13.6 Multiple comparisons

If $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ is rejected we have found differences between expected values, but not how they differ from each other. A useful tool for this is pairwise comparisons.

LSD-method (Least significant difference)

Make confidence intervals for $\mu_i - \mu_j$.

$C_{ij} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{\frac{\alpha}{2},\nu}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

There are $\binom{k}{2}$ such intervals. If c_{ij} does not cover 0, we have proven that μ_i and μ_j differ.

Bonferroni's method

The method above has proven to be useful with the F-test, but one should be aware of the following.

Let $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$. H_1: at least two are different

Let $\bar{C}_{ij}$ be the event that C_{ij} covers 0. $P(\bar{C}_{ij}) = 1-\alpha$ if H_0 is true. Then

$P(\bar{C}_{12} \cap \bar{C}_{13} \cap \cdots \cap \bar{C}_{k-1,k}) = (1-\alpha)^{\binom{k}{2}}$ if we can assume independence between the intervals.

and $P(\text{reject } H_0) = 1 - (1-\alpha)^{\binom{k}{2}} \approx \binom{k}{2}\alpha \geq \alpha$.

Therefore the significance level in pairwise comparisons is often modified. Bonferroni's method suggests we should use significance level $\frac{\alpha}{\binom{k}{2}}$ i.e substitute $t_{\frac{\alpha}{2},\nu}$ with

$t_{\frac{\alpha}{2\binom{k}{2}},\nu}$ in the confidence intervals for comparing two and two means. This works well if k is not too large. The assumption about independence between intervals is not realistic.

Tukey's test.

Substitute $t_{\frac{\alpha}{2},\nu}$ with $\frac{q(\alpha,k,\nu)}{\sqrt{2}}$, Table. A12, p. 752 W.M.M.Y. The method is exact o: $P(\bar{C}_{12} \cap \bar{C}_{13} \cap \cdots \cap \bar{C}_{k-1,k}) = 1-\alpha$ if. $m_1 = m_2 = \cdots = m_k$.

```
> TukeyHSD(aov(y~type, data=filter)
+ )
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ type, data = filter)

$type
          diff        lwr       upr     p adj
B-A -61.200000 -148.18821  25.78821 0.2640093
C-A -99.616667 -186.60488 -12.62846 0.0192815
D-A   8.333333  -78.65488  95.32154 0.9985086
E-A 142.680000   51.44600 233.91400 0.0009770
C-B -38.416667 -125.40488  48.57154 0.6930953
D-B  69.533333  -17.45488 156.52154 0.1625944
E-B 203.880000  112.64600 295.11400 0.0000077
D-C 107.950000   20.96179 194.93821 0.0099683
E-C 242.296667  151.06266 333.53067 0.0000004
E-D 134.346667   43.11266 225.58067 0.0019024
```

Here: C and B cannot be distinguished (interval covers 0).
B, A and D cannot be distinguished for the same reason.
E can be distinguished from all the others.

The randomized complete block design

In one-way analysis of variance, heterogeneity in experimental units can be so large that it is difficult to detect differences among populations. One way to to obtain homogeneous conditions is by ~~blocking~~ blocking.

Another way to express it is as follows. When you know or suspect there are specific sources that cause undesirable change in the observed values, you may reduce or eliminate their effects by blocking. The experiments will now be randomized within blocks.

A model for the randomized block design is:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \left\{ \begin{array}{l} N(0, \sigma^2) \text{ and independent} \\ j = 1, 2, \dots, b, \quad i = 1, 2, \dots, k. \end{array} \right.$$

where b is the number of blocks and i is the number of observations within each block. (number of treatment levels).

Define $\mu_{ij} = \mu' + \alpha_i' + \beta_j'$

Then $\mu_{ij} = \mu' + \bar{\alpha}_.' + \bar{\beta}_.' + \alpha_i' - \bar{\alpha}_.' + \beta_j' - \bar{\beta}_.'$

where $\bar{\alpha}_.' = \dfrac{\sum_{i=1}^{k} \alpha_i'}{k}$ and $\bar{\beta}_.' = \dfrac{\sum_{j=1}^{b} \beta_j'}{b}$.

Define $\mu = \mu' + \bar{\alpha}_.' + \bar{\beta}_.'$, $\alpha_i = \alpha_i' - \bar{\alpha}_.'$, $\beta_j = \beta_j' - \bar{\beta}_.'$

We get $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $\Big\}$ $N(0,\sigma^2)$ and independent

$j = 1, 2, \dots, b, \quad i = 1, 2, \dots, k$

where $\sum_{i=1}^{k} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$

Partitioning of variation

$$Y_{ij} = \bar{Y}_{..} + \bar{Y}_{i.} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.} = \bar{Y}_{..} + \underbrace{\bar{Y}_{i.} - \bar{Y}_{..}}_{\hat{\alpha}_i} + \underbrace{\bar{Y}_{.j} - \bar{Y}_{..}}_{\hat{\beta}_j} + Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

where $\bar{Y}_{i.} = \frac{1}{b} \sum_{j=1}^{b} Y_{ij}$, $\bar{Y}_{.j} = \frac{1}{k} \sum_{i=1}^{k} Y_{ij}$ and $\bar{Y}_{..} = \frac{1}{kb} \sum_{i=1}^{k} \sum_{j=1}^{b} Y_{ij}$

$$= \frac{1}{k} \sum_{i=1}^{k} \bar{Y}_{i.} = \frac{1}{b} \sum_{j=1}^{b} \bar{Y}_{.j}$$

From a one-way ~~an~~ anova partitioning we have:

$$\sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

$$+ \underbrace{2 \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})}_{= R}$$

$$R = 2 \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{.j})}_{0} - 2 \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})}_{0}$$

We get

$$\sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{b} (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

or $$SS_T = SS_A + SS_B + SS_E$$