UNBALANCED DESIGNS

Unbalanced experimental design:

Sample sizes for the treatment combinations are not all equal.

Reasons why balanced designs are better:

- The test statistic is less sensitive to small departures from the equal variance assumption.
- The power of the test is largest when sample sizes are equal.

Reasons why you may need to be able to work with unbalanced designs:

- Balanced designs produce unbalanced data when something goes wrong:
  Plants die
  Machinery breaks down
  Shipments of raw materials don't come in on time
  Operators or subjects get sick
  Etc.
- Some treatments may be more expensive or more difficult to run than others.
- Some treatment combinations may be of particular interest, so the experimenter chooses to sample more heavily from them.

Ways of analyzing unbalanced designs:

- A minor variation to the method for balanced designs works for "proportional" data:

$$r_{ij} = (r_{i\bullet}r_{\bullet j})/r_{\bullet\bullet}$$

(See Montgomery, p. 601 for details.)

- Various approximate procedures for data that are only slightly unbalanced:

Estimating missing observations.

Omitting observations from cells with larger sizes.

Methods involving adjusting weights.

(See Montgomery, pp. 601- 603 for details.)

- The "Exact Method": Representing the analysis of variance model as a regression model.

This is the only method we will discuss for unbalanced factorial designs.

Cautions:

- The same problem might be done in more than one way, resulting in different sums of squares.
- The hypotheses tested might be different from those tested in balanced ANOVA.
- The tests sometimes create their own problems in interpretation.

In Minitab and many other software packages:

Use *General Linear Model* ("GLM")
(a regression approach)

Illustration with Battery data:

1. Use Balanced ANOVA:

Analysis of Variance for LPUC

Source	DF	SS	MS	F	P
duty	1	252004	252004	106.43	0.000
brand	1	124609	124609	52.63	0.000
duty*brand	1	51302	51302	21.67	0.000
Error	12	28413	2368		
Total	15	456328			

2. Use GLM, specifying "dutylbrand" or "duty brand duty*brand":

Analysis of Variance for LPUC

Source	DF	Seq SS	Adj SS	Adj MS	F	P
duty	1	252004	252004	252004	106.43	0.000
brand	1	124609	124609	124609	52.63	0.000
duty*brand	1	51302	51302	51302	21.67	0.000
Error	12	28413	28413	2368		
Total	15	456328				

Note: Outputs are identical, *except* GLM output has an additional column "Adj SS" repeating all but the last line of the SS column.

3. Use GLM, specifying "duty*brand duty brand":

Analysis of Variance for LPUC

Source	DF	Seq SS	Adj SS	Adj MS	F	P
duty*brand	1	51302	51302	51302	21.67	0.000
duty	1	252004	252004	252004	106.43	0.000
brand	1	124609	124609	124609	52.63	0.000
Error	12	28413	28413	2368		
Total	15	456328				

Note: Output is the same as in (2), *except* the order of the rows is different.

This reflects the order in which the variables were entered into the software.

Battery data with last row of data deleted:

Sample sizes are now unequal: the treatment "heavy-duty, name brand" only has three observations, while the other three treatment combinations still have four observations.

1. Using Balanced ANOVA:

```
 * ERROR * Unequal cell counts.
```

2. Using GLM, specifying "duty brand duty*brand " or "duty|brand":

Analysis of Variance for LPUC

Source	DF	Seq SS	Adj SS	Adj MS	F	P
duty	1	221585	226482	226482	89.36	0.000
brand	1	123214	110720	110720	43.69	0.000
duty*brand	1	50185	50185	50185	19.80	0.001
Error	11	27879	27879	2534		
Total	14	422863				

Note: The Seq SS and Adj SS columns are no longer the same! This is typical for unequal sample sizes.

3. Using GLM, specifying "duty*brand duty brand":

Analysis of Variance for LPUC

Source	DF	Seq SS	Adj SS	Adj MS	F	P
duty*brand	1	80282	50185	50185	19.80	0.001
duty	1	203982	226482	226482	89.36	0.000
brand	1	110720	110720	110720	43.69	0.000
Error	11	27879	27879	2534		
Total	14	422863				

Comparing to case (2):

- Rows are in a different order.
- The sums of squares in the Seq SS column corresponding to each term are different (except for error)
- The sums of squares in the Adj SS column are the same. This contrasts with the case of equal sample sizes, where both columns (Seq SS and Adj SS) were the same.

How GLM Works

GLM first creates a *design matrix*. For the full battery data:

```
Matrix XMAT1

  1   1   1   1
  1   1  -1  -1
  1   1   1   1
  1  -1  -1   1
  1   1   1   1
  1   1   1   1
  1   1  -1  -1
  1  -1   1  -1
  1  -1  -1   1
  1   1  -1  -1
  1   1  -1  -1
  1  -1   1  -1
  1  -1  -1   1
  1  -1   1  -1
  1  -1  -1   1
  1  -1   1  -1
```

(For the battery data with the last observation deleted, the design matrix is obtained from this one by deleting the last row.)

Notice about the design matrix:

- 4 columns
- 16 rows -- one row for each observation.
- The first column is all 1's.
- The second has 1 for each observation with duty = 1 and -1 for each observation with duty = 2.
- The third column has 1 for each observation with brand = 1 and -1 for each observation with brand = 2.
- The fourth column is the product of the second and third columns.

This is the matrix describing a regression using constant term and regressors X_1, X_2, and X_1*X_2, where X_1 is an indicator variable defined by

$$X_1 = \begin{cases} 1, \text{ if duty} = 1 \\ -1, \text{ if duty} = 2, \end{cases}$$

and X_2 is an indicator variable defined similarly for the factor brand.

Minitab performs a regression on the response variable (LPUC in this example) with predictor variables corresponding to the columns of the design matrix (constant, X_1, X_2, and X_1*X_2 in this example).

The *sequential sum of squares* (in the column Seq SS) is:

- (in regression terms) the sum of the sums of squares for all indicator variables corresponding to the item listed, given all terms corresponding to items previously listed. (In the battery example, there is only one indicator variable for each of the items duty, brand, and duty*brand.)

- (in analysis of variance terms) the sum of squares obtained by taking the difference between the sum of squares for error for the model including all previously listed items (reduced model) with the one obtained by adding the new item to those previously listed.

The *adjusted sum of squares* (in the column Adj SS) is:

- (in regression terms) the sum of the sums of squares for all indicator variables corresponding to the item, given all the other items.

- (in analysis of variance terms) the difference in error sums of squares when comparing the full model with the reduced model obtained by omitting the item in question.

For equal sample sizes: The columns of the design matrix are uncorrelated (dot product as vectors is 0). This leads to the two sums of squares columns being equal.

Deleting the last row (in the battery example) resulted in matrix with columns that have non-zero correlation.

For unbalanced data, the *adjusted sum of squares* is the one that is important for hypothesis testing.

The *adjusted mean square* is obtained by dividing the corresponding adjusted sum of squares by its degrees of freedom. The resulting F-statistic is then this mean square divided by the mean square for error.

Example: Battery data with one observation deleted.