



BIOMETRICS INFORMATION

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

PAMPHLET NO. # 59

DATE: September 16, 1999

SUBJECT: ANOVA: Coefficients for contrasts and means of incomplete factorial designs

Contrasts can always be used to examine scientifically relevant relationships between two or more sets of treatments regardless of the factorial structure. This is particularly useful if some treatment combinations don't occur (a partial factorial). Some simple examples of this were described in Biometrics Information Pamphlets #14 and #45. This pamphlet will extend those discussions to show how to use a diagrammatic method¹ to determine the appropriate coefficients for any desired contrast. This method is also useful for calculating means of interest that would not normally be output by statistical software².

For discussion purposes we will use two sets of treatments or factors, A and B which are assumed to have the same experimental units. Thus any contrasts of their main effects or interaction are all tested by the same error term. Treatments A and B can share the same error term within completely randomized and randomized block designs, and within split-plot designs where they are applied at the same level of split (i.e., they are both mainplot treatment factors, or both split-plot treatment factors). On the other hand, if A is a mainplot treatment while B is a split-plot treatment then they have different error terms and the appropriate error term for any particular contrast may be complicated to determine. This situation is discussed, for instance, in Milliken and Johnson (1984, Chapter 24) and Littell et. al. (1996).

To show how this diagrammatic method works, let's use the example from Biometrics Information Pamphlet #45: an example modified from Milliken and Johnson (1984, page 129). This was an unbalanced two-way design with treatment A having 2 levels and treatment B having 3 levels. We will further modify the data so that cell A = 2 and B = 3 is missing to create a partial factorial design. Suppose that the scientifically relevant relationships under consideration are³:

- Q1) Difference between the levels of A but calculated only for the first two levels of B.
- Q2) Difference between the first level of B and the last two levels, but only within the first level of A.
- Q3) Difference between the first two levels of B within the first level of A and the second level of B within the second level of A.
- Q4) Difference between the differences of the first two levels of B within each level of A (Contrast #4 in Biometrics Information Pamphlet #45).

¹ Described in Littell et. al. (1991), pages 103-104.

² Note that the coefficients obtained are appropriate for the effects form of the ANOVA model while Biometrics Information Pamphlet #45 described coefficients for the means form of the ANOVA model.

³ While an orthogonal set of contrasts would be desirable, a limited set of scientifically relevant contrasts may better match the study objectives. The number of contrasts should remain small to minimize the experiment-wise error rate (most textbooks on ANOVA describe orthogonal contrasts).

Contrasts can be used to test all of these differences against the usual null hypothesis of zero whether or not the cell, $A = 2$ and $B = 3$, was ever included in the study or if all the data for that cell was lost. This is because the coefficient for this cell is always zero within all four contrasts.

The easiest way to determine the required contrast coefficients is to use the diagrammatic method. The first step is to create a table showing the factorial arrangement. Cells corresponding to missing treatment combinations should be marked with an 'X'. All contrasts must work around the X's, treating each X as a zero coefficient. For the example, an X is placed in cell $A = 2$, $B = 3$ since that treatment combination is missing.

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1			
2			
3		X	
Sum:			

The next step is to put the appropriate contrast coefficients into the various cells of the table. For question number 1 (Q1), the coefficients are⁴:

Q1:

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1	1	-1	
2	1	-1	
3	0	X	
Sum:			

Finally, sum up the column and row coefficients:

Q1:

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1	1	-1	0
2	1	-1	0
3	0	X	0
Sum:	2	-2	divisor = 2

The column totals, 2 and -2 are the contrast coefficients for the A main effects while the row totals, 0, 0, 0 are the contrast coefficients for the B main effects. The numbers in the body of the table are the contrast coefficients for the A×B interaction effects⁵. Notice that the divisor is the sum of the positive numbers within the table.

⁴ If you are not sure where these numbers come from, read Biometrics Information Pamphlet #16 or look for discussions about contrasts or pre-planned comparisons in any ANOVA textbook.

⁵ The intercept is also an effect in the effects model but is not explicitly included in this table. It has a coefficient of zero for contrasts, but not for means. We shall see this more clearly when we use the ESTIMATE statement to estimate specific least square means.

Contrasts can be calculated using PROC's GLM and/or MIXED using either the ESTIMATE⁶ or CONTRAST statements. The resulting statistical tests are essentially the same for either statement. The CONTRAST or ESTIMATE statement requires that coefficients be specified for all effects which do not have all zero values (otherwise, zeroes are assumed), namely:

```
** Assuming order of levels used is7
                                     A = 1      A = 2
                                     B = 1, 2, 3  B = 1, 2 ;
Contrast 'Q1:A=1 vs 2 w B=1&2' A 2 -2 A*B 1 1 0      -1 -1 ;
Estimate 'Q1:A=1 vs 2 w B=1&2' A 2 -2 A*B 1 1 0      -1 -1 / divisor = 2;
```

Due to limited space, I have used a 'w' in the statement label to stand for 'within'.

Notice that if A 2 -2 was left out of the CONTRAST or ESTIMATE statement then it would become non-estimable. This would also happen if the order of the numbers was reversed without also reversing the signs of the A×B coefficients.

For the second contrast, the contrast coefficient table and corresponding SAS code are:

Q2:

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1	2	0	2
2	-1	0	-1
3	-1	X	-1
Sum:	0	0	divisor = 2

```
** Assuming order of levels used is
                                     A = 1      A = 2
                                     B = 1, 2, 3  B = 1, 2 ;
Contrast 'Q2:B=1 vs 2&3 w A=1' B 2 -1 -1 A*B 2 -1 -1      0 0 ;
Estimate 'Q2:B=1 vs 2&3 w A=1' B 2 -1 -1 A*B 2 -1 -1      0 0 / divisor = 2;
```

Note that the trailing zeroes at the end of the A×B-list of coefficients could be left out.

For the third contrast, the table and corresponding SAS code are:

Q3:

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1	-1	0	-1
2	-1	2	1
3	0	X	0
Sum:	-2	2	divisor = 2

⁶ The ESTIMATE statement is generally used to estimate a weighted sum of the effects model parameters, where the weights are given by the coefficients in the above table and the parameters are those of the various effects: intercept, A, B and their interaction A×B. The weights given by the table for Q1 will estimate a weighted *sum* of the treatment means. To estimate a contrast, which is a weighted *mean* of the treatments means, the integer weights must be converted to fractions. While fractions could be used in the ESTIMATE statement, it is easier to use the divisor option. For this example, divisor = 2. This results in calculating the difference between two *means* of means instead of a difference between two *sums* of means.

⁷ Within SAS this order is determined by how A and B are listed in the CLASS statement and by the sort order of their levels. This can be checked by looking at the first page output by the procedures.

```

** Assuming order of levels used is          A = 1          A = 2
                                           B = 1, 2, 3  B = 1, 2 ;
Contrast 'Q3:B=1&2 wA1 vs B=2wA2' A -2 2 B -1 1 0 A*B -1 -1 0    0 2 ;
Estimate 'Q3:B=1&2 wA1 vs B=2wA2' A -2 2 B -1 1 0 A*B -1 -1 0    0 2 / divisor = 2;

```

The first three contrasts have required that coefficients for the interaction and one or both of the main effects must be specified in the CONTRAST and/or ESTIMATE statements. The fourth contrast is an example that does not. It is the 2 by 2 interaction between the first two levels of B and the two levels of A. Since it is a contrast solely of the interaction between A and B the coefficients for both main effects cancel out. This contrast can be thought of as a test for the difference of two differences between means, namely the means of cells: $(A1B1 - A1B2) - (A2B1 - A2B2)$. For this last contrast, the table and corresponding SAS code are:

Q4:

Treatment B	Treatment A		Sum:
	A = 1	A = 2	
1	1	-1	0
2	-1	1	0
3	0	X	0
Sum:	0	0	divisor = 2

```

** Assuming order of levels used is          A = 1          A = 2
                                           B = 1, 2, 3  B = 1, 2 ;
Contrast 'Q4:B=1 vs 2 same for A'          A*B 1 -1 0          -1 1 ;
Estimate 'Q4:B=1 vs 2 same for A'          A*B 1 -1 0          -1 1 / divisor = 2;

```

The SAS program to read in the data was presented in the Appendix of Biometrics Information Pamphlet #45⁸. The additional SAS code from above produces the following output⁹, with the results of the CONTRAST statements presented before those of the ESTIMATE statements.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Q1:A=1 vs 2 w B=1&2	1	9.6000000	9.6000000	4.80	0.0562
Q2:B=1 vs 2&3 w A=1	1	38.8800000	38.8800000	19.44	0.0017
Q3:B=1&2 wA1 vs B=2wA2	1	0.4615385	0.4615385	0.23	0.6424
Q4:B=1 vs 2 same for A	1	38.4000000	38.4000000	19.20	0.0018

Parameter	Estimate ¹⁰	T for H0: Parameter=0	Pr > T	Std Error of Estimate
Q1:A=1 vs 2 w B=1&2	-2.00000000	-2.19	0.0562	0.91287093
Q2:B=1 vs 2&3 w A=1	-4.50000000	-4.41	0.0017	1.02062073
Q3:B=1&2 wA1 vs B=2wA2	0.50000000	0.48	0.6424	1.04083300
Q4:B=1 vs 2 same for A	-4.00000000	-4.38	0.0018	0.91287093

⁸ A statement must be added to the DATA step to delete the data for cell A=2 and B=3.

⁹ The mean square error, its df, and resulting standard error may be a little different from that of Biometrics Information Pamphlet #45 because of the missing data.

¹⁰ See the Appendix for a description of how to check that the parameter estimates output by the ESTIMATE statement are correct.

Notice that the p-values for each question, whether calculated via a CONTRAST or the ESTIMATE statement, are identical. This is not surprising since the tests are identical although one is a t-test and the other an F-test¹¹.

So far, we have discussed how the diagrammatic method can be used to correctly code for contrasts. Let us now consider how to use it to calculate means using the ESTIMATE statement. To start, let us suppose that we want to summarize the results of our example study with a table of the least square means:

Treatment B:	Treatment A		LSmean:
	A = 1	A = 2	
B = 1	20.0† (0.816)	26.0 (1.00)	23.0 (0.645)
B = 2	25.0 (1.000)	23.0 (0.816)	24.0 (0.645)
LSmean: (SE):	22.5 * (0.645)	24.5 * (0.645)	23.5 * (0.456)
B = 3	24.0 (0.707)	na	non-est
LSmean: (SE):	23.0 (0.491)	Non-est	23.6 *‡ (0.392)

na = not available; non-est = non-estimable

† For A by B means the SE is variable because of differing sample sizes

* Standard errors requiring a separate ESTIMATE statement.

‡ Calculated for all five existing A by B means

Most of these means can be calculated using the using an LSMEANS¹² statement. The calculations for four of the means, indicated by an ‘*’, cannot be obtained this way and we must use an ESTIMATE statement instead. The coefficients for the ESTIMATE statements can be determined by using the diagrammatic method. As an example, let’s look at the diagram for the mean of all 5 of the A×B means:

Grand Mean:	Treatment	Treatment A		Sum:
	B	A = 1	A = 2	
1	1	1	1	2
2	1	1	1	2
3	1	1	X	1
Sum:	3	3	2	divisor = 5

To avoid fractions, a ‘1’ has been put into each cell used to calculate the specified mean. Then the rows and columns are added and the total of all the cell coefficients are put into the divisor box. The divisor converts the integer coefficients into fractions so that a mean of the means is calculated (instead of a sum of the means). Since none of the rows or columns have all zeroes the

¹¹ See Biometrics Information Pamphlet # 27: When the t-test and F-test are equivalent.

¹² It would look like: LSMEANS A B A*B ;

ESTIMATE statement will need to include all the effects: the intercept, A, B and A×B. The intercept is required in this case since it is not subtracted out when calculating a mean (as occurs when calculating contrasts). The SAS code for all four means is:

```
Estimate 'All 5 means'          intercept 5  A 3 2  B 2 2 1  A*B 1 1 1 1 1 / divisor=5;
Estimate 'Mean of A=1 for B=1,2' intercept 2  A 2 0  B 1 1 0  A*B 1 1 0 0 0 / divisor=2;
Estimate 'Mean of A=2 for B=1,2' intercept 2  A 0 2  B 1 1 0  A*B 0 0 0 1 1 / divisor=2;
Estimate 'Mean of A=1,2 & B=1,2' intercept 4  A 2 2  B 2 2 0  A*B 1 1 0 1 1 / divisor=4;
```

CAUTIONS:

- 1) PROC GLM will correctly calculate the standard errors for means (using either the LSMEANS or the ESTIMATE statements) only for completely randomized designs, with and without subsampling. Designs with a random block or other random effects, including split-plot designs, should use PROC MIXED instead for standard error calculations.
- 2) Further, PROC GLM will correctly calculate contrast tests for partial factorial designs when the factors involved share the same error term, otherwise PROC MIXED should be used.
- 3) The coefficients determined here are appropriate for any software which uses an effects ANOVA model with a design matrix parameterized in the same way as does SAS's PROC's GLM and MIXED (with 0's and 1's). They would need to be modified if the software uses a different parameterization (for example, SAS's PROC CATMOD uses 0's, 1's and -1's).

Prepared by:

Wendy Bergerud. Phone: 250-387-5676; e-mail: wendy.bergerud@gems8.gov.bc.ca
B.C. Ministry of Forests, Research Branch.

References:

- Littell, R. C., Freund, R., and Spector, P. C., 1991, SAS[®] system for linear models, 3rd ed., SAS Institute Inc., Cary, NC.
- Littell, R.C., Milliken, G. A., Stroup, W.W., and Wolfinger, R. D., 1996, SAS[®] system for mixed models, SAS Institute Inc., Cary, NC.
- Milliken, G.A. and D. E. Johnson, 1984, Analysis of messy data, Vol. I: Designed experiments, Van Nostrand Reinhold, NY, NY.

Appendix:

We can check that the ESTIMATE statements for the contrasts have worked the way we expected by calculating the parameter estimates by hand. We can do this using the estimated A×B lsmeans from the summary table (on page 5). The corresponding hand calculations are:

$$\begin{array}{l}
 \text{Q1: } A = 1 \text{ vs } 2 \text{ w } B=1\&2: \quad (20+25)/2 - (26+23)/2 = 22.5 - 24.5 = -2.0 \\
 \text{Q2: } B = 1 \text{ vs } 2 \& 3 \text{ w } A=1: \quad 20 - (25+24)/2 = 20 - 24.5 = -4.5 \\
 \text{Q3: } B = 1\&2 \text{ w } A1 \text{ vs } B=2\text{w}A2: \quad (20 + 25)/2 - 23 = 22.5 - 23 = 0.5 \\
 \text{Q4: } B = 1 \text{ vs } 2 \text{ same for } A: \quad (20 - 25)/2 - (26-23)/2 = -2.5 - 1.5 = -4.0
 \end{array}$$

These results match those of the output on page 4.