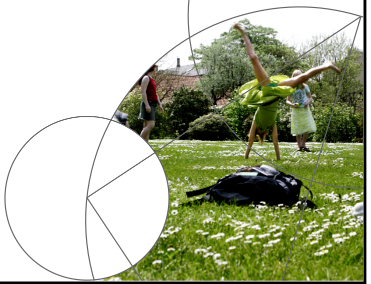
Faculty of Life Sciences

One-way analysis of variance

Ib Skovgaard & Claus Ekstrøm
E-mail: ims@life.ku.dk



Program

- One-way ANalysis Of VAriance (ANOVA)
  - Problem and type of data
  - Variation between groups and within groups
  - Residuals

Antibiotics and decomposition of organic material

Data

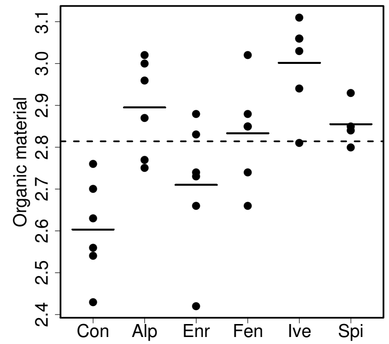
- Five types antibiotics and a control treatment.
- 36 heifers in 6 treatment groups. Feed with antibiotics added.
- Dung deposits in bags in the ground. After 8 weeks amount of organic material measured.
- For spiramycin: only four usable measurements,

Problem(s):

- Do the antibiotics affect the decomposition of organic material?
- How do the five antibiotics compare with the control?
- They seem to give higher values, but can we conclude that they counteract the decomposition?

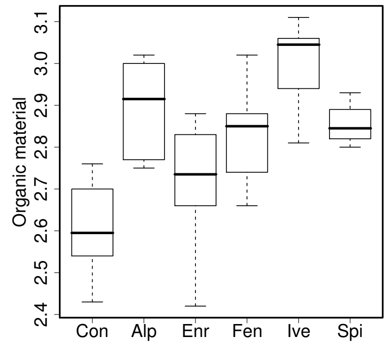
Graphs

Data



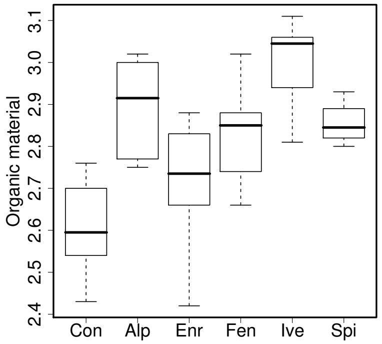
Parallel box-plots

`boxplot(org~treat)`



Group means and group-wise standard deviations

Type	n_j	$\bar{y}_j$	s_j
Control	6	2.603	0.119
α-cyperm.	6	2.895	0.117
Enrofloxacin	6	2.710	0.162
Fenbendaz.	6	2.833	0.124
Ivermectin	6	3.002	0.109
Spiramycin	4	2.855	0.054



- Do we need anything but the numbers and the graphs?
- What would you conclude?

Populations, samples and estimates

Population vs. sample

- The 34 heifers is a sample from the population of heifers.
- More precisely we imagine that we could continue sampling heifers to each of the treatment groups belonging to six (infinite) treatment populations: heifers given treatment 1, heifers given treatment 2, etc.
- Our sample is assumed to be representative for its population.
- Computations necessarily are done on the sample
- but conclusions should regard the populations to be useful.

Population and sample means

- Let α_j denote the population mean for heifers given treatment j
- The sample mean $\bar{y}_j$ is the estimate for α_j: $\hat{\alpha}_j = \bar{y}_j$
- What does it mean if there is no effect of antibiotics?

Notation

- $k =$ number of groups, here $k = 6$
- $n_j =$ number of obs. in group j, here $n_1 = \cdots = n_5 = 6$, $n_6 = 4$.
- $g(i)$ denotes the group for observation i. For example

$$g(1) = \cdots = g(6) = \text{control}, \quad g(31) = \cdots = g(34) = \text{Spiramycin}$$

or

$$g(1) = \cdots = g(6) = 1, \qquad g(31) = \cdots = g(34) = 6.$$

- Sample mean and sample standard deviation in group j:

$$\bar{y}_j = \frac{1}{n_j} \sum_{i:g(i)=j} y_i \qquad s_j = \sqrt{\frac{1}{n_j - 1} \sum_{i:g(i)=j} (y_i - \bar{y}_j)^2}$$

but really just the mean and standard deviation for group j.

Pooled standard deviation

If it is reasonable to assume similar variation in all groups, it is better to use all the groups to compute a single standard deviation reflecting the within-group variation.

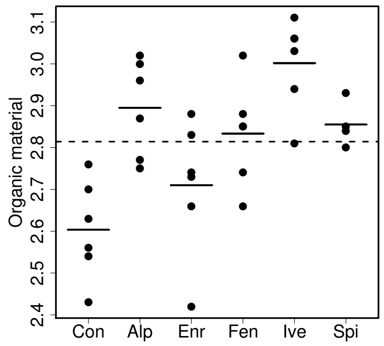
Pooled within-group sample standard deviation:

$$s = \sqrt{\frac{1}{n-k}\sum_{j=1}^{k}(n_j-1)s_j^2}$$

$$= \sqrt{\frac{1}{28}(5\cdot s_1^2 + 5\cdot s_2^2 + \cdots + 3\cdot s_6^2)} = 0.1217$$

The pooled within-group sample variance is s^2, and it is a weighted mean of the group sample variances.

Variation within and between groups



- Variation within groups — points around the lines.
  $$SS_e = \sum_{i=1}^{n}(y_i - \bar{y}_{g(i)})^2$$
- Variation between groups — Lines around the dashed line
  $$SS_{grp} = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$
- Total variation — Points around the dashed line.
  $$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Variation between groups is compared with variation within groups:

Analysis of variance (ANOVA) table

Variation	SS	df (degrees of freedom)	MS = SS/df
Between types	0.5908	$k-1=5$	0.1182
Residual	0.4150	$n-k=28$	0.0148
Total	1.0058	$n-1=33$	

The table splits the total variation into two parts, because

$$SS_{total} = SS_{grp} + SS_e$$

and

$$df_{total} = df_{grp} + df_e$$

Residuals

Recall the residuals from the linear regression: $r_i = y_i - \hat{\alpha} - \hat{\beta}\cdot x_i$.

One-way ANOVA:

- Residuals
  $$r_i = y_i - \bar{y}_{g(i)} = \text{observation} - \text{estimate}$$
- Residual sum of squares is SS_e:
  $$SS_e = \sum_{i=1}^{n}(y_i - \bar{y}_{g(i)})^2 = \sum_{i=1}^{n} r_i^2$$
- The pooled standard deviation can be obtained from the residual sum of squares:
  $$s = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n} r_i^2} = \sqrt{\frac{1}{df_e}\sum_{i=1}^{n} r_i^2}$$

  This holds for all linear models (coming later ...!)

Two unpaired or paired samples

Unpaired samples: 2 groups — one-way ANOVA.

Paired samples: ToDo!

One-way ANOVA: summary

- Observations divided into k groups, such as treatments strains, product or age groups.
- Purpose: comparison of the groups
- Partition of the total variation into variation between groups and variation within groups
- Pooled standard deviation, s
- Still need statistical assessment of some kind to conclude if the population groups are different.