Lesson 1.1. The nature of unbalanced data and method of unweighted means

The nature of unbalanced data

Recall the analysis of balanced factorial designs—that is, cases where there are an equal number of observations n in each combination of the factors (cell). However, it is not unusual to encounter situations where the number of observations in the cells is unequal.

These unbalanced factorial designs occur for various reasons. For example, the experimenter may have designed a balanced experiment initially, but because of unforeseen problems in running the experiment, resulting in the loss of some observations, he or she ends up with unbalanced data. On the other hand, some unbalanced experiments are deliberately designed that way. For instance, certain treatment combinations may be more expensive or more difficult to run than others, so fewer observations may be taken in those cells. Alternatively, some treatment combinations may be of greater interest to the experimenter because they represent new or unexplored conditions, and so the experimenter may elect to obtain additional replication in those cells.

The orthogonality property of main effects and interactions present in balanced data does not carry over to the unbalanced case. This means that the usual analysis of variance techniques do not apply. Consequently, the analysis of unbalanced factorials is much more difficult than that for balanced designs. Methods for dealing with unbalanced factorials, concentrating on the case of the two-factor fixed effects model will be our focus here.

The case of proportional data

One situation involving unbalanced data presents little difficulty in analysis; this is the case of proportional data. That is, the number of observations in the ij^{th} cell is given by

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

This condition implies that the number of observations in any two rows or columns is proportional. When proportional data occur, the standard analysis of variance can be employed with minor modifications in the computing formulas.

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} Y_{ijk}^2 - \frac{Y_{...}{}^2}{n_{..}}$$

$$SSA = \sum_{i=1}^{a}\frac{Y_{i..}^2}{n_{i.}} - \frac{Y_{...}{}^2}{n_{..}}$$

$$SSB = \sum_{j=1}^{b}\frac{Y_{.j.}^2}{n_{.j}} - \frac{Y_{...}{}^2}{n_{..}}$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{Y_{ij.}^2}{n_{ij}} - \frac{Y_{...}{}^2}{n_{..}} - SSA - SSB$$

$$SSE = SST - SSA - SSB - SSAB$$

Example:

An engineer is designing a battery for use in a device that will be subjected to some extreme variations in temperature. He uses three types of plate material for the battery. Since battery life is affected by temperature, he decides to test all three plate materials at three temperature levels which are consistent with the product end-use environment. Suppose the data (battery life) collected from the test experiment are as follows:

Material Type	Temperature (°F)		
	15	70	125
1	130 155 74 180	34 40 80 75	70 58
2	159 126	136 115	45
3	138 160	150 139	96

The ANOVA table for this data is shown below [VERIFY!].

Source of Variation	df	Sum of squares	Mean squares	F
Material (A)	2	7811.60	3905.80	4.784
Temperature (B)	2	16090.88	8045.44	9.854
A x B	4	6266.53	1566.63	1.919
Residual	11	8981.00	816.45	
Total	19	39150		

Approximate methods for disproportionate class frequencies

When unbalanced data are not too far from the balanced case, it is sometimes possible to use approximate procedures that convert the unbalanced problem into a balanced one. This, of course, makes the analysis only approximate.

In practice, we must decide when the data are not sufficiently different from the balanced case to make the degree of approximation introduced relatively unimportant. We assume

that every cell has at least one observation (i.e. $n_{ij} \geq 1$). See examples below.

The n_{ij} Values for an Unbalanced Design

	Columns		
Rows	**1**	**2**	**3**
1	4	4	4
2	4	3	4
3	4	4	4

The n_{ij} Values for an Unbalanced Design

	Columns		
Rows	**1**	**2**	**3**
1	4	4	4
2	4	5	4
3	4	4	4

Approximate methods include (1) estimating missing Observations, (2) setting data aside, and (3) the method of unweighted means.

A. **Estimating missing observations**

Estimating missing observations is reasonable if only a few n_{ij}'s are different. For a model with interaction, the estimate of the missing value in the ij^{th} cell that minimizes the error sum of squares is $\overline{y}_{ij.}$. That is, we estimate the missing value by taking the average of the observations that are available in that cell. The estimated value is treated just like actual data. The calculations of the sums of squares are the same, but we have reduce the error degrees of freedom by the number of missing observations that have been estimated.

Example:

Consider the following data. estimate the missing observation using the cell mean and construct the analysis of variance table. Is there a reason to believe that factors A and B interact with each other?

Factor A	**Factor B**					
	1			**2**		
1	1.2	2.8	3.2	1.9	3.4	3.4
2	1.8	3.0	◯	2.6	2.8	1.9

B. **Setting data aside**

Consider the data below. Note that cell (2,2) has one more observation than the others. Estimating missing values for the remaining eight cells is probably not a good idea here because this would result in estimates constituting about 88 percent of the final data. An alternative is to set aside one of the observations in cell (2,2), giving a balanced design with $n = 4$ replicates.

The observation that is set aside should be chosen randomly. Furthermore, rather than completely discarding the observation, we could return it to the design, and then randomly

choose another observation to set aside and repeat the analysis. And, we hope, these two analyses will not lead to conflicting results. If they do, we inspect if the observation that was set aside is an outlier and should be handled accordingly. In practice, this confusion is unlikely to occur when only small numbers of observations are set aside and the variability within the cells is small.

The n_{ij} Values for an Unbalanced Design

	Columns		
Rows	**1**	**2**	**3**
1	4	4	4
2	4	5	4
3	4	4	4

Example:

Consider the following data. Is there a reason to believe that factors A and B interact with each other?

Factor A	Factor B	
	1	**2**
1	2 2 4 5	2 3 4 3
2	3 2 2 4	4 3 3 5 5
3	3 5 6 3	4 6 4 4

C. **Method of unweighted means**

In this approach, introduced by Yates (1934), the cell averages are treated as data and are subjected to a standard balanced data analysis to obtain sums of squares for rows, columns, and interaction. This approach is considered approximate because the sums of squares are not distributed as chi-square random variables. The method is the most computationally simple, but works reasonably well only when the cell frequencies n_{ij} are not too far apart.

Example:

Consider the following data.

Material Type	Temperature (°F)		
	15	70	125
1	130 155 74	34 40 80	70 58
2	159 126	136 115	45
3	138 160	150 139	96

The first thing we need to do is compute the table of totals and means. This is shown below.

Material Type	Temperature (°F) 15	70	125	Total of Means (Material Type)
1	359 (119.67)	154 (51.33)	128 (64)	(235)
2	285 (142.5)	251 (125.5)	45 (45)	(313)
3	298 (149)	289 (144.5)	96 (96)	(389.5)
Total of Means (Temp)	(411.17)	(321.33)	(205)	1905 (937.5)

A. Calculations based on the observed Y_{ij}'s:

$$CF = \frac{1905^2}{18} = 201612.5$$

$$SST = (130^2 + 155^2 + \cdots + 96^2) - CF = 234405 - 201612.5 = 32792.5$$

$$SSTr = \left(\frac{359^2}{3} + \frac{154^2}{3} + \cdots + \frac{96^2}{1}\right) - CF = 228574.17 - 201612.5 = 26961.67$$

$$SSE = SST - SSTr = 32792.5 - 26961.67 = 5830.83$$

$$MSE = \frac{SSE}{\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1)} = \frac{5830.83}{9} = 647.87$$

B. Calculations based on cell means:

$$CF' = \frac{937.5^2}{9} = 97656.25$$

$$SST' = (119.67^2 + 51.33^2 + \cdots + 96^2) - CF' = 111430.43 - 97656.25 = 13774.18$$

$$SSA' = \left(\frac{235^2}{3} + \frac{313^2}{3} + \frac{389.5}{3}\right) - CF' = 101634.75 - 97656.25 = 3978.5$$

$$SSB' = \left(\frac{411.17^2}{3} + \frac{321.33^2}{3} + \frac{205^2}{3}\right) - CF' = 104779.38 - 97656.25 = 7123.13$$

$$SSAB' = SST' - SSA' - SSB' = 13774.18 - 3978.5 - 7123.13 = 2672.55$$

$$MSE' = \frac{MSE}{\overline{n}} = \frac{647.87}{0.57}$$

where

$$\overline{n} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}} = \frac{1}{3(3)}\left(\frac{1}{3}+\frac{1}{3}+\cdots+\frac{1}{1}\right) = 0.57$$

SoV	df	SS	MS	F	p
Mat (A')	2	3978.50	1989.25	1.75	0.2280
Temp (B')	2	7123.13	3651.57	3.13	0.0930
A' x B'	4	2672.55	668.14	0.59	0.6780
Error'	9		1136.61		