Multivariate Statistics in Ecology and Quantitative Genetics

2. More about ANOVA

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/StatGen.html

18. Mai 2010

Let i be the index for the row of a data table. The data are subdivided into groups and G_i is the group row i (or patient i) belongs to; e.g. G_i can be the treatment of patient i. Let Y_i be the response variable, e.g. the blood pressure of patient i. We can apply an anova to check whether Y depends on G, and the model behind it is:

$$Y_i = b_{G_i} + \varepsilon_i$$

where the ε_i are assumed to be independent and normally distributed with expectation 0, and all ε_i have the same variance σ^2. During the ANOVA we estimate the influence b_{G_i} of the group on Y_i by the group mean $\widehat{b_g}$. Thus, the residuals $r_i := Y_i - \widehat{b_{G_i}} \approx Y_i - b_{G_i} = \varepsilon_i$ should be approximately normally distributed.

More than one factor can play a role. For example we may take into account that the blood pressure Y_i of a patient may depend on the sex S_i of the patient. In this case the model behind the anova takes the form

$$Y_i = b_{G_i} + c_{S_i} + \varepsilon_i.$$

b_{G_i} depends only on the treatment group and c_{S_i} only on the sex of the female. If we also want allow in *interaction* between the treatment and the sex, we need another variable d_{G_i,S_i} that may depend on both:

$$Y_i = b_{G_i} + c_{S_i} + d_{G_i,S_i} + \varepsilon_i.$$

This makes possible, for example, that a certain treatment has a stronger effect for males than for females.

A *balanced design* means, that the sample size are the same for each combination of factors. E.g. 10 males and 10 females in each treatment group. Some ANOVA-based method will only work for balanced designs. Therefore, it is preferable to use a balanced design when planning an experiment. If the data, however, are observations from nature, the “design” is usually unbalanced and this has to be taken into account in the analysis.

One of the methods for which you need a balanced design is Tukey’s HSD (honest significat differences). From an anova it computes confidence intervals for the pairwise differences between the group means with mulptiple-testing correction (cf. R-script).

Another thing to be careful with is the interpretation of ANOVA tables. The R command anova, applied to a single model gives a so-called “Type I Anova”, where each line take only the variables in the lines above into account:

```
> anova(model4)
Analysis of Variance Table

Response: log(ccrt)
           Df  Sum Sq Mean Sq F value    Pr(>F)
line        1  1.2224 1.22238 13.1486 0.0003812 ***
day        11  2.8471 0.25883  2.7841 0.0023769 **
person      1  0.0850 0.08504  0.9147 0.3402393
[...]
```

For example, the p-value 0.0023769 tells how much better the model with line line and day can explain the data compared to a model that only takes line into account. Thus, the values assigned to variables depend on the input order.

If you use the R command drop1 with the option test="F", you get a so-called "Type II Anova", in which each line shows the influence of one variable, given the estimates of *all* other variables.

```
> drop1(model4,test="F")
[...]
       Df Sum of Sq    RSS     AIC F value    Pr(F)
<none>              15.618 -418.91
line    1   0.05860 15.677 -420.23  0.6304 0.428338
day    11   2.47080 18.089 -414.18  2.4161 0.008177 **
person  1   0.08504 15.703 -419.92  0.9147 0.340239
```

For example, the *p*-value 0.008177 says that a model that takes line, day and person into account explains the data significantly better than a model that uses only line and person.

It is often important to rescale (i.e. transform) the data. For example, if a comparison between fitted values (group means) and the residuals show that the larger values have larger standard deviations, this may mean that the random error ist rather multiplicative than additive (as it should be). In this case, a log transform may help. Other transformations are shown in the R-script. Sometimes, there is a good explantation why a certain transformation should be applied. Sometimes the Box-Cox-Transform can help, which can take various shapes, depending on a parameter to be optimized.

Nested ANOVA: What if the data are not really independent?