     46.34     0.0001
```

```
                                The SAS System                              3
                       General Linear Models Procedure
Source            Type III Expected Mean Square
MACHINE           Var(Error) + 2.137 Var(MACHINE*PERSON) + Q(MACHINE)
PERSON            Var(Error) + 2.2408 Var(MACHINE*PERSON) + 6.7224 Var(PERSON)
MACHINE*PERSON    Var(Error) + 2.3162 Var(MACHINE*PERSON)
```

```
                                The SAS System                              4
                       General Linear Models Procedure
          Tests of Hypotheses for Mixed Model Analysis of Variance
Dependent Variable: SCORE

Source: MACHINE
Error: 0.9226*MS(MACHINE*PERSON) + 0.0774*MS(Error)
                             Denominator     Denominator
    DF      Type III MS               DF              MS       F Value    Pr > F
     2     619.09881279            10.04    37.370383818        16.567    0.0007

Source: PERSON
Error: 0.9674*MS(MACHINE*PERSON) + 0.0326*MS(Error)
                             Denominator     Denominator
    DF      Type III MS               DF              MS       F Value    Pr > F
     5      202.2107668            10.01    39.143708026         5.166    0.0133

Source: MACHINE*PERSON
Error: MS(Error)
                             Denominator     Denominator
    DF      Type III MS               DF              MS       F Value    Pr > F
    10     40.431502803               26    0.8725641026        46.336    0.0001
```