# Handout 5

## Unbalanced two-factor ANOVA

Consider the Hay Fever Relief data, but suppose that a few observations are missing and this an example of an unbalanced two-factor study.

| | Factor B (ingredient 2 | | |
|---|---|---|---|
| Factor A (ingredient 1) | $j=1$ | $j=2$ | $j=3$ |
| $i=1$ | $2.4, 2.7, 2.3, 2.5$ | $4.6, 4.2, 4.9$ | $4.8, 4.4$ |
| $i=2$ | $5.8, 5.2, 5.5$ | $8.9, 9.1, 9$ | $9.3, 9.4, 9.1$ |
| $i=3$ | $6.1, 5.9, 6.2$ | $9.9, 10.6, 10.1$ | $13.5, 13.3, 13.2$ |

The model is of the form

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, k = 1, \ldots, n_{ij}, j = 1, \ldots, b, i = 1, \ldots, a, \end{aligned}$$

where $\{\varepsilon_{ijk}\}$ are iid $N(0, \sigma^2)$, and the main effects and the interactions satisfy

$$\begin{aligned} \sum \alpha_i = 0,\ \sum \beta_j = 0, \\ \sum_j (\alpha\beta)_{ij} = 0 \text{ for each } i, \text{ and} \\ \sum_i (\alpha\beta)_{ij} = 0 \text{ for each } j. \end{aligned}$$

Note that in this case

$$\begin{aligned} \mu_{..} = \sum\sum \mu_{ij}/(ab),\ \mu_{i.} = \sum_j \mu_{ij}/b,\ \mu_{.j} = \sum_i \mu_{ij}/a, \\ \alpha_i = \mu_{i.} - \mu_{..},\ \beta_j = \mu_{.j} - \mu_{..}, \\ (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} = \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j). \end{aligned}$$

Estimates of the parameters are

$$\begin{aligned} \hat{\mu}_{ij} = \bar{Y}_{ij.},\ \hat{\mu}_{..} = \sum\sum \hat{\mu}_{ij}/(ab), \mu_{i.} = \sum_j \hat{\mu}_{ij}/b,\ \mu_{.j} = \sum_i \hat{\mu}_{ij}/a, \\ \hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu}_{..},\ \hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu}_{..}, \\ \widehat{(\alpha\beta)_{ij}} = \hat{\mu}_{ij} - \hat{\mu}_{i.} - \hat{\mu}_{.j} + \hat{\mu}_{..} = \hat{\mu}_{ij} - (\hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j). \end{aligned}$$

Residuals, fitted values, SSE and MSE are

$$\hat{Y}_{ijk} = \hat{\mu}_{ij} = \bar{Y}_{ij\cdot},\ e_{ijk} = Y_{ijk} - \hat{Y}_{ijk},$$
$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} e_{ijk}^2,\ df = n_T - ab,$$
$$MSE = SSE/(n_T - ab),$$

where $n_T = \sum\sum n_{ij}$ is the total number of observations. The following result is easy to prove.

**Fact** $E(MSE) = \sigma^2$.

The last result shows that $MSE$ is an unbiased estimate of $\sigma^2$.

**Important Note: For the unbalanced two-factor study, it is no longer true that**

$$\hat{\mu}_{\cdot\cdot} = \bar{Y}_{i\cdot\cdot},\ \hat{\mu}_{i\cdot} = \bar{Y}_{i\cdot\cdot},\ \hat{\mu}_{\cdot j} = \bar{Y}_{\cdot j\cdot},$$
$$SSTO = SSA + SSB + SSAB + SSE.$$

# Hypothesis Testing.

**Test for Interactions:** Suppose we wish to test for no interaction, i.e., $H_0 : (\alpha\beta)_{ij} = 0$ against $H_1$ :not all $(\alpha\beta)_{ij}$ are zero, we need to use the framework of testing using the full and the reduced models. Here the full and the reduced models are

$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}\ (full)$$
$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \varepsilon_{ijk}\ (reduced).$$

Note that these models can be fitted using the regression approach given in Supplementary Note 2. Let $SSE(A, B, AB)$ (i.e., $SSE_{full}$) and $SSE(A, B)$ (i.e., $SSE_{red}$) denote the residual sums of squares for the full and the reduced models respectively. The degrees of freedom are

$$df(full) = n_T - ab,\ df(red) = n_T - (a + b - 1).$$

The test statistic is

$$F^* = \frac{[SSE_{red} - SSE_{full}]/[df(red) - df(full)]}{SSE_{full}/df(full)}$$
$$= \frac{[SSE(A, B) - SSE(A, B, AB)]/[(a-1)(b-1)]}{SSE(A, B, AB)/[n_T - ab].}.$$

We can reject $H_0$ if $F^* > F(1-\alpha; (a-1(b-1), n_T - ab)$.

For the unbalanced Hay Fever data given in this handout, $a = b = 3$ and $n_T = 29$ and we have

$$\begin{aligned} SSE_{full} &= SSE(A, B, AB) = 1.014, \\ SSE_{red} &= SSE(A, B) = 22.301, \\ F^* &= \frac{[22.301 - 1.014]/4}{1.014/18} = \frac{21.287/4}{0.056} = \frac{5.322}{0.056} = 94.452. \end{aligned}$$

The degrees of freedom are $(4, 18)$ and the p-value=0.000. So we reject the null hypothesis of no interaction

**Test for the main effects of factor A**

Suppose that we wish to test $H_0 : \alpha_i = 0$ for all $i$ against $H_1$ :not all $\alpha_i$ are zero, then we need the full model and the reduced model

$$\begin{aligned} Y_{ijk} &= \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ (full) \\ Y_{ijk} &= \mu_{..} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ (reduced). \end{aligned}$$

We can fit these models using the regression approach given in Supplementary Note 2. Let $SSE(A, B, AB)$ (i.e., $SSE_{full}$) and $SSE(B, AB)$ (i.e., $SSE_{red}$) denote the residual sums of squares for the full and the reduced models respectively. The degrees of freedom are

$$\begin{aligned} df(full) &= n_T - ab, \\ df(red) &= n_T - [1 + b - 1 + (a-1)(b-1)] \\ &= n_T - 1 - a(b-1) \end{aligned}$$

The test statistic is

$$\begin{aligned} F^* &= \frac{[SSE_{red} - SSE_{full}]/[df(red) - df(full)]}{SSE_{full}/df(full)} \\ &= \frac{[SSE(B, AB) - SSE(A, B, AB)]/(a-1)}{SSE(A, B, AB)/[n_T - ab].}. \end{aligned}$$

We can reject $H_0$ if $F^* > F(1-\alpha; a-1, n_T - ab)$.

For the data set in this handout, we have

$$\begin{aligned} SSE_{full} &= SSE(A, B, AB) = 1.014, \\ SSE_{red} &= SSE(B, AB) = 160.924, \\ F^* &= \frac{[160.924 - 1.014]/2}{1.014/18} = \frac{159.91/2}{0.056} = \frac{79.955}{0.056} = 1427.77. \end{aligned}$$

The degrees of freedom for this test are $(2, 18)$ and the p-value is 0.000. Thus we reject $H_0$.

## Sequential vs Adjusted Sums of Squares.

The ANOVA tables obtained by using "lm" or "aov" commands in R produce "sequential sum of squares" which is also known as "type I sum of squares", and this is true also for regression. The first two tables given below show the sequential sums of squares.

With the command aov(Relief~A+B+A*B, the following sequence of models are fitted

$$\begin{aligned} Y_{ijk} &= \mu_{..} + \varepsilon_{ijk}, \text{ model 1} \\ Y_{ijk} &= \mu_{..} + \alpha_i + \varepsilon_{ijk}, \text{ model 2} \\ Y_{ijk} &= \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}, \text{ model 3} \\ Y_{ijk} &= \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \text{ model 4.} \end{aligned}$$

Note that $SSTO$, $SSE(A)$, $SSE(A, B)$ and $SSE(A, B, AB)$ are the SSE's of models 1 through 4 respectively. Thus sequential SS for A is $SSTO - SSE(A)$, i.e., it is the reduction in the residual sum of squares from model (1) to model (2). Similarly, $SSE(A) - SSE(A, B)$ is the sequential SS for B and it is the reduction in SSE from model (2) to model (3). For the balanced case, it can be easily shown that $SSTO - SSE(A) = SSA$ and $SSE(A) - SSE(A, B) = SSB$, where $SSA$ and $SSB$ are as defined in Handout 1. However, these equalities do not hold for the unbalanced cases.

When testing $H_0 : \alpha_i = 0$ for all $i$ against $H_1$ :not all $\alpha_i$ are zero, we need to test if the main effects of A can be dropped from model 4. But $SSTO - SSE(A)$ cannot test for this in the the unbalanced case - it tests if $\{\alpha_i\}$ can be dropped from model 2. However, in the balanced case $STO - SSE(A)$ equals $SSE(B, AB) - SSE(A, B, AB)$.

ANOVA Table [Sequential SS (Type I SS)] [R command: aov(Relief~A+B+A*B)]

| Source | df | SeqSS | MS | F | p-val |
|---|---|---|---|---|---|
| A | $a-1=2$ | $182.389 = SSTO - SSE(A)$ | 91.194 | 1618.570 | 0.000 |
| B | $b-1=2$ | $94.777 = SSE(A) - SSE(A,B)$ | 47.388 | 841.077 | 0.000 |
| A*B | $(a-1)(b-1)=4$ | $21.287 = SSE(A,B) - SSE(A,B,AB)$ | 5.322 | 94.452 | 0.000 |
| Error | $n_T - ab = 18$ | $1.014 = SSE(A,B,AB)$ | 0.056 | | |
| Total | $n_T - 1 = 26$ | 299.467 | | | |

For the second table, the R command aov(Relief~B+A+A*B) assumes that you start with the model $Y_{ijk} = \mu_{..} + \varepsilon_{ijk}$ and then you add $\{\beta_j\}$, then $\{\alpha_i\}$ and finally add $\{(\alpha\beta_j)\}$.

ANOVA Table[Sequential SS (Type I SS)] [R command: aov(Relief~B+A+A*B)]

| Source | df | SeqSS | MS | F | p-val |
|---|---|---|---|---|---|
| B | $b-1=2$ | $126.732 = SSTO - SSE(B)$ | 63.366 | 1124.656 | 0.000 |
| A | $a-1=2$ | $150.434 = SSE(B) - SSE(A,B)$ | 75.217 | 1334.990 | 0.000 |
| A*B | $(a-1)(b-1)=4$ | $21.287 = SSE(A,B) - SSE(A,B,AB)$ | 5.322 | 94.452 | 0.000 |
| Error | $n_T - ab = 18$ | $1.014 = SSE(A,B,AB)$ | 0.056 | | |
| Total | $n_T - 1 = 26$ | 299.467 | | | |

The final table here is the table of adjusted sums of squares. This table gives you the sums of squares, mean squares and the F-values that are needed for testing for $\{\alpha_i\}$, $\{\beta_j\}$ and $\{(\alpha\beta)_{ij}\}$ in the full model $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$. Note that for the balanced case, adjusted sums of squares and sequential sums of squares are the same.

ANOVA Table[Adjusted SS (Type III SS)]

| Source | df | AdjSS | MS | F | p-val |
|---|---|---|---|---|---|
| A | $a-1=2$ | $159.910 = SSE(B,AB) - SSE(A,B,AB)$ | 79.995 | 1428.482 | 0.000 |
| B | $b-1=2$ | $93.342 = SSE(A,AB) - SSE(A,B,AB)$ | 46.671 | 833.411 | 0.000 |
| A*B | $(a-1)(b-1)=4$ | $21.287 = SSE(A,B) - SSE(A,B,AB)$ | 5.322 | 94.452 | 0.000 |
| Error | $n_T - ab = 18$ | $1.014 = SSE(A,B,AB)$ | 0.056 | | |
| Total | $n_T - 1 = 26$ | 299.467 | | | |

# Inference

In order to construct confidence intervals, the following formulas are useful

| Parameter | Estimate | Var(estimate) |
|---|---|---|
| $\mu_{ij}$ | $\hat{\mu}_{ij} = \bar{Y}_{ij.}$ | $\sigma^2(\hat{\mu}_{ij}) = \sigma^2/n_{ij}$ |
| $\mu_{i.}$ | $\hat{\mu}_{i.} = \sum_j \hat{\mu}_{ij}/b$ | $\sigma^2(\hat{\mu}_{i.}) = (\sigma^2/a^2)\sum_j n_{ij}^{-1}$ |
| $\mu_{.j}$ | $\hat{\mu}_{.j} = \sum_i \hat{\mu}_{ij}/a$ | $\sigma^2(\hat{\mu}_{.j}) = (\sigma^2/b^2)\sum_i n_{ij}^{-1}$ |
| $L = \sum c_i\mu_{i.}$ | $\hat{L} = \sum c_i\hat{\mu}_{i.}$ | $\sigma^2(\hat{L}) = (\sigma^2/a^2)\sum_i c_i^2 \sum_j n_{ij}^{-1}$ |
| $L = \sum c_j\mu_{.j}$ | $\hat{L} = \sum c_j\hat{\mu}_{.j}$ | $\sigma^2(\hat{L}) = (\sigma^2/b^2)\sum_j c_i^2 \sum_i n_{ij}^{-1}$ |
| $L = \sum\sum c_{ij}\mu_{ij}$ | $\hat{L} = \sum\sum c_{ij}\hat{\mu}_{ij}$ | $\sigma^2(\hat{L}) = \sigma^2 \sum_i \sum_j c_{ij}^2 n_{ij}^{-1}$ |

It should be noted in passing that pairwise differences are special cases of the linear combinations. For instance $\mu_{i.} - \mu_{i'.} = c_1\mu_{1.} + \cdots + c_a\mu_{a.} = L$ in which $c_i = 1, c_{i'} = -1$ and all other $c's$ are zeros. Thus

$$\hat{L} = \hat{\mu}_{i.} - \hat{\mu}_{i'.},\ \sigma^2(\hat{L}) = (\sigma^2/a^2)\left[\sum_j n_{ij}^{-1} + \sum_j n_{i'j}^{-1}\right], \text{ and}$$

$$s^2(\hat{L}) = (MSE/a^2)\left[\sum_j n_{ij}^{-1} + \sum_j n_{i'j}^{-1}\right]$$

One can thus carry out simultaneous inferences using Bonferroni, Scheffe or Tukey method. However, one needs to be careful about using the Tukey method since it requires a balanced design. A little departure from being balanced is perhaps alright, but application of this method is not advisable if there is a substantial departure from balancedness.

**For the Hay Fever Data given in this handout**, we want to find which combinations of the ingredients lead to most relief. The table of estimated means are given here.

| | Factor B (ingredient 2 | | |
|---|---|---|---|
| Factor A (ingredient 1) | $j = 1$ | $j = 2$ | $j = 3$ |
| $i = 1$ | 2.475 | 4.567 | 4.6 |
| $i = 2$ | 5.5 | 9 | 9.267 |
| $i = 3$ | 6.067 | 10.2 | 13.333 |

We need to do all pairwise comparisons since $s^2(\hat{\mu}_{ij} - \hat{\mu}_{i'j'}) = MSE(1/n_{ij} + 1/n_{i'j'})$ are not the same for any $9i, j) \neq (i', j')$. However, since $n_{ij} \geq 2$ for all $i$ and $j$, we conclude that

$$s^2(\hat{\mu}_{ij} - \hat{\mu}_{i'j'}) = MSE(1/n_{ij} + 1/n_{i'j'}) \leq MSE = 0.056, \text{ and}$$
$$s(\hat{\mu}_{ij} - \hat{\mu}_{i'j'}) \leq 0.237,$$

for any $i, j$ and $i', j'$. Let $\alpha = 0.01$. Then the Scheffe multiplier is

$$S = \sqrt{(9-1)F(1-0.05, 9, 18)} = \sqrt{(8)(2.4563)} = 4.433.$$

[The Scheffe multiplier for all **contrasts** in $\{\mu_{ij}\}$ is $S = \sqrt{(ab-1)F(1-\alpha, ab-1, n_T - ab)}$].

This tells us, for any pairwise difference $\mu_{ij}$ and $\mu_{i'j'}$, is at

$$Ss(\hat{\mu}_{ij} - \hat{\mu}_{i'j'}) \leq (4.433)(0.237) = 1.051.$$

As in the balanced case, let us compare $\mu_{33}$ and $\mu_{32}$ since $\hat{\mu}_{33}$ and $\hat{\mu}_{32}$ are the two largest sample means. Since $\hat{\mu}_{33} - \hat{\mu}_{32} = 13.333 - 10.2 = 3.133$, we can conclude with at least 99% overall confidence that $\mu_{33}$ is the largest of $\{\mu_{ij}\}$.

# Missing observations in two-factor model with observation per cell.

Let us consider the data set in Handout 4. If some observations are missing, then we would like to compare the treatments.

| Block | Training method ($j$) | | | | | Block | Training method ($j$) | | |
|---|---|---|---|---|---|---|---|---|---|
| $i$ | $j=1$ | $j=2$ | $j=3$ | | | $i$ | $j=1$ | $j=2$ | $j=3$ |
| 1 | 73 | 81 | 92 | | | 6 | 73 | | 86 |
| 2 | 76 | | 89 | | | 7 | 68 | 72 | |
| 3 | 75 | 76 | 87 | | | 8 | 64 | | 82 |
| 4 | 74 | 77 | 90 | | | 9 | | 73 | 81 |
| 5 | | 71 | 88 | | | 10 | 62 | 69 | 78 |

The models is $Y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij}$, with $\sum \rho_i = 0$ and $\sum \tau_j = 0$. Here the number of blocks $n_b = 10$ and the number of treatments $r = 3$. However, not all $n_b r = 30$ observations are available, only 24 are available. Such a study is called incomplete (block) design. However, every block (row) has at least one observation and every column (treatment) has at least one observation. Thus it is possible to estimate all the parameters. The following output is obtained by via R

Analysis of Variance Table

Response: Proficiency

| Source | df | (Seq)SS | MS | F | p-val |
|---|---|---|---|---|---|
| Block | 9 | 498.12 | 55.35 | 12.413 | 0.000 |
| Training | 2 | 1005.99 | 503.00 | 112.807 | 0.000 |
| Residuals | 12 | 53.51 | 4.46 | | |
| Total | 23 | | | | |

How does R calculate the parameters, obtain the ANOVA table etc.? We may now want to use a regression framework when the observations are missing since we will not find explicit estimates of the parameters when some of the observations are missing. For notational convenience, let us call block and treatment as factors A and B respectively. Once again, we can create factor $a-1$ factor A variables and $b-1$ factor B (training method) variables as follows

$$X_{ij,1}^{(A)} = \begin{cases} 1 & i = 1 \\ -1 & i = n_b \\ 0 & \text{otherwise} \end{cases}, \cdots, X_{ij,a-1}^{(A)} = \begin{cases} 1 & i = a-1 \\ -1 & i = a \\ 0 & \text{otherwise} \end{cases},$$

$$X_{ij,1}^{(B)} = \begin{cases} 1 & j = 1 \\ -1 & j = b \\ 0 & \text{otherwise} \end{cases}, \cdots, X_{ij,b-1}^{(B)} = \begin{cases} 1 & j = b-1 \\ -1 & j = b \\ 0 & \text{otherwise} \end{cases}.$$

Hence the model is

$$Y_{ij} = \mu_{..} + \rho_1 X_{ij,1}^{(A)} + \cdots + \rho_{a-1} X_{ij,a-1}^{(A)} + \tau_1 X_{ij,1}^{(B)} + \cdots + \tau_{b-1} X_{ij,b-1}^{(B)} + \varepsilon_{ij}.$$

We may nor rewrite this as a Gauss-Markov model $Y = X\beta + \varepsilon$, where $Y$ is the vector of all the available observations, the matrix $X$ has $a+b-1$ columns, $\beta$ consists of $\mu_{..}, \rho_i$'s and $\tau_j$'s and $\varepsilon$ is the vector of random errors. Thus $\hat{\beta} = (X^T X)^{-1} X^T Y$ and we can carry out inferences for any linear function $f^T \beta$.