Stat 526
5-27-25

Unbalanced data in factorial designs

①

		Temp (°F) 15	70	125
Material Type	1	130 155 74 180	34 40 80 75	70 58
	2	159 26	136 115	45
	3	138 160	150 139	96

②

Look at the cell counts

4	4	2	10
2	2	1	5
2	2	1	5
8	8	4	20

Notice that the rows are proportional

and the columns are proportional

and $n_{ij} = \dfrac{n_{i\cdot}\, n_{\cdot j}}{n_{\cdot\cdot}}$

③

This is called <u>proportional</u> <u>data</u>

You <u>may</u> be able to run this using the balanced ANOVA module in the stat package

<u>Exact Analysis</u>

Model: $y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{(ij)k}$

$i = 1, 2, 3 \quad j = 1, 2, 3 \quad k = 1, \ldots, n_{ij}$

$\sum_i \tau_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_i \tau\beta_{ij} = \sum_j \tau\beta_{ij} = 0$

④

$$\underset{Y}{\begin{bmatrix} 130 \\ 155 \\ 74 \\ 180 \\ 34 \\ 40 \\ 80 \\ 75 \\ 70 \\ 58 \\ 159 \\ 126 \\ 136 \\ 115 \\ 45 \\ 138 \\ 160 \\ 150 \\ 139 \\ 96 \end{bmatrix}}_{20\times 1} = \begin{array}{c} \begin{array}{ccccccccc} x_0 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}_{20\times 9} \end{array} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \tau\beta_{11} \\ \tau\beta_{12} \\ \tau\beta_{21} \\ \tau\beta_{22} \end{bmatrix}_{9\times 1} + \vec{\varepsilon}$$

⑤

Note:
$$\tau\beta_{13} = -\tau\beta_{11} - \tau\beta_{12}$$
$$\tau\beta_{23} = -\tau\beta_{21} - \tau\beta_{22}$$
$$\tau\beta_{31} = -\tau\beta_{11} - \tau\beta_{21}$$
$$\tau\beta_{32} = -\tau\beta_{12} - \tau\beta_{22}$$
$$\tau\beta_{33} = -\tau\beta_{31} - \tau\beta_{32}$$
$$= \tau\beta_{11} + \tau\beta_{21} + \tau\beta_{12} + \tau\beta_{22}$$

⑥

To get an exact analysis, run this as a regression, entering the variables $Y, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ as your data set.

Look at the sequential sums of squares.

$$SS_A = SS_{X_1} + SS_{X_2|X_1}$$

$$SS_{B|A} = SS_{X_3|X_1X_2} + SS_{X_4|X_1X_2X_3}$$

$$SS_{AB|A,B} = SS_{X_5|X_1X_2X_3X_4} + \ldots$$

Approximate methods:

Imputation: In Stat 565, we considered the case of an empty cell.
If a cell has some values, but fewer than the other cells, you can use $\bar{y}_{ij\cdot}$ as the imputed value.
You must manually subtract an error df for each imputed value.

Setting aside a data value:

Randomly set aside values from the over-populated cells.

Yates' Method: Compute the cell means + treat those as if they were the data set.

(9)

Type	Temp 15	70	125	
1	134.75	57.25	64	
2	142.5	125.5	45	
3	149	144.5	96	

Run this as a 2-way ANOVA with interaction. There will 0 df for error.

Compute $\hat{\sigma}^2 = \dfrac{\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2}{n_{\cdot\cdot} - ab}$

(10)

This estimates $\sigma^2 = V(y_{ijk})$

However, Yates' Method is an ANOVA on the cell means, i.e. the $\bar{y}_{ij\cdot}$, not the original y_{ijk}.

$V(\bar{y}_{ij\cdot}) = \dfrac{\sigma^2}{n_{ij}}$ (not the same for every cell)

Their average is $\bar{V}(\bar{y}_{ij\cdot}) = \dfrac{\sum_i \sum_j \frac{\sigma^2}{n_{ij}}}{ab}$

(11)

This can be estimated by

$$\frac{\hat{\sigma}^2}{ab} \sum_i \sum_j \frac{1}{n_{ij}}$$

Yates said to use this as the manually-computed MSE, with $n_{..} - ab$ df.

(12)

Cell-means ANOVA

Source	df
A	$a-1$
B	$b-1$
AB	$(a-1)(b-1)$
Error	0
Total	$ab-1$

Then, manually compute the MSE + reconstruct the ANOVA table

Source	df
A	$a-1$
B	$b-1$
AB	$(a-1)(b-1)$
Error	$n_{..} - ab$
Tot	$n_{..} - 1$