Unequally-replicated case

Sometimes, by design or accident (e.g. missing data), our data has an unequal degree of replication:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, \ldots, r; j = 1, 2, \ldots, c; k = 1, 2, \ldots, \ell_{ij}$, where the ℓ_{ij} are not all equal. The total number of observations is $n = \sum_i \sum_j \ell_{ij}$.

With appropriate changes (the ℓ_{ij} go inside the summations in the definitions of the sums of squares, and the residual degrees of freedom becomes $\sum_i \sum_j (\ell_{ij} - 1)$), the ideas in slides 56 to 58 mostly still apply, except that we lose orthogonality, e.g.

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \ell_{ij} (\overline{Y}_{i..} - \overline{Y}_{...})(\overline{Y}_{.j.} - \overline{Y}_{...})$$

is not in general zero. The ANOVA then becomes more complicated, and must be interpreted sequentially, as in the case of general regression.

Blocks and treatments

Suppose we want to compare a set of *treatments*, for example, drug therapies, fertilisers or varieties in agriculture, industrial process settings, or teaching methods, and we want to make our comparisons on a broad range of experimental units, for example, patients of different ages, fields in different locations, etc. Groups of units that are similar are called , and the standard arrangement is the randomised block design, where we compare say different treatments, applying each to different blocks of r units, the units being randomly assigned to the treatments in each block in order to protect against any accidental bias. Sometimes there are replicates, e.g. there might be ℓr units in each block, with assigned to each treatment.

In this situation, a analysis of variance is appropriate, even though we are typically only interested in differences between treatments, not between blocks.

But you might think that a analysis of variance would be sufficient. Does it matter which you do?

Suppose we have unreplicated data and that the response to treatment in block j is Y_{ij}.

The test of the hypothesis of no treatment effect refers

$$F = \frac{SS_{\rm row}/}{/((r-1)(c-1))} \quad \text{to} \quad F_{(r-1),(r-1)(c-1)}$$

according to slide 64. If we did a one-way analysis test of the same hypothesis we would refer

$$F^{[1]} = \frac{SS_{\rm row}/(r-1)}{SS_{\rm E}^{[1]}/} \quad \text{to} \quad F_{(r-1),r(c-1)}$$

from slide 55, changing the notation there to match the present situation.

The numerators in the F ratios are the same, but the denominators are different. The residual sums of squares are $SS_{\rm E} = \sum_{i=1}^{r}\sum_{j=1}^{c}(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$ and $SS_{\rm E}^{[1]} = \sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - \overline{Y}_{i.})^2$ from slides 64 and 55 respectively, and the degrees of freedom are correspondingly .

If there are substantial differences between block means, then $SS_{\rm E}^{[1]} \gg SS_{\rm E}$. The test is then much less sensitive (less likely to reject the null hypothesis, even if it is false). The distinction is the same as that between the paired comparison -test and the two-sample t-test, which is what this all reduces to if .

Connection with t tests

A point about connections between t and F tests was made in the regression context in question 1 of sheet 4, but it is more general.

The percentage points in the tables satisfy $t_\nu(\alpha/2)^2 = F_{1,\nu}(\alpha)$ for all ν, α, or in **R**, `qt(1-p/2,nu)^2=qf(1-p,1,nu)` for all `(p,nu)`, and recalling the definitions of the t and F distributions in terms of independent normal and χ^2 random variables, it is not hard to see why.

As for the test statistics, consider the non-replicated two-way analysis, with $r = 2$. Let $d_j = Y_{1j} - Y_{2j}$. Then $SS_{\rm row} = c\sum_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$ can be written as $(c/2)\overline{d}^2$, while $SS_{\rm E} = \sum_i\sum_j(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} - \overline{Y}_{..})^2$ is the same as $(1/2)\sum_j(d_j - \overline{d})^2$. So

$$F = \frac{SS_{\rm row}/(r-1)}{SS_{\rm E}/((r-1)(c-1))} = \frac{c\overline{d}^2}{\sum_j(d_j - \overline{d})^2/(c-1)}$$

which is clearly just the square of the paired-comparison t statistic.