

## TWO-WAY ANOVA FOR UNEQUAL SAMPLE SIZES

This analysis needs to be done by the General Linear Model (GLM) command, which essentially uses a regression approach. This will be explained first by looking at an example.

### 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 "duty|brand" 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				

Notice that the outputs are identical, except that the GLM output has an additional column "Adj SS" that repeats 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 that the output is the same as in (2), except that the order of the rows is different. This reflects the difference in specifying the model: in (3), `duty*brand` was specified before `duty` and `brand`, so is listed before them.

### Battery data with last row of data deleted:

We now have unequal sample sizes -- 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				

Notice that 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 this case to case (2), we see that in addition to having a different order to the rows, the sums of squares in the Seq SS column corresponding to each term are different (except for error), but are still the same in the Adj SS column. This difference contrasts with the case of equal sample sizes, where both columns were the same.

### How GLM Works

GLM first creates a *design matrix*. For the full battery data, this matrix is

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 that the design matrix has 4 columns and 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.

Those of you who have had regression will recognize this as 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 (here, LPUC) with predictor variables corresponding to the columns of the design matrix ( here,  $X_1$ ,  $X_2$ , and  $X_1 * X_2$ ).

- 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 variable listed, given all terms corresponding to variables previously listed. (In the battery example, there is only one indicator variable for each of the terms duty, brand, and duty\*brand.) In analysis of variance terms: the sequential sum of squares is the sum of squares obtained by taking the difference between the sum of squares for error for the model including all previously listed variables (reduced model) with the one obtained by adding the new variable 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 term, given all the other terms. In analysis of variance terms: the adjusted sum of squares is the difference in error sums of squares when comparing the full model with the reduced model obtained by omitting the variable in question.

The reason that the two columns are equal for equal sample sizes turns out to be that the columns of the design matrix are uncorrelated; deleting the last row (in the battery example) resulted in 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.