Multiple comparison of group means

$y_{ij} = \mu_i + \epsilon_{ij}$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d.

The classic ANOVA test is the test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, which is uninteresting and the hypothesis is usually not true. What an experimentor usually wants to find out is which treatments are better, so rejection of H_0 is usually not the end of the analysis. Once it is rejected, further work is needed to find out why it was rejected.

Definition. A linear parametric function $\sum_{i=1}^k a_i\mu_i = a'\mu$ with known constants $a_1, \ldots, a_k$ satisfying $\sum_{i=1}^k a_i = a'\mathbf{1} = 0$ is called a contrast (linear contrast).

Example. If $a = (1, -1, 0, \ldots, 0)'$, then $a'\mu = \mu_1 - \mu_2$.

Result. $\mu_1 = \mu_2 = \cdots = \mu_k$ if and only if $a'\mu = 0$ for all $a \in \mathcal{A} = \Big\{a = (a_1, \ldots, a_k)' : \sum_{i=1}^k a_i = 0\Big\}$.

Remark. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true iff $H_a : a'\mu = 0$ for all $a \in \mathcal{A}$, or all linear contrasts are zero.

Proof. $\mu_1 = \mu_2 = \cdots = \mu_k$ iff $\mu = \alpha\mathbf{1}$ for some α, or $\mu \in \mathcal{M}_C(\mathbf{1})$. Note, $\mathcal{A} = \mathcal{M}_C^{\perp}(\mathbf{1})$.

Thus, if H_0 fails, atleast one of the H_a must fail for $a \in \mathcal{A}$. i.e., $a'\mu \neq 0$. The experimenter may be interested in this contrast, and its inference. Consider inference of any linear parametric function, $a'\mu = \sum_{i=1}^k a_i\mu_i$. We have the model,
$y_{ij} \sim N(\mu_i, \sigma^2), j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$ independent. Then, $\bar{y}_{i.} \sim N(\mu_i, \sigma^2/n_i)$, $i = 1, 2, \ldots, k$ independent, and

$$E(\sum_{i=1}^k a_i\bar{y}_{i.}) = \sum_{i=1}^k a_i\mu_i = a'\mu, \quad Var(\sum_{i=1}^k a_i\bar{y}_{i.}) = \sigma^2\sum_{i=1}^k \frac{a_i^2}{n_i},$$

so that

$$\frac{\sum_{i=1}^k a_i\bar{y}_{i.} - \sum_{i=1}^k a_i\mu_i}{\sqrt{\sigma^2\sum_{i=1}^k \frac{a_i^2}{n_i}}} \sim N(0, 1).$$

Let $S_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$, $i = 1, 2, \ldots, k$. Then $S_i^2 \sim \sigma^2\chi^2_{n_i-1}$ independent of $\bar{y}_{i.}$, $i = 1, 2, \ldots, k$. Also, $(S_1^2, \ldots, S_k^2)$ is independent of $\bar{\mathbf{y}} = (\bar{y}_{1.}, \ldots, \bar{y}_{k.})$. Let $S_p^2 = \sum_{i=1}^k S_i^2$. Then $S_p^2 \sim \sigma^2\chi^2_{\sum_{i=1}^k n_i - k}$ independent of $\bar{\mathbf{y}}$. Note that this is just a repeat of our old result that RSS $= \sum_{i=1}^k\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = S_p^2$ is independent of $\hat{\beta} = \hat{\mu}$. Thus, as discussed previously,

$$\frac{a'\bar{\mathbf{y}} - a'\mu}{\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}} \sim t_{\sum_{i=1}^k n_i - k},$$

so that

$$a'\bar{\mathbf{y}} \pm t_{\sum_{i=1}^k n_i - k}(1-\alpha/2)\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}$$

is a $100(1-\alpha)\%$ confidence interval for $a'\mu$. Also, reject $H_{a,0} : a'\mu = 0$ in favour of $H_{a,1} : a'\mu \neq 0$ if

$$\left| \frac{a'\bar{\mathbf{y}}}{\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{a_i^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}} \right| > t_{\sum_{i=1}^k n_i - k}(1-\alpha/2).$$

What if we want investigate a set of contrasts simultaneously? From Boole's Inequality,
$P\left(\cup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty P(A_i)$, so $P\left(\cup_{i=1}^\infty A_i^c\right) \leq \sum_{i=1}^\infty P(A_i^c)$.
Since $\cup_{i=1}^\infty A_i^c = \left(\cap_{i=1}^\infty A_i\right)^c$,

$$1 - P(\cap_{i=1}^n A_i) \leq \sum_{i=1}^n (1 - P(A_i)) = n - \sum_{i=1}^n P(A_i), \text{ or}$$

$$P\left(\cap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

This is known as the Bonferroni Inequality. Apply this to the above problem. If we want a simultaneous confidence set for $a^{(1)'}\mu, \ldots, a^{(d)'}\mu$, consider

$$C = \left\{ a^{(j)'}\bar{\mathbf{y}} \pm t_{\sum_{i=1}^k n_i - k}(1-\frac{\alpha}{2d})\sqrt{S_p^2 \left(\sum_{i=1}^k \frac{(a_i^{(j)})^2}{n_i}\right) / (\sum_{i=1}^k n_i - k)}, j = 1, 2, \ldots, d \right\}.$$

Then

$$P(C) = P(\cap_{l=1}^d A_l) \geq \sum_{l=1}^d P(A_l) - (d-1) = \sum_{l=1}^d (1 - \frac{\alpha}{d}) - (d-1) = d - \alpha - d + 1 = 1 - \alpha.$$

This procedure is useful when d is not too large.