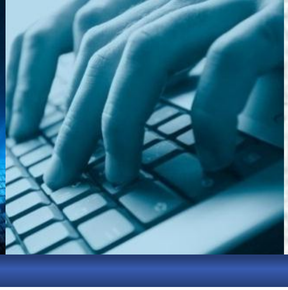


Chapter 5: Analysis of Variance

An Online Course

Sponsored by The Georgia R School

Presented by Geoffrey S. Hubona

Weight Gain

- Experiment to study the ***gain in weight of rats*** fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal).
- Ten rats are randomized to each of the four treatments and weight gain in grams is recorded.
- The question is: ***how diet affects weight gain.***
  - Is a ***balanced factorial design***: How do **four treatments** affect weight gain?
    - ‘Balanced’ means same number of observations in each cell.

Weight Gain

- 40 Observations, 3 Data Variables:
  - **Source of protein:** Beef or cereal (as factors);
  - **Amount of protein:** Low and high (as factors);
  - **Weight gain:** in grams.

Foster Feeding in Rats

- Foster feeding experiment with rat mothers and litters of four different genotypes: A, B, I and J.
- Measurement of interest is the litter weight (in grams) after a trial feeding period.
- Investigator’s interest: ***uncovering the effects of genotype of mother and genotype of litter on litter weight.***
  - Is an ***unbalanced factorial design***: are a different number of observations in the 16 cells of the design.

Foster Feeding in Rats

- 61 Observations, 3 Data Variables:
  - **litgen:** A, B, I or J;
  - **motgen:** A, B, I or J;
  - **weight:** in grams.

Water Hardness and Mortality

- Extend the previous analysis of ***water hardness*** and ***mortality*** for 61 large towns in England and Wales:
  - Assess differences of both hardness and mortality in the North or South.
  - Hypothesis is that ***the two-dimensional mean-vector of water hardness and mortality is the same for cities in the North and the South*** using MANOVA and the *Hotelling-Lawley* test.

Male Egyptian Skulls

- Four different skull measurements made on Egyptian skulls from five different epochs.
- Question is: ***are there any differences between the skulls from the five different epochs?***
    - Non-constant measurements of the skulls over time would indicate interbreeding with immigrant populations.

Male Egyptian Skulls

- 150 Observations, 5 Data Variables:
  - **epoch:** millenium BC (i.e. c4000BC);
  - **mb:** maximum breadth of the skull;
  - **bh:** basibregmatic heights of the skull;
  - **bl:** basialiveolar length of the skull;
  - **nh:** nasal heights of the skull.

Statistical Tests

- For each of these data sets, question of interest is whether certain populations differ in mean value for, in the first two domains (weight gain; foster feeding in rats), a single variable, or with the third and fourth domains (water hardness, male Egyptian skulls), for a set of variables.
    - First two domains use ***analysis of variance*** (ANOVA).
    - Third and fourth domains use ***multivariate analysis of variance*** (MANOVA).

ANOVA

The model used in the analysis of each ANOVA is:

$$y_{ijk} = \mu + \gamma_i + \beta_j + (\gamma\beta)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} represents the kth measurement made in cell (i , j) of the factorial design, μ is the overall mean, γ_i is the main effect of the first factor, β_j is the main effect of the second factor, $(\gamma\beta)_{ij}$ is the interaction effect of the two factors and ε_{ijk} is the residual or error term assumed to have a normal distribution with mean zero and variance σ^2.

Formula Specification in R

In R, the model is specified by a model formula. The two-way layout with interactions specified on the previous slide reads

$$y \sim a + b + a{:}b$$

where the variable a is the first and the variable b is the second factor. The interaction term $(\gamma\beta)_{ij}$ is denoted by a:b.

MANOVA

The linear model used is:

$$y_{ijh} = \mu_h + \gamma_{jh} + \varepsilon_{ijh}$$

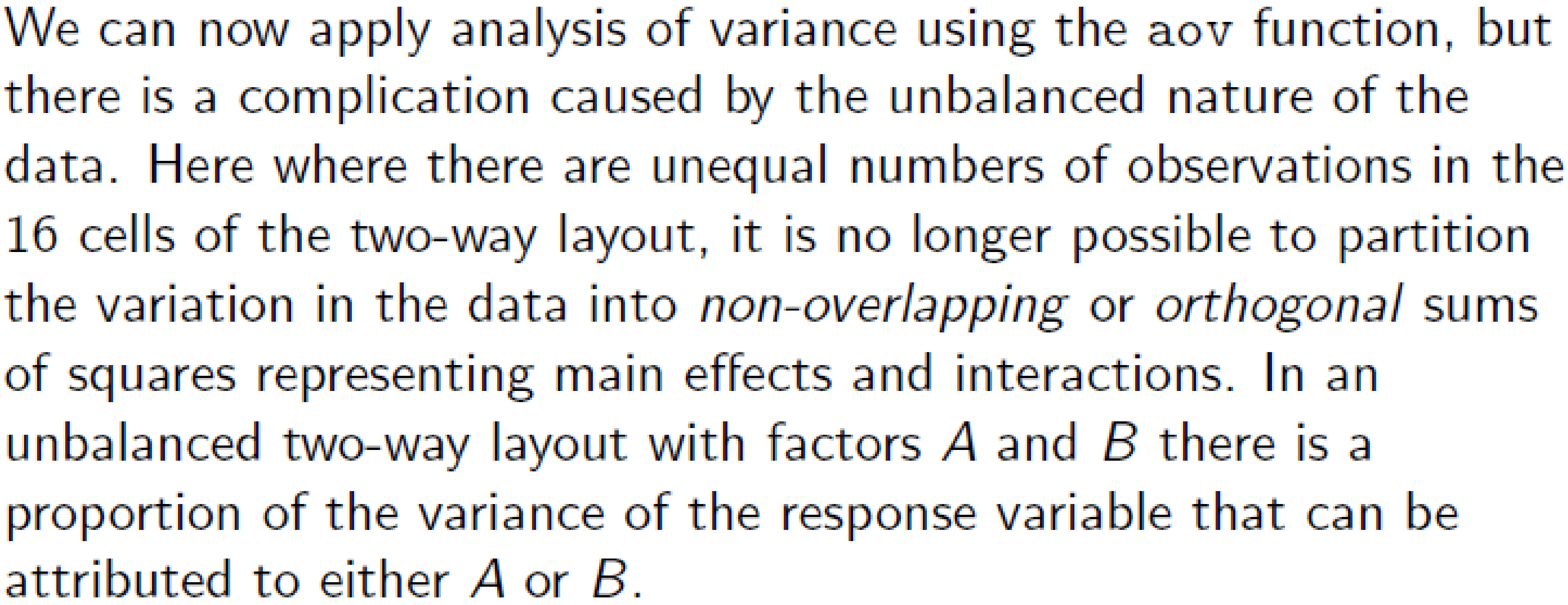
where μ_h is the overall mean for variable h, γ_{jh} is the effect of the jth level of the single factor on the hth variable, and ε_{ijh} is a random error term. The vector $\varepsilon^T = (\varepsilon_{ij1}, \varepsilon_{ij2}, \ldots, \varepsilon_{ijq})$ where q is the number of response variables (four in the skull example) is assumed to have a multivariate normal distribution with null mean vector and covariance matrix, Σ, assumed to be the same in each level of the grouping factor. The hypothesis of interest is that the population mean vectors for the different levels of the grouping factor are the same.

MANOVA Inference

A number of different test statistics are available which may give different results when applied to the same data set, although the final conclusion is often the same. The principal test statistics for the multivariate analysis of variance are

- Hotelling-Lawley trace,
- Wilks' ratio of determinants
- Roy's greatest root,
- Pillai trace.

Unbalanced ANOVA



We can now apply analysis of variance using the aov function, but there is a complication caused by the unbalanced nature of the data. Here where there are unequal numbers of observations in the 16 cells of the two-way layout, it is no longer possible to partition the variation in the data into *non-overlapping* or *orthogonal* sums of squares representing main effects and interactions. In an unbalanced two-way layout with factors A and B there is a proportion of the variance of the response variable that can be attributed to either A or B.

Multiple Comparisons

We can investigate the effect of genotype B on litter weight in more detail by the use of *multiple comparison procedures*.

Such procedures allow a comparison of all pairs of levels of a factor whilst maintaining the nominal significance level at its selected value and producing adjusted confidence intervals for mean differences. One such procedure is called *Tukey honest significant differences* suggested by Tukey (1953), see Hochberg and Tamhane (1987) also.