6. Mathematics of ANOVA

It is no accident that $SS_T = SS_{Tr} + SS_E$. The following technique is basic; learn it now.

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{n} ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}_{..}))^2$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_i - \bar{y}_{..})^2$$

$$+ 2 \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_i - \bar{y}_{..})(y_{ij} - \bar{y}_i)$$

$$= SS_E + SS_{Tr} + 0,$$

since the last term is

$$\sum_{i=1}^{a} \left\{ (\bar{y}_i - \bar{y}_{..}) \sum_{j=1}^{n} (y_{ij} - \bar{y}_i) \right\}$$

and each inner sum is 0 (why?).
A more methodical and rigorous development of ANOVA starts with a statistical model of the data, and goes on to derive procedures according to some optimality principle. In this case there are two popular models:

**Means model:** \[ y_{ij} = \mu_i + \epsilon_{ij}, \]

**Effects model:** \[ y_{ij} = \mu + \tau_i + \epsilon_{ij}. \]

In both cases we assume that the random errors \( \epsilon_{ij} \) are independently and normally distributed, with common variance \( \sigma^2 \). Thus the means model is equivalent to

\[ y_{ij} \overset{ind.}{\sim} N \left( \mu_i, \sigma^2 \right), \]

and we seek estimates of the parameters \( \mu_1, \ldots, \mu_a, \sigma^2 \). The hypothesis that all treatments are equally effective becomes ‘\( \mu_1 = \cdots = \mu_a \)’. In the effects model, we interpret the ‘overall mean’ \( \mu \) as the common effect of the treatments, i.e. as the mean response we would see if all the treatments were equally effective. The ‘treatment effects’ \( \tau_i \)
must then \( = 0 \) on average, since a non-zero average would be included in \( \mu \). Thus in the effects model,

\[
y_{ij} \sim_{\text{ind.}} N \left( \mu + \tau_i, \sigma^2 \right), \quad \text{and} \quad a \sum_{i=1}^{a} \tau_i = 0.
\]

Here we concentrate on the effects model, and derive parameter estimates. A common and in certain senses optimal method of estimation is Least Squares, by which we minimize the total squared discrepancy between the observations and their means:

\[
S(\mu, \tau) = \sum_{i=1}^{a} \sum_{j=1}^{n} \left( y_{ij} - E[y_{ij}] \right)^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} \left( y_{ij} - \mu - \tau_i \right)^2.
\]

We aim to find estimates \( \hat{\mu}, \hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_a) \), with \( \sum \hat{\tau}_i = 0 \), minimizing \( S(\mu, \tau) \). By calculus we get the minimizers

\[
\hat{\mu} = \bar{y}_{..}, \\
\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}.
\]
(Show that $\sum \hat{\tau}_i = 0$.) Here is an algebraic proof that these are the minimizers; the technique used is important. Decompose $S (\mu, \tau)$ as

$$S (\mu, \tau) = \sum_{i=1}^{a} \sum_{j=1}^{n} \left[ (y_{ij} - \bar{y}_i) - (\mu - \bar{y}_..) - (\tau_i - (\bar{y}_i - \bar{y}_..)) \right]^2$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (\mu - \bar{y}_..)^2$$

$$+ \sum_{i=1}^{a} \sum_{j=1}^{n} (\tau_i - (\bar{y}_i - \bar{y}_..))^2 + 3 \text{ cross-products}$$

$$= SS_E + N (\mu - \hat{\mu})^2 + n \sum_{i=1}^{a} (\tau_i - \hat{\tau}_i)^2 + 0,$$

since, e.g.,

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (\mu - \bar{y}_..) (\tau_i - \hat{\tau}_i) = n (\mu - \bar{y}_..) \sum_{i=1}^{a} (\tau_i - \hat{\tau}_i) = 0.$$

(Why? Show that the other two cross-products vanish.) Thus $S (\mu, \tau)$ is the sum of the non-negative terms $SS_E$, $N (\mu - \hat{\mu})^2$ and $n \sum_{i=1}^{a} (\tau_i - \hat{\tau}_i)^2$. The parameters occur only in the final two, which are clearly minimized by $\mu = \hat{\mu}$ and $\tau_i = \hat{\tau}_i$. The minimum attainable value of $S (\mu, \tau)$ is $S (\hat{\mu}, \hat{\tau}) = SS_E$. 
A basic principle of hypothesis testing. We see how much the minimum value of $S$ increases if we assume that the hypothesis is true (this is the SS ‘due to the hypothesis’) and compare this to its absolute minimum value. Formally:

$$F_0 = \frac{\left(\min_{H_0} S - \min S\right)}{\text{(change in d.f.)}} / \min S / \text{d.f.} \sim F_{\text{change in d.f.}}^{\text{d.f.}}.$$  

The $F$-distribution holds when the errors are $\sim N(0, \sigma^2)$ and when $H_0$ is true.

As an example, to test $H_0$: $\tau_1 = \cdots = \tau_a = 0$ vs. $H_1$: ‘not all = 0’ we first assume $H_0$ is true, so that

$$S = S(\mu, 0) = SS_E + N(\mu - \hat{\mu})^2 + n \sum_{i=1}^{a} \hat{\tau}_i^2,$$

$$\min_{H_0} S = SS_E + n \sum_{i=1}^{a} \hat{\tau}_i^2 = SS_E + SS_{Tr} = SS_T,$$

on $N-1$ d.f.

Then since $\min S = SS_E$ on $N-a$ d.f., we have

$$F_0 = \frac{SS_{Tr}/(a-1)}{SS_E/(N-a)} = \frac{MS_{Tr}}{MS_E} \sim F_{N-a}^{a-1}.$$
• Sometimes \( \min S \) and \( \min_{H_0} S \) are written as \( SS_{Full} \) and \( SS_{Reduced} \): ‘full’ model containing all terms, ‘reduced’ if \( H_0 \) holds.

• It is shown in the text (p. 68) that \( E [MS_E] = \sigma^2 \). Also (asst. 1) \( E [MS_{Tr}] = \sigma^2 + \frac{n \sum_{i=1}^{a} \tau_i^2}{a-1} \); thus we expect \( F_0 \) to be near 1 under \( H_0 \), and \( > 1 \) otherwise.

• **Unbalanced case.** If the treatment groups have differing sizes \( n_i \) we say the design is unbalanced. In this case we impose the condition \( \sum_{i=1}^{a} n_i \tau_i = 0 \) and find that the LSEs are the same as before, but that now \( SS_{Tr} = \sum_{i=1}^{a} n_i \hat{\tau}_i^2 \). Everything else remains the same.
7. Model checking

- **Model adequacy testing.** Are our assumptions on the random errors satisfied? To answer this we note that the behaviour of the random errors \( e_{ij} = y_{ij} - E[y_{ij}] \) should be reflected in the residuals

\[
e_{ij} = y_{ij} - \text{est. of } E[y_{ij}]' = y_{ij} - \hat{y}_{ij}.
\]

Here the ‘fitted values’ \( \hat{y}_{ij} \) are

\[
\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_. + (\bar{y}_i - \bar{y}_.) = \bar{y}_i.
\]

- Note that, since

\[
e_{ij} = y_{ij} - \bar{y}_i,
\]

we have that \( SS_E = \sum_{i,j} e_{ij}^2 \); for this reason it is often written \( SS_{Resid} \).
We look at qq-plots, and plots of residuals versus $\tilde{y}_i$, or versus time, ... looking for anything that might contradict the assumptions of normality, independence, equality of variances.

Fig. 3.2
Formal tests of equality of variances. We suppose that $\varepsilon_{ij} \sim (0, \sigma_i^2)$ and test $H_0: \sigma_1^2 = \cdots = \sigma_a^2$. A test which is optimal if the data are Normal, but can be quite misleading otherwise, is ‘Bartlett’s test’. It is based on the fact that if $T_1, \ldots, T_a$ are any positive numbers, and $w_1, \ldots, w_a$ are any weights summing to 1, then the ratio of the weighted arithmetic mean to the weighted geometric mean is always $\geq 1$:

$$\frac{\sum w_i T_i}{\prod T_i^{w_i}} \geq 1,$$

with equality iff $T_1 = \cdots = T_a$. This is applied with $T_i = S_i^2$, the sample variance from the $i^{th}$ group, and $w_i = (n_i - 1)/(N - a)$. The log of the ratio is computed; it is positive and large values support $H_1$.

```r
> bartlett.test(g$resid, content)
Bartlett test for homogeneity of variances
data: g$resid and content
Bartlett’s K-squared = 0.9331, df = 4,
p-value = 0.9198
```
A test which is more robust against non-normality is 'Levene's test'. First calculate absolute deviations from the group medians $\tilde{y}_i$:

$$d_{ij} = |y_{ij} - \tilde{y}_i|.$$ 

Equality of the variances is indicated by equality of the group means of the absolute deviations, which is tested by the usual F-test. So we compute the $d_{ij}$ (see the commands on the web site), fit a linear model (called g.d) as before, and do the ANOVA:

```r
> anova(g.d)
Analysis of Variance Table

Response: abs.deviations

            Df Sum Sq  Mean Sq  F value Pr(>F)
content.d   4  4.960  1.24000  0.3179 0.8626
Residuals 20 78.000  3.90000
```
If our assumptions seem to be violated we might:

1. try a transformation of the $y_{ij}$ (e.g. $\sqrt{y_{ij}}$, of log $y_{ij}$) hoping that the transformed values are ‘more normal’ - read §3-4.3 for details; or

2. apply a nonparametric procedure. In the latter case we work with the ranks $R_{ij}$ of the $y_{ij}$ rather than their numerical values (so this is another kind of transformation - the *rank transformation*). Nonparametric tests are more robust; if they give the same results as the standard tests, we take this as evidence that the assumptions are met.

```r
> strength
  7  12  14  19  7  7  17  18
 25  10  15  12  18  22  11  11
 18  19  19  15  9  18  19  23  11

R <- rank(strength)
 2.0  9.5 11.0 20.5  2.0  2.0 14.0 16.5
25.0  5.0 12.5  9.5 16.5 23.0  7.0  7.0
16.5 20.5 20.5 12.5  4.0 16.5 20.5 24.0  7.0
```
Then do an ANOVA on the ranks:

```r
> g.rank <- lm(R~content)
> anova(g.rank)
Analysis of Variance Table
Response: R

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>content</td>
<td>4</td>
<td>1020.7</td>
<td>255.17</td>
<td>19.310</td>
<td>1.212e-06</td>
</tr>
<tr>
<td>Residuals</td>
<td>20</td>
<td>264.3</td>
<td>13.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Note that, apart from ties, $SS_T$ is just $N-1$ times the variance of the ranks 1, ..., $N$. It is not random and can be calculated without knowing the data. The only randomness is in $SS_{Tr}$ (since $SS_E = SS_T - SS_{Tr}$) and there is a $\chi^2$-approximation to the distribution of $SS_{Tr}$. This is used in the Kruskal-Wallis test (see §3-10.1), which typically gives results very similar to the rank transformation followed by the $F$-test.

```r
> kruskal.test(strength,content)
Kruskal-Wallis rank sum test
data:  strength and content
Kruskal-Wallis chi-squared = 19.0637, df = 4, p-value = 0.0007636
```