How SAS does it.

- **Another option: Force the <u>last</u> $\tau_i = 0$.**

- **These options give different numerical estimates for the parameters, but all conclusions about <u>effects and contrasts</u> will be <u>the same</u>.**

<u>Unbalanced Data</u>

- **Using the standard ANOVA formulas is easy, but it will give wrong results when data are unbalanced (different numbers of observations across cells).**

- **Dummy variable approach always gives correct answers.**

<u>Illustration</u>: A unbalanced 2-factor factorial study. (Table 11.3 data, p. 585)

A \ C	1	2
1	4,5,6	8
2	5	7,9

- **<u>Question</u>: Does factor A have a significant effect on the response? (For simplicity, <u>ignore</u> any interaction between A and C for this example).**

Recall: Our F-statistic formula for this type of test was:

$$F^* = \frac{MSA}{MSW} = \frac{SSA/(a-1)}{SSW/ac(n-1)}$$

and SSA = $cn\sum_i (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$

- **This formula is based on the variation between the marginal means $\bar{Y}_{1\cdot\cdot}$ and $\bar{Y}_{2\cdot\cdot}$**

- **For the Table 11.3 data:**

$$\bar{Y}_{1\bullet\bullet} = \frac{4+5+6+8}{4} = 5.75$$

$$\bar{Y}_{2\bullet\bullet} = \frac{5+7+9}{3} = 7$$

→ Based on this, there is <u>some</u> sample variation between the means for levels 1 and 2 of factor A.

- **However, let's look at the sample means for levels 1 and 2 of A, <u>separately at each level of C</u>:**

For level 1 of C:

$$\bar{Y}_{11\bullet} = \frac{4+5+6}{3} = 5$$

$$\bar{Y}_{21\bullet} = 5$$

For level 2 of C:

$$\bar{Y}_{12\bullet} = 8$$

$$\bar{Y}_{22\bullet} = \frac{7+9}{2} = 8$$

- **These results imply that (at each level of C) there is <u>no</u> sample variation between the means for levels 1 and 2 of factor A.**

- **Which conclusion is correct?**

- **Our model is (recall there is no interaction term):**

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$$

Note: $\bar{Y}_{11\bullet} - \bar{Y}_{21\bullet}$ **is an estimate of:** $E(Y_{11k}) - E(Y_{21k}) =$

$$(\mu + \alpha_1 + \gamma_1) - (\mu + \alpha_2 + \gamma_1) = \alpha_1 - \alpha_2$$

Also, $\bar{Y}_{12\bullet} - \bar{Y}_{22\bullet}$ **is an estimate of:** $E(Y_{12k}) - E(Y_{22k}) =$

$$(\mu + \alpha_1 + \gamma_2) - (\mu + \alpha_2 + \gamma_2) = \alpha_1 - \alpha_2$$

- **So these <u>do</u> estimate the true difference in the means for levels 1 and 2 of factor A.**

But ... $\bar{Y}_{1\bullet\bullet} - \bar{Y}_{2\bullet\bullet}$**, for these data, is:**

$$\frac{3\bar{Y}_{11\bullet} + \bar{Y}_{12\bullet}}{4} - \frac{\bar{Y}_{21\bullet} + 2\bar{Y}_{22\bullet}}{3}$$

which estimates:

$$\frac{3}{4}(\mu + \alpha_1 + \gamma_1) + \frac{1}{4}(\mu + \alpha_1 + \gamma_2) - \frac{1}{3}(\mu + \alpha_2 + \gamma_1) - \frac{2}{3}(\mu + \alpha_2 + \gamma_2)$$

$$= \alpha_1 - \alpha_2 + \frac{5}{12}\gamma_1 - \frac{5}{12}\gamma_2$$

- **This is <u>not</u> the true difference in factor A's level means that we wanted to estimate.**

- **For balanced data, the magnitudes of all the coefficients would be the same and everything would cancel out properly.**

- **With unbalanced data, we need to adjust for the fact that the various <u>cell means</u> are based on <u>different numbers of observations</u> per cell.**

- **Using a dummy variable regression model implies the effect of factor A is estimated <u>holding factor C constant</u> → produces correct results.**

- **Analysis for unbalanced data involves the <u>least squares means</u>, not the ordinary <u>factor level means</u>.**

LS mean for level 1 of A is: $\frac{5+8}{2} = 6.5$

- **The least squares mean (for, say, level 1 of factor A) is the <u>unweighted average</u> of the cell sample means corresponding to level 1 of factor A. With unbalanced data, this is different than simply averaging all response values for level 1 of factor A. (see example)**

LS mean for level 2 of A is: $\frac{5+8}{2} = 6.5$

- **With unbalanced data in the two-way ANOVA, our F-tests about the factors use the Type III sums of squares, rather than the ordinary (Type I) ANOVA SS.**

- **See example for calculating these F-statistics correctly.**

Example: (Table 11.3 data)

- **Least squares means:**

for Level 1 of Factor A: $\frac{5+8}{2} = 6.5$

for Level 2 of Factor A: $\frac{5+8}{2} = 6.5$

for Level 1 of Factor C: $\frac{5+5}{2} = 5$

for Level 2 of Factor C: $\frac{8+8}{2} = 8$

- **Correct F-tests about factor effects:**

- Done based on Type III SS which uses a dummy variable regression model and full vs. reduced model F-tests.
- No significant A×C interaction.
- Factor A main-effects F-test had a P-value of 1.
- Factor C main-effect F-test had a P-value of .054

} based on Type III SS.

- **More complicated example: Suppose A has 3 levels and C has 2 levels.**

- **Now we need to use 3 – 1 = 2 dummy variables for A and 2 – 1 = 1 dummy variable for C.**

Example:

Factor A	Factor C = 1	Factor C = 2
1	4, 5, 6	8
2	5	7, 9
3	11	12

Again, Type III SS does the dummy variable regression approach:

A×C interaction was not significant.

Factor A main-effects F-test had P-value .0305.

Factor C main-effects F-test had P-value .0703.