7

Analysis of variance and the Kruskal–Wallis test

In this section, we consider comparisons among more than two groups parametrically, using analysis of variance, as well as nonparametrically, using the Kruskal–Wallis test. Furthermore, we look at two-way analysis of variance in the case of one observation per cell.

7.1 One-way analysis of variance

We start this section with a brief sketch of the theory underlying the one-way analysis of variance. A little bit of notation is necessary. Let x_{ij} denote observation no. j in group i, so that x_{35} is the fifth observation in group 3; $\bar{x}_i$ is the mean for group i, and $\bar{x}_.$ is the grand mean (average of all observations).

We can decompose the observations as

$$x_{ij} = \bar{x}_. + \underbrace{(\bar{x}_i - \bar{x}_.)}_{\substack{\text{deviation of}\\\text{group mean from}\\\text{grand mean}}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\substack{\text{deviation of}\\\text{observation from}\\\text{group mean}}}$$

informally corresponding to the model

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_7, © Springer Science+Business Media, LLC 2008

in which the hypothesis that all the groups are the same implies that all α_i are zero. Notice that the error terms ϵ_{ij} are assumed to be independent and have the same variance.

Now consider the sums of squares of the underbraced terms, known as *variation within groups*

$$\mathrm{SSD}_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

and *variation between groups*

$$\mathrm{SSD}_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_.)^2 = \sum_i n_i (\bar{x}_i - \bar{x}_.)^2$$

It is possible to prove that

$$\mathrm{SSD}_B + \mathrm{SSD}_W = \mathrm{SSD}_{\text{total}} = \sum_i \sum_j (x_{ij} - \bar{x}_.)^2$$

That is, the total variation is split into a term describing differences between group means and a term describing differences between individual measurements within the groups. One says that the grouping explains part of the total variation, and obviously an informative grouping will explain a large part of the variation.

However, the sums of squares can only be positive, so even a completely irrelevant grouping will always "explain" some part of the variation. The question is how small an amount of explained variation can be before it might as well be due to chance. It turns out that in the absence of any systematic differences between the groups, you should expect the sum of squares to be partitioned according to the degrees of freedom for each term, $k - 1$ for SSD_B and $N - k$ for SSD_W, where k is the number of groups and N is the total number of observations.

Accordingly, you can normalize the sums of squares by calculating *mean squares*:

$$\begin{aligned} \mathrm{MS}_W &= \mathrm{SSD}_W / (N - k) \\ \mathrm{MS}_B &= \mathrm{SSD}_B / (k - 1) \end{aligned}$$

MS_W is the pooled variance obtained by combining the individual group variances and thus an estimate of σ^2. In the absence of a true group effect, MS_B will also be an estimate of σ^2, but if there *is* a group effect, then the differences between group means and hence MS_B will tend to be larger. Thus, a test for significant differences between the group means can be performed by comparing two variance estimates. This is why the procedure is called *analysis of variance* even though the objective is to compare the group means.

A formal test needs to account for the fact that random variation will cause some difference in the mean squares. You calculate

$$F = \mathrm{MS}_B/\mathrm{MS}_W$$

so that F is ideally 1, but some variation around that value is expected. The distribution of F under the null hypothesis is an F distribution with $k-1$ and $N-k$ degrees of freedom. You reject the hypothesis of identical means if F is larger than the 95% quantile in that F distribution (if the significance level is 5%). Notice that this test is one-sided; a very small F would occur if the group means were very similar, and that will of course not signify a difference between the groups.

Simple analyses of variance can be performed in R using the function `lm`, which is also used for regression analysis. For more elaborate analyses, there are also the functions `aov` and `lme` (linear mixed effects models, from the `nlme` package). An implementation of Welch's procedure, relaxing the assumption of equal variances and generalizing the unequal-variance t test, is implemented in `oneway.test` (see Section 7.1.2).

The main example in this section is the "red cell folate" data from Altman (1991, p. 208). To use `lm`, it is necessary to have the data values in one vector and a factor variable (see Section 1.2.8) describing the division into groups. The `red.cell.folate` data set contains a data frame in the proper format.

```
> attach(red.cell.folate)
> summary(red.cell.folate)
     folate              ventilation
 Min.   :206.0   N2O+O2,24h:8
 1st Qu.:249.5   N2O+O2,op :9
 Median :274.0   O2,24h    :5
 Mean   :283.2
 3rd Qu.:305.5
 Max.   :392.0
```

Recall that `summary` applied to a data frame gives a short summary of the distribution of each of the variables contained in it. The format of the summary is different for numeric vectors and factors, so that provides a check that the variables are defined correctly.

The category names for `ventilation` mean "N_2O and O_2 for 24 hours", "N_2O and O_2 during operation", and "only O_2 for 24 hours".

In the following, the analysis of variance is demonstrated first and then a couple of useful techniques for the presentation of grouped data as tables and graphs are shown.