Handout 2

Pooling of sums of squares:

When the interactions effects can be ignored either because of prior knowledge or because of an F-test for interactions, an alternative estimator of σ^2 can be provided. Without the interactions the two factor ANOVA model becomes: $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$.

For this model (additive model), the estimates of $\mu_{..}$, α_i and β_j are the same as before, i.e.,

$$\hat{\mu}_{..} = \bar{Y}_{...}, \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}.$$

Hence the fitted values and the residuals are:

$$\hat{Y}_{ijk} = \hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j, \text{ and}$$
$$e'_{ijk} = Y_{ijk} - \hat{Y}_{ijk} = Y_{ijk} - (\hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j).$$

Hence the new residual sum of squares is

$$SSE_{new} = \sum\sum\sum e'^2_{ijk} = \sum\sum\sum (Y_{ijk} - (\hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j))^2.$$

It can be shown that

$$SSE_{new} = SSTO - SSA - SSB = SSAB + SSE,$$
$$df(SSE_{new}) = df(SSAB) + df(SSE) = (a-1)(b-1) + (n-1)ab = nab - a - b + 1.$$

So

$$MSE_{new} = SSE_{new}/df(SSE_{new}) = SSE_{new}/(nab - a - b + 1).$$

Fact: MSE_{new} is an estimate of σ^2.

Strategy for analysis of factor effects

First let us note a few important identities

$$\mu_{i.} = \mu_{..} + \alpha_i \text{ and } \mu_{.j} = \mu_{..} + \beta_j, \text{ and hence}$$
$$\alpha_i - \alpha_{i'} = \mu_{i.} - \mu_{i'.}, i \neq i' \text{ and } \beta_j - \beta_{j'} = \mu_{.j} - \mu_{.j'}, j \neq j'.$$

Similarly a contrast in α_i's is a contrast in $\mu_{i.}$'s, and a contrast in β_j's is a contrast in $\mu_{.j}$'s

(1) First test to investigate if the interaction effects are present.

(2) If the interactions are not present, then we are really interested in comparing α_i's and β_j's.

(3) If the interactions effects are not negligible, then the interest often is in comparing μ_{ij}'s.

[Note that if interaction effects are present, one may try to investigate if transformation are effective in obtaining an additive model (i.e., a model without interaction).]

Case I. Interaction not present

If the interactions are not present, then the model is

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}, \text{with} \sum \alpha_i = 0, \sum \beta_j = 0,$$

and ε_{ijk}'s independent $N(0, \sigma^2)$. In this case estimate σ^2 can be estimated by either

$$MSE = \frac{\sum\sum\sum (Y_{ijk} - \bar{Y}_{ij.})^2}{(n-1)ab} \text{ with } df = (n-1)ab, \text{ or by}$$

$$MSE_{new} = \frac{\sum\sum\sum (Y_{ijk} - (\hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j))^2}{nab - a - b + 1} \text{ with } df = nab - a - b + 1.$$

For notational convenience we will continue to call both MSE and MSE_{new} by MSE and denote its degrees of freedom by m with the understanding that $m = (n-1)ab$ if there is no pooling of sums of squares and $m = nab - a - b + 1$ if there is a pooling.

The following table lists the estimates of various parameters and their properties

parameter	estimate	$\sigma^2(estimate)$	$s^2(estimate)$
$\mu_{i.}$	$\hat{\mu}_{i.} = \bar{Y}_{i..}$	$\sigma^2(\hat{\mu}_{i.}) = \frac{\sigma^2}{nb}$	$s^2(\hat{\mu}_{i.}) = \frac{MSE}{nb}$
α_i	$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$	$\sigma^2(\hat{\alpha}_i) = \frac{a-1}{nab}\sigma^2$	$s^2(\hat{\alpha}_i) = \frac{a-1}{nab}MSE$
$\mu_{.j}$	$\hat{\mu}_{.j} = \bar{Y}_{.j.}$	$\sigma^2(\hat{\mu}_{.j}) = \frac{\sigma^2}{na}$	$s^2(\hat{\mu}_{.j}) = \frac{MSE}{na}$
β_j	$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$	$\sigma^2(\hat{\beta}_j) = \frac{b-1}{nab}\sigma^2$	$s^2(\hat{\beta}_j) = \frac{b-1}{nab}MSE$
$D = \mu_{i.} - \mu_{i'.}$	$\hat{D} = \bar{Y}_{i..} - \bar{Y}_{i'..}$	$\sigma^2(\hat{D}) = \frac{2\sigma^2}{nb}$	$s^2(\hat{D}) = \frac{2MSE}{nb}$
$D = \mu_{.j} - \mu_{.j'}$	$\hat{D} = \bar{Y}_{.j.} - \bar{Y}_{.j'.}$	$\sigma^2(\hat{D}) = \frac{2\sigma^2}{na}$	$s^2(\hat{D}) = \frac{2MSE}{na}$
$L = \sum c_i \mu_{i.}$	$\hat{L} = \sum c_i \bar{Y}_{i..}$	$\sigma^2(\hat{L}) = \frac{\sum c_i^2}{nb}\sigma^2$	$s^2(\hat{L}) = \frac{\sum c_i^2}{nb}MSE$
$L = \sum d_j \mu_{.j}$	$\hat{L} = \sum d_j \bar{Y}_{.j.}$	$\sigma^2(\hat{L}) = \frac{\sum d_j^2}{na}\sigma^2$	$s^2(\hat{L}) = \frac{\sum d_j^2}{na}MSE$

We will now write down three multiple comparison methods.

Bonferroni

If we have g linear combinations $L_1, ..., L_g$ of $\mu_{i.}$'s or $\mu_{.j}$'s, then simultaneous $(1-\alpha)100\%$ confidence intervals for $L_1, ..., L_g$ are given by

$$\hat{L} \pm Bs(\hat{L}), \text{ where } B = t(1 - \frac{\alpha}{2g}; m),$$

where m is the df of the MSE.

Tukey:

In the table below it is understood that $i \neq i'$ and $j \neq j'$.

all differences in means	simultaneous $(1-\alpha)100\%$ confidence intervals
$D = \mu_{i\cdot} - \mu_{i'\cdot}$'s	$\hat{D} \pm Ts(\hat{D})$ where $T = \frac{1}{\sqrt{2}} q(1-\alpha; a, m)$
$D = \mu_{\cdot j} - \mu_{\cdot j'}$'s	$\hat{D} \pm Ts(\hat{D})$ where $T = \frac{1}{\sqrt{2}} q(1-\alpha; b, m)$
$D_1 = \mu_{i\cdot} - \mu_{i'\cdot}$'s and $D_2 = \mu_{\cdot j} - \mu_{\cdot j'}$'s	$\hat{D}_1 \pm T_1 s(\hat{D}_1)$ and $\hat{D}_2 \pm T_2 s(\hat{D}_2)$ where $T_1 = \frac{1}{\sqrt{2}} q(1-\alpha/2; a, m)$ and $T_2 = \frac{1}{\sqrt{2}} q(1-\alpha/2; b, m)$

Scheffe

all contrasts	simultaneous $(1-\alpha)100\%$ confidence intervals
$L = \sum c_i \mu_{i\cdot}$ with $\sum c_i = 0$	$\hat{L} + Ss(\hat{L})$ where $S^2 = (a-1)F(1-\alpha; a-1, m)$
$L = \sum d_j \mu_{\cdot j}$ with $\sum d_j = 0$	$\hat{L} + Ss(\hat{L})$ where $S^2 = (b-1)F(1-\alpha; b-1, m)$
$L_1 = \sum c_i \mu_{i\cdot}$ with $\sum c_i = 0$ and $L_2 = \sum d_j \mu_{\cdot j}$ with $\sum d_j = 0$	$\hat{L}_1 + Ss(\hat{L}_1)$ and $\hat{L}_2 + Ss(\hat{L}_2)$ where $S^2 = (a+b-2)F(1-\alpha; a+b-2, m)$

Case II Interaction present

When interactions are present, we can construct confidence intervals for α_i's, β_j's, their differences, their contrasts etc., but they may not be always meaningful. In this case one should perhaps focus on the means μ_{ij}'s, their differences or their contrasts. It should be noted that an estimate of σ^2 here is given by

$$MSE = \frac{\sum\sum\sum (Y_{ijk} - \bar{Y}_{ij\cdot})^2}{(n-1)ab} \text{ with } df = (n-1)ab.$$

The following table lists the estimates of various parameters and their properties.

parameter	estimate	$\sigma^2(estimate)$	$s^2(estimate)$
μ_{ij}	$\hat{\mu}_{ij} = \bar{Y}_{ij\cdot}$	$\sigma^2(\hat{\mu}_{ij}) = \frac{\sigma^2}{n}$	$s^2(\hat{\mu}_{ij}) = \frac{MSE}{n}$
$D = \mu_{ij} - \mu_{i'j'}, (i,j) \neq (i',j')$	$\hat{D} = \bar{Y}_{ij\cdot} - \bar{Y}_{i'j'\cdot}$	$\sigma^2(\hat{D}) = \frac{2\sigma^2}{n}$	$s^2(\hat{D}) = \frac{2MSE}{n}$
$L = \sum\sum c_{ij}\mu_{ij}$	$\hat{L} = \sum\sum c_{ij}\bar{Y}_{ij\cdot}$	$\sigma^2(\hat{L}) = \frac{\sum\sum c_{ij}^2}{n}\sigma^2$	$s^2(\hat{L}) = \frac{\sum\sum c_{ij}^2}{n} MSE$

Bonferroni

Simultaneous $(1-\alpha)100\%$ confidence intervals for g linear combinations $L_1, ..., L_g$ of μ_{ij}'s are given by

$$\hat{L} \pm Bs(\hat{L}) \text{ where } B = t(1 - \frac{\alpha}{2g}; (n-1)ab).$$

Tukey:

Simultaneous $(1-\alpha)100\%$ confidence intervals for all pairwise difference of means $D = \mu_{ij} - \mu_{i'j'}, (i,j) \neq (i',j')$, are given by

$$\hat{D} \pm Ts(\hat{D}), \text{ where } T = \frac{1}{\sqrt{2}} q(1-\alpha; ab, (n-1)ab).$$

Scheffe

Simultaneous $(1-\alpha)100\%$ confidence intervals for all contrasts $L = \sum\sum c_{ij}\mu_{ij}$ (i.e., $\sum\sum c_{ij} = 0$) are given by

$$\hat{L} \pm Ss(\hat{L}), \text{ where } S^2 = (ab-1)F(1-\alpha; ab-1, (n-1)ab).$$

Hay fever data

We have seen from Handout 1 that interaction effects are present for this data. Recall that we have $a = 3, b = 3, n = 4$ so that $n_T = nab = 36$. We will construct a few confidence intervals here. Suppose that we wish to construct a 95% confidence intervals for $D = \mu_{33} - \mu_{32}$. Then

$$\hat{D} = \bar{Y}_{33\cdot} - \bar{Y}_{32\cdot} = 13.250 - 10.275 = 2.975,$$
$$s^2(\hat{D}) = \frac{2}{n} MSE = \frac{2}{4}(0.0602) = 0.0301,\ s(\hat{D}) = 0.1735.$$

From the t-table, $t(0.975; (n-1)ab) = t(0.975; 27) = 2.052$. Hence a 95% confidence interval for $D = \mu_{33} - \mu_{32}$ is

$$2.975 \pm (2.052)(0.1735), \text{ i.e., } 2.975 \pm 0.356, \text{ i.e., } (2.619, 3.331).$$

Now suppose we wish to construct a confidence interval for $L = \mu_{33} - (\mu_{23} + \mu_{32})/2$. So

$$\hat{L} = \bar{Y}_{33\cdot} - (\bar{Y}_{23\cdot} + \bar{Y}_{32\cdot})/2 = 13.250 - (9.125 + 10.275)/2 = 3.550,$$
$$s^2(\hat{L}) = \frac{\sum\sum c_{ij}^2}{n} MSE = \frac{1^2 + (-1/2)^2 + (-1/2)^2}{4}(0.0602) = 0.0226,$$
$$s(\hat{L}) = 0.1502.$$

So a 95% confidence interval for $L = \mu_{33} - (\mu_{23} + \mu_{32})/2$ is

$$\hat{L} \pm t(0.975; 27) s(\hat{L}), \text{ i.e., } 3.550 \pm (2.052)(0.1502), \text{ i.e., } 3.550 \pm 0.308, \text{ i.e., } (3.242, 3.858).$$

Now suppose that we wish to know which combinations of the ingredients is most effective in providing relief. For this we will need to compare all possible difference of μ_{ij}'s and there are $\binom{9}{2} = 36$ such differences. Fortunately, we do not need not construct all possible confidence intervals. We need to compare only the two means corresponding to the two largest values of $\hat{\mu}_{ij}$'s. These are $\hat{\mu}_{33}$ and $\hat{\mu}_{32}$. So we can construct a 95% confidence interval for $D = \mu_{33} - \mu_{32}$, but we will need to use a multiplier that comes from a multiple comparison method which compares all possible difference of μ_{ij}'s. Let us use the Tukey multiplier T as it is the most efficient when comparing all possible differences of the means. From the Studentized Range distribution, we have $q(1 - \alpha; ab, (n-1)ab) = q(0.95; 9, 27) \approx 4.795$. [From the table, $q(0.95; 9, 24) = 4.81, q(0.95; 9, 30) = 4.72$. So $q(0.95; 9, 27) \approx 4.795$.]
So using the Tukey multiplier we have

$$(\hat{\mu}_{33} - \hat{\mu}_{32}) \pm T s(\hat{\mu}_{33} - \hat{\mu}_{32}), \text{ i.e., } 2.975 \pm (4.795)(0.1735), \text{ i.e., } 2.975 \pm 0.832.$$

Since this confidence interval does not include zero, we can say that the most effective compound is obtained when both the factors are level 3.

Appendix.

There are many results in the handout on construction of confidence intervals. Let us focus on a few of them since the proofs of the others are similar.

Additive Model.

We focus on estimating $L = \sum c_i \mu_{i\cdot}$.

The least squares estimate of L is $\hat{L} = \sum c_i \bar{Y}_{i\cdot\cdot}$. Noting that $E(\bar{Y}_{i\cdot\cdot}) = \mu_{i\cdot}$, we have

$$E(\hat{L}) = \sum c_i E(\bar{Y}_{i\cdot\cdot}) = \sum c_i \mu_{i\cdot} = L.$$

Since $\{\bar{Y}_{i\cdot}\}$ are independent, we have

$$\sigma^2(\hat{L}) = Var(\hat{L}) = \sum c_i^2 Var(\bar{Y}_{i\cdot}) = \sum c_i^2 \sigma^2/(nb) = \frac{\sum c_i^2}{nb}\sigma^2.$$

Since $\{Y_{ijk}\}$ are independent and normally distributed any linear combination of $\{Y_{ijk}\}$ is normally distributed. Thus we have $L \sim N(L, \sigma^2(\hat{L}))$, where $\sigma^2(\hat{L})$ is as given above.

Distribution of $\hat{L}$: Since $\hat{L}$ is a function of the fitted Y values, $\hat{L}$ is independent of the residuals $\{e'_{ijk}\}$ and thus of MSE_{new} (defined at the beginning of this handout). Let us denote MSE_{new} and SSE_{new} by MSE and SSE for notational simplicity. Following the arguments used in the proof of Fact 2 in Handout 1, we conclude that $R_0^2 = SSE_{new}/\sigma^2 \sim \chi^2_{nab-a-b+1}$. Since $Z = (\hat{L} - L)/\sigma(\hat{L}) \sim N(0,1)$ and Z and R_0^2 are independent, the variable

$$\frac{\hat{L} - L}{s(\hat{L})} = \frac{(\hat{L} - L)/\sigma(\hat{L})}{\sqrt{\frac{s^2(\hat{L})}{\sigma^2(\hat{L})}}} = \frac{Z}{\sqrt{R_0^2/(nab - a - b + 1)}},$$

has a t-distribution with df=$nab - a - b + 1$.

Model with Interactions.

We focus on estimating $L = \sum\sum c_{ij}\mu_{ij}$.

The least squares estimate of L is $\hat{L} = \sum\sum c_{ij}\bar{Y}_{ij\cdot}$. Since $E({}_{ij}\bar{Y}_{ij\cdot}) = \mu_{ij}$, we have

$$E(L) = \sum\sum c_{ij} E(\bar{Y}_{ij\cdot}) = \sum\sum c_{ij}\mu_{ij} = L.$$

Note that $\{\bar{Y}_{ij\cdot}\}$ are independent and thus we have

$$\sigma^2(L) = Var(\hat{L}) = \sum\sum c_{ij}^2 Var(\bar{Y}_{ij\cdot}) = \sum\sum c_{ij}^2 \sigma^2/n = \frac{\sum\sum c_{ij}^2}{n}\sigma^2.$$

Once again $\hat{L} \sim N(L, \sigma^2(\hat{L}))$ since $\hat{L}$ is a linear combination of independent normal random variables $\{Y_{ijk}\}$.

Distribution of $\hat{L}$. Following the arguments given above for the additive model, we can conclude that $(\hat{L} - L)/s(\hat{L}) \sim t_{nab-ab}$.