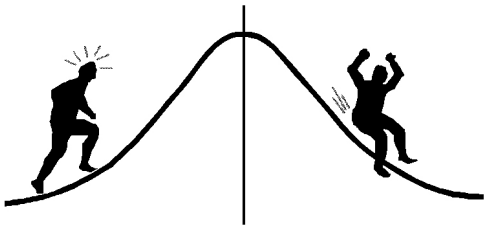

BIOMETRICS INFORMATION

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

PAMPHLET NO. # 45 DATE: May 13, 1993

SUBJECT: Calculating Contrast F-tests when SAS will not [1]

Most experimental studies end up with unbalanced datasets even though a balanced design was planned. The unequal sample sizes (unbalancedness) complicate the analysis when ANOVA and corresponding contrasts are appropriate for data analysis. If the missing data are due to random factors unassociated with the treatments then the original analysis can proceed with some modifications. This pamphlet will describe how to calculate and test contrasts for the general case of unbalanced sample sizes with no completely missing cells. First, we will discuss getting SAS to do the calculations and then how to do them by hand if SAS deems them non-estimable[2].

Generally, we use PROC GLM with CONTRAST statements to do the ANOVA calculations. It is important to consider whether the E = option is required in the CONTRAST statement to specify the correct error term for the contrast. SAS may determine that some of the contrasts are non-estimable (the appropriate error messages can be found in the SAS log). When SAS produces this error message you should check that:

1) the CONTRAST statement has been typed correctly;
2) the contrast coefficients add up to zero;
3) the number of coefficients is less than, or equal to, the number of levels for that source of variation (missing values at the end of a list are treated as zero);
4) the coefficients are in the right order (look at the first page of the PROC GLM output where the factors and their levels are listed in the implicit order used by SAS). This is particularly important for contrasts on interactions where the variables are sorted in the order that they appear on the CLASS statement;
5) all the population marginal means or least square means[3], μ_i, are estimable by submitting the corresponding LSMEANS statement. Milliken and Johnson state that (pg 107): '*If every cell has at least one observation, then all population marginal means are estimable, whereas they are not necessarily estimable if some cells are empty.*' This suggests that if the LSMEANS statement for the source of interest produces least square means that are estimable then any contrast using those means should be estimable. Nevertheless, SAS frequently informs us that those contrasts are non-estimable.

[1] Contrasts have also been discussed in BI #12, 16, 23, and 26.

[2] I do not yet understand all the reasons why SAS will deem some contrasts non-estimable. Non-estimable messages can also occur with balanced but nested designs.

[3] These are means of means and are more clearly defined in the theoretical section below.

Milliken and Johnson also discuss the Means Model and the Effects Model as two different general approaches which can be used with unbalanced data. The Means Model provides the foundation for the following discussion. I feel that this model is simpler and more intuitively appealing. See Milliken and Johnson for more discussion of this model.

If the CONTRAST statement produces a non-estimable error message even though all the marginal means needed for the contrast (i.e. with non-zero contrast coefficients) are deemed estimable by the LSMEANS output, then the calculations can be done by hand using the following steps.

1) Include the LSMEANS statement in the program with the STDERR option;

2) Check whether a TEST statement is necessary to correctly test the source of interest and if so, use the E = option in the LSMEANS statement to specify the correct error term. (This should also be done for any multiple comparisons requested with the MEANS statement. Note that the MEANS statement tests different means than does the LSMEANS statement);

3) Determine the proper coefficients, c_k, for the contrast of interest. Remember that $\sum c_k = 0$;

4) Calculate the contrast $\hat{\gamma} = \sum c_k \hat{\mu}_k$ where the $\hat{\mu}_k$ are obtained from the LSMEANS output under the column titled LSMEAN. **Note that if the output indicates that a mean is non-estimable and it has a non-zero coefficient then the contrast is non-estimable.** This should only occur if there are missing cells in the design;

5) Calculate the contrast's variance by using the standard error values from the LSMEANS output under the column titled Std Err, namely $\text{var}(\hat{\gamma}) = \sum c_k^2 SE_k^2$ where SE_k is the standard error for $\hat{\mu}_k$;

6) Calculate the F-test by: $F = \hat{\gamma}^2/\text{var}(\hat{\gamma}) = \dfrac{(\sum c_k \hat{\mu}_k)^2}{\sum c_k^2 SE_k^2}$ (see note 4 below) .

Notice that the contrast SS's output by the CONTRAST statement is NOT directly used in these calculations. An easy way to calculate them for comparison purposes is by SSC = F*MS, where MS is the proper error term for the source of interest.

Let's work through a two-way factorial for an example. The appendix contains a complete SAS program and output for this analysis. Suppose that the following data have been obtained from a completely randomized design with two treatment (fixed) factors A and B (this example is a modification of Table 9.1 in Milliken and Johnson). The $\hat{\mu}_{ij}$ in the table are the interaction means. Their calculation and the calculation of the main effect (or marginal) means is described in the **theoretical section** below.

---

[4] Note that this formula is the square of the t-value described in BI# 16.

	A = 1			A = 2		
	B = 1	B = 2	B = 3	B = 1	B = 2	B = 3
	19	24	22	25	21	31
	20	26	25	27	24	33
	21		25		24	
			24			
$\hat{\mu}_{ij}$ =	20	25	24	26	23	32
SE_{ij} =	0.816	1.000	0.707	1.000	0.816	1.000

So let's calculate some contrasts for the example and compare the results with the SAS output. Suppose the contrasts of interest have the following contrast coefficents for the interaction means:

	A = 1			A = 2		
	B = 1	B = 2	B = 3	B = 1	B = 2	B = 3
A main effect:						
#1:	1	1	1	-1	-1	-1
B main effects:						
#2:	1	-1	0	1	-1	0
#3:	1	1	-2	1	1	-2
A*B interaction effects:						
#4:	1	-1	0	-1	1	0
#5:	1	1	-2	-1	-1	2

These contrasts correspond to the following questions about the experiment.

#1: Are the means for the two levels of A different?

#2: Are the first two levels of B different?

#3: Are the first two levels of B different from the third level?

#4: Are the first two levels of B different for the two levels of A?

#5: Are the first two levels of B different from the third level at the different levels of A?

Contrast #1 is the main effect for A while #2 & #3 together represent the main effect for B. The remaining contrasts, #4 & #5, represent the interaction of A and B.

The contrast F-test is easily calculated using the information output by the corresponding LSMEANS statement. It is important, when specifying the LSMEANS statement, that the E= option be used if the default error is not the correct denominator for the F-test. In this case, a TEST statement should also be included in the SAS code. The calculations for three of the contrasts of the example are shown on the next page.

Contrast for the A main effect using output from `LSMEANS A / stderr;` :

A	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0
1	23.0000000	0.4906534	0.0001
2	27.0000000	0.5443311	0.0001

The F-value for contrast #1 is calculated by:

$$F = \frac{(\hat{\mu}_{1.} - \hat{\mu}_{2.})^2}{(1)^2*SE_{1.}^2 + (-1^2)*SE_{2.}^2} = \frac{(23 - 27)^2}{0.491^2 + 0.544^2} = 29.79$$

Contrast for the B main effect using output from `LSMEANS B / stderr;` :

B	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0
1	23.0000000	0.6454972	0.0001
2	24.0000000	0.6454972	0.0001
3	28.0000000	0.6123724	0.0001

The F-value for contrast #2 is calculated by:

$$F = \frac{(\hat{\mu}_{.1} - \hat{\mu}_{.2} + 0*\hat{\mu}_{.3})^2}{(1)^2*SE_{.1}^2 + (-1)^2*SE_{.2}^2 + (0)^2*SE_{.3}^2} = \frac{(23 - 24)^2}{2*0.645^2} = 1.20$$

Note that these main effect contrasts have been calculated from the main effect means. They could also have been calculated directly from the interaction means using the larger set of coefficients presented in the above contrast table.

Contrasts for the A*B main effect using output from `LSMEANS A*B / stderr;` :

A	B	Y LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0
1	1	20.0000000	0.8164966	0.0001
1	2	25.0000000	1.0000000	0.0001
1	3	24.0000000	0.7071068	0.0001
2	1	26.0000000	1.0000000	0.0001
2	2	23.0000000	0.8164966	0.0001
2	3	32.0000000	1.0000000	0.0001

The F-value for contrast #5 is calculated by:

$$F = \frac{(\hat{\mu}_{11} + \hat{\mu}_{12} - 2*\hat{\mu}_{13} - \hat{\mu}_{21} - \hat{\mu}_{22} + 2*\hat{\mu}_{23})^2}{SE_{11}^2 + SE_{12}^2 + 4*SE_{13}^2 + SE_{21}^2 + SE_{22}^2 + 4*SE_{23}^2}$$

$$= \frac{(20 + 25 - 2*24 - 26 - 23 + 2*32)^2}{0.816^2 + 1^2 + 4*0.707^2 + 1^2 + 0.816^2 + 4*1^2} = 15.43$$

Exact probability values for these F-values can be obtained by using a simple SAS program described in BI # 15. Note that these values agree with the contrast F-tests on the SAS output in the appendix.

Theoretical Section (optional): This next section describes how the lsmeans and their standard errors are calculated for the example so that you may do them by hand if you prefer not to rely on the LSMEANS statement in SAS. The method and formulae described below are **NOT** obviously transferable to more complex designs. See discussions in SAS For Linear Models and the SAS User's Guide: Statistics. It is most important to realize that the variances or SE's for any lsmean are based on the sample sizes of the means of the very highest order interaction in the model.

Each cell mean of the highest order interaction (two-way interaction for the example) and its standard error are calculated in the usual manner, namely: $\hat{\mu}_{ij} = \sum Y_{ijm}/n_{ij}$

where $\hat{\mu}_{ij}$ = estimated mean for A = i and B = j (i.e. for cell i, j)

Y_{ijm} = one of the m responses in cell i, j

and n_{ij} = number of responses in cell i, j.

The variance or square of the standard error for $\hat{\mu}_{ij}$ is calculated by: $Var(\hat{\mu}_{ij}) = [SE(\hat{\mu}_{ij})]^2 = SE_{ij}^2 = MS/n_{ij}$ where MS is the Mean Square used for testing the A*B interaction and is obtained from the ANOVA table. The MS for the example is 2.000 and so, for example, $SE_{11} = \sqrt{2.000/3} = 0.8164966$.

The Means Model estimates the main effects means for A and B by simply averaging the corresponding cell means, regardless of sample size. For the example the least square means are:

	B = 1	B = 2	B = 3	Least Square[5] Means for A
A = 1:	20	25	24	23.0
A = 2:	26	23	32	27.0
LS Means for B:	23.0	24.0	28.0	25.0

The lsmeans for A, for example, are calculated by: $\hat{\mu}_{i.} = \sum \hat{\mu}_{ij}/b$, where b is the number of levels for factor B. Its variance is calculated by:

$$Var(\hat{\mu}_{i.}) = SE_{i.}^2 = \frac{1}{b^2}\left(\sum_j SE_{ij}^2\right) = \frac{MSE}{b^2}\left(\sum_j (1/n_{ij})\right).$$

In general, a contrast is a weighted sum of means which we have denoted by:

$$\hat{\gamma} = \sum c_k \hat{\mu}_k \quad \text{where the weights } c_k \text{ sum to 0 (i.e. } \sum c_k = 0).$$

[5] Milliken and Johnson call these the *Estimated Population Marginal Means* or marginal means for short.

In unbalanced situations using the Means Model, the $\hat{\mu}_k$ are the lsmeans obtained from SAS with the LSMEANS statement and **NOT** the ordinary means obtained from PROC MEANS or from the MEANS statement in PROC GLM. Using the traditional assumption of independence[6] of the means, μ_k, the variance of a contrast[7] is:

$$\mathrm{Var}(\hat{\gamma}) = \mathrm{Var}(\sum c_k \hat{\mu}_k) = \sum c_k^2 \, \mathrm{Var}(\hat{\mu}_k) = \sum c_k^2 \, \mathrm{SE}^2(\hat{\mu}_k).$$

The required means and their standard errors can be obtained directly from SAS by using the LSMEANS statement, namely `LSMEANS A B A*B / stderr;`. The variances, $\mathrm{Var}(\hat{\mu}_k)$, can be easily obtained from this output by simply squaring the values in the Std Err column beside the means. See the example contrast calculations above.

References:

Milliken, G.A. and D.E. Johnson. 1984. Analysis of Messy Data, Vol. I: Designed Experiments Lifetime Learning Publ., Belmont Calif.

Freund, R.J. & R.C. Littell. 1981. SAS For Linear Models. SAS Institute Inc., Cary, N.C.

SAS Institute Inc. 1988. SAS/STAT User's Guide, Release 6.03 ed. SAS Institute Inc. Cary, N.C.

Contact: Wendy Bergerud
387-5676

NEW PROBLEMS

Calculate the F-tests for contrasts #3 and #4 of the example. Calculate the contrast SS's for all five contrasts. Calculate all the lsmeans and their standard errors. Check your answers with the output in the appendix.

---

[6] Which can be assured by proper random assignment of treatments to experimental units.

[7] or indeed any weighted sum, such as the lsmeans for A above.

Program:

```
/* contrast.sas */

/* From Milliken & Johnson, page 129 -- data modified to increase imbalance */

title 'Calculating contrasts for unbalanced means';
data twoway;
  do a = 1 to 2;
     do b = 1 to 3;
        do rep = 1 to 4;
        input y @@;
  output; end; end; end;
cards;
19 20 21  .
24 26  .  .
22 25 25 24 .
25 27  .  .
21 24 24  .
31 33  .  .
;
proc glm;
  class a b;
  model y = a|b / ss3;
  contrast '#1: A main effect'     a  1 -1;
  contrast '#2: B main effect1'    b  1 -1  0;
  contrast '#3: B main effect2'    b  1  1 -2;
  contrast '#4: A*B 1'            a*b 1 -1  0 -1  1  0;
  contrast '#5: A*B 2'            a*b 1  1 -2 -1 -1  2;
  lsmeans a|b / stderr;        <== see page 4 for output.
  means a b;                   <== compare means with output from LSMEANS.
title2 'Basic GLM run';
run;
```

OUTPUT:

```
                    Calculating contrasts for unbalanced means

                        General Linear Models Procedure
                            Class Level Information

                           Class    Levels    Values

                           A             2    1 2

                           B             3    1 2 3


                    Number of observations in data set = 25


NOTE: Due to missing values, only 16 observations can be used in this analysis.
```

Calculating contrasts for unbalanced means

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	185.9375000	37.1875000	18.59	0.0001
Error	10	20.0000000	2.0000000		
Corrected Total	15	205.9375000			

R-Square	C.V.	Root MSE	Y Mean
0.902883	5.787063	1.414214	24.4375000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	59.58620690	59.58620690	29.79	0.0003
B	2	71.82857143	35.91428571	17.96	0.0005
A*B	2	69.25714286	34.62857143	17.31	0.0006

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
#1: A main effect	1	59.58620690	59.58620690	29.79	0.0003
#2: B main effect1	1	2.40000000	2.40000000	1.20	0.2990
#3: B main effect2	1	69.42857143	69.42857143	34.71	0.0002
#4: A*B 1	1	38.40000000	38.40000000	19.20	0.0014
#5: A*B 2	1	30.85714286	30.85714286	15.43	0.0028

Calculating contrasts for unbalanced means

See page 4 for output from the LSMEANS statement taken from this page.

Calculating contrasts for unbalanced means

General Linear Models Procedure

Level of A	N	Y Mean	Y SD
1	9	22.8888889	2.47206616
2	7	26.4285714	4.23702501

Level of B	N	Y Mean	Y SD
1	5	22.4000000	3.43511281
2	5	23.8000000	1.78885438
3	6	26.6666667	4.32049380