Anova theory even simpler

The 2 and a half pages from Peter Dalgaard's "Introductory Statistics with R" even more simplified.

Ramon Diaz-Uriarte

2024-09-03

What the *MS* estimate

- MS_W: an estimate of the variance, σ^2.
  - (The variance of what? Of the error term. Recall the first page, that says the error terms are assumed to come from a normal distribution of mean 0 and variance σ^2). This is the residual variance left after we fit the model.

- MS_B:
  - If the null hypothesis is true (i.e., all group means are equal, i.e., $\mu_1 = \mu_2 = \mu_3 = \ldots$): MS_B is also an estimate of the variance, σ^2.
  - If not all means are the same (i.e., at least one true mean is different from the rest): MS_B is "an estimate of the variance, σ^2, plus something else"
    - That "something else" is the contribution that the differences between means make to the SSD_B.

What we expect the *MS* ratio to be

Thus:

- If the null hypothesis is true, both MS_B and MS_W are estimating the same thing. Therefore, the ratio $\frac{MS_B}{MS_W}$ should be about 1.
- If the null hypothesis is not true (i.e., not all means are equal), MS_B should be larger than MS_W. So the ratio $\frac{MS_B}{MS_W}$ should be larger than 1.

The F-statistic is that ratio $\frac{MS_B}{MS_W}$. We will compare it against the F distribution (the distribution of that ratio under the null).

The F distribution has, as parameters, the degrees of freedom of the MS in the numerator and the degrees of freedom of the MS in the denominator.

What do SSD and MS stand for

- The “SSD” stands for “sum of squared deviations”. It is often written (as in much computer output) as “Sum Sq” or “sums of squares” or “SS”.
- “MS”: mean squares.