Data analysis with R

Lecture 9

Statistical modelling

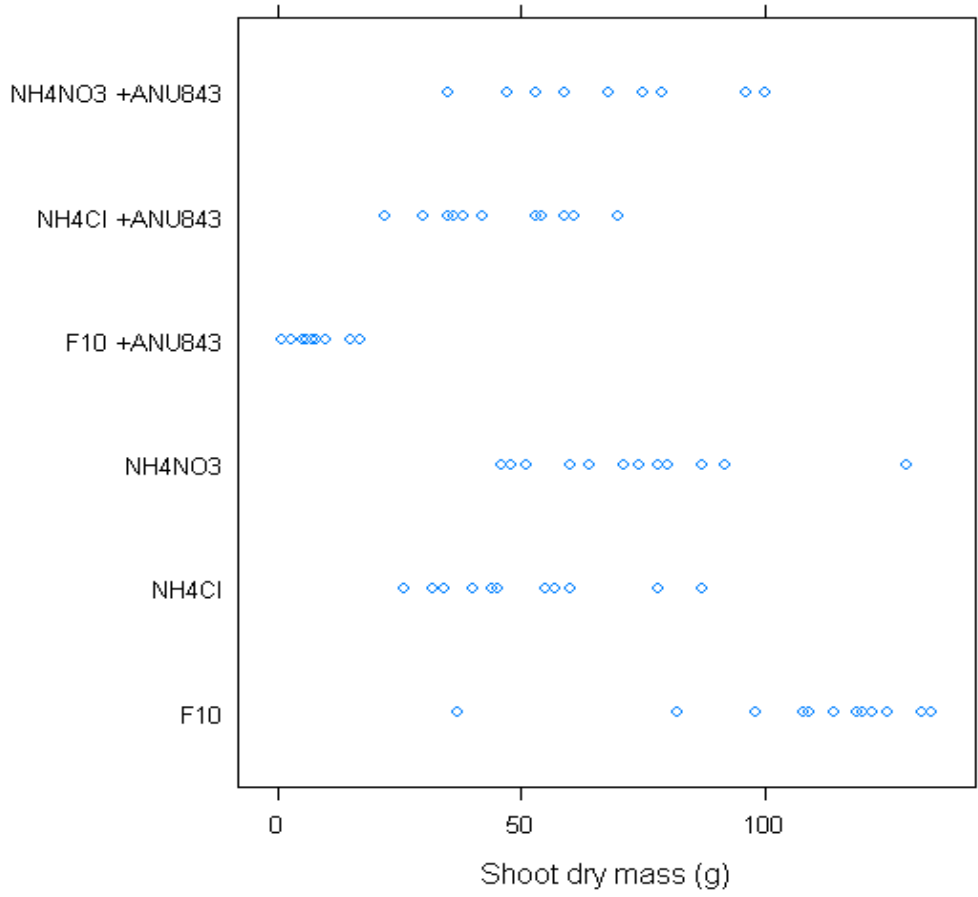
Jouni Junnila

Example data

- Let’s use an example-data to get us acquainted with statistical modelling.
- The example-data is a shoot dry mass data from an experiment that compared wild type *(wt)* and genetically modified rice plant (*ANU843*), each with three different chemical treatments. We have 72 observations in total.

First view

- `> stripplot(trt ~ShootDryMass, data=rice, xlab="Shoot dry mass (g)")`



First view (2)

- The bottom three strips are the results of ”ild type” plants, the final three strips repeat the treatments but for ANU843.
- The stripplot displays ”within group” variation, as well as gives an indication of variability between the group means.
- For now, let’s ignore the two-way structure in the data and carry-out a one-way analysis of the results.

One-way analysis of variance

- One-way analysis of variance formally tests whether the variation among the means is greater than what might occur simply because of the natural variation within each group.
- An F-statistic much larger than 1, indicates that the means are different.
  - P-value is designed to assist this judgement

One-way analysis of variance in R

- Easiest way to conduct a one-way ANOVA in R is to use *aov*-function.

```
model <- aov(ShootDryMass ~ trt, data=rice)
anova(model)
Analysis of Variance Table

Response: ShootDryMass
           Df Sum Sq Mean Sq F value    Pr(>F)
trt         5  68326 13665.1  36.719 < 2.2e-16 ***
Residuals 66  24562   372.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

One.way ANOVA; interpretation

- The very small *p*-value for the *F*-statistic strongly indicates that there are indeed differences between the treatment means.
- Interest now lays in determining the nature of the differences.
- A first-step could be to print out the coefficients of the model.

Coefficients

```
>summary.lm(model)$coef

                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)      108.33333   5.568917  19.453214 4.918624e-29
trtNH4Cl         -58.08333   7.875638  -7.375064 3.474516e-10
trtNH4NO3        -35.00000   7.875638  -4.444084 3.454197e-05
trtF10 +ANU843  -101.00000   7.875638 -12.824358 1.381337e-19
trtNH4Cl +ANU843 -61.75000   7.875638  -7.840635 5.106634e-11
trtNH4NO3 +ANU843 -36.83333  7.875638  -4.676870 1.488189e-05
```

Coefficients; interpretation

- The initial level, which in here is *F10*, has the role of a reference or baseline level.
  - The ”Intercept” line gives the estimate for *F10*.
- Other treatment estimates are differences from the estimates for *F10*.
- The standard errors are, after the first row, standard errors for differences between *F10* and later treatments.

Changing the reference-level

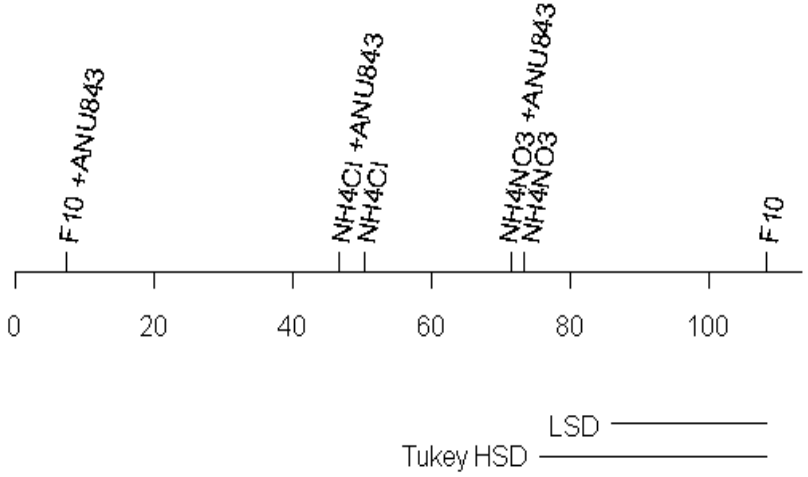
- We can easily change the reference level to some other with the *relevel*-function.
- For example, if we want the reference-level to be ”NH4Cl”, we can type:

```
› rice$trt <- relevel(rice$trt,
  ref=”NH4Cl”)
```

- And then run the analysis again.

One-way plot

- A special plot in the DAAG-library for one-way layouts is called *oneway.plot.*
- For out example we can type
  - `oneway.plot(aov(ShootDryMass~trt, data=rice=`



Interpretation of the plot

- From the plot we see, that results come in pairs. For F10 there is a huge difference between wild type and ANU843 variety, on for the two other chemicals there is no detectable difference.
  - This highlights the two-way structure we actually have in the data.
  - If we have a two-way structure, running a one-way model is undesirable. We may miss out important features.

Multiple comparisons

- When doing multiple comparisons, we have to worry about multiplicity-issue.
- Tukey's HSD-test (Honestly significant differences) does a quite strict and conservative comparison, i.e it is somewhat biased against finding differences.
- For example The Least Significant Difference (LSD) –test does the opposite, it is anti-conservative and biased towards finding differences.
- Usually prefer to be conservative than anti-conservative.

Tukey HSD-test

- Function for doing Tukey's HSD-test is *TukeyHSD*.

```
>TukeyHSD(model)
  Tukey multiple comparisons of means 95% family-wise confidence level
Fit: aov(formula = ShootDryMass ~ trt, data = rice)
$trt
```

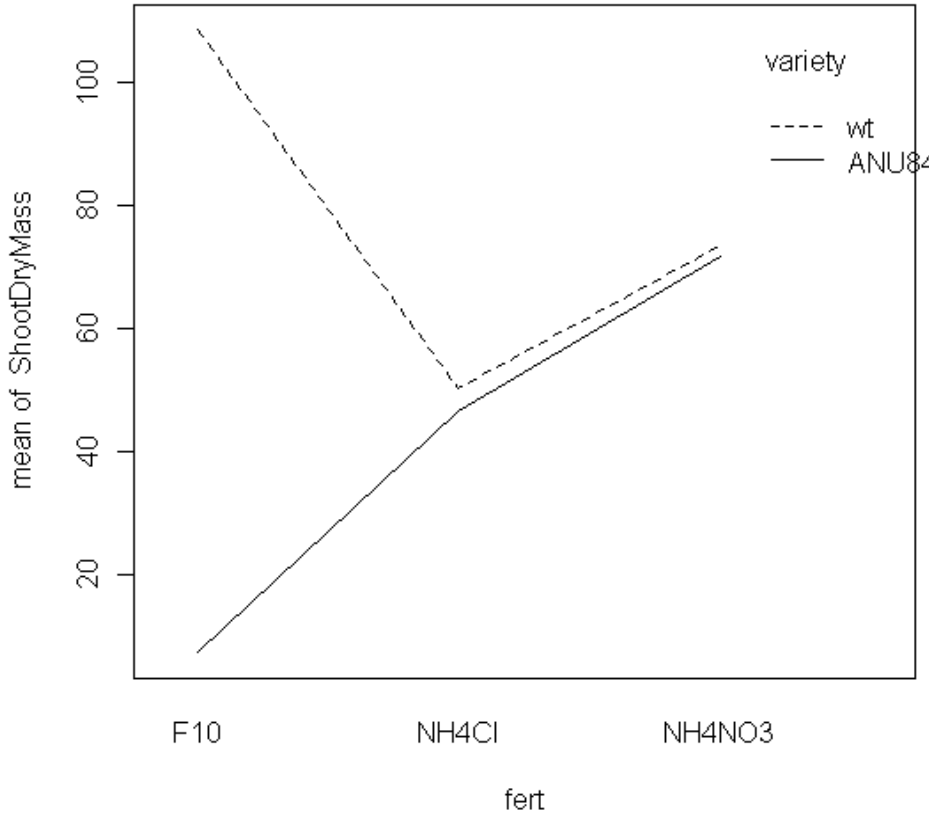
	diff	lwr	upr	p adj
NH4Cl-F10	-58.083333	-81.1990766	-34.967590	0.0000000
NH4NO3-F10	-35.000000	-58.1157432	-11.884257	0.0004789
F10 +ANU843-F10	-101.000000	-124.1157432	-77.884257	0.0000000
NH4Cl +ANU843-F10	-61.750000	-84.8657432	-38.634257	0.0000000
NH4NO3 +ANU843-F10	-36.833333	-59.9490766	-13.717590	0.0002094
NH4NO3-NH4Cl	23.083333	-0.0324099	46.199077	0.0505271
F10 +ANU843-NH4Cl	-42.916667	-66.0324099	-19.800923	0.0000117
NH4Cl +ANU843-NH4Cl	-3.666667	-26.7824099	19.449077	0.9971514
NH4NO3 +ANU843-NH4Cl	21.250000	-1.8657432	44.365743	0.0892143
F10 +ANU843-NH4NO3	-66.000000	-89.1157432	-42.884257	0.0000000
NH4Cl +ANU843-NH4NO3	-26.750000	-49.8657432	-3.634257	0.0141406
NH4NO3 +ANU843-NH4NO3	-1.833333	-24.9490766	21.282410	0.9999020
NH4Cl +ANU843-F10 +ANU843	39.250000	16.1342568	62.365743	0.0000682
NH4NO3 +ANU843-F10 +ANU843	64.166667	41.0509234	87.282410	0.0000000
NH4NO3 +ANU843-NH4Cl +ANU843	24.916667	1.8009234	48.032410	0.0273045

Data with a two-way structure

- The example data, rice-data has in fact a two-way structure.
- The first factor relates to whether F10, NH4Cl or NH4NO3 is applied.
- Second factor relates to whether the plant is wild type or ANU843.
- An interaction plot represents nicely this structure.

Interaction plot

- `attach(rice)`
- `interaction.plot (fert,variety, ShootDryMass)`



Interaction plot; interpretation

- The interaction plot shows a large difference between ANU843 and wt for the F10 treatment.
- For the other treatments there is now detectable difference.
- A two-way analysis would show us a large interaction.
- Let’s analyze the data with a two-way variance analysis model.

Two-way ANOVA in R

```
model2 <- aov(ShootDryMass ~ fert + variety + fert*variety, data=rice); anova(model2)
Analysis of Variance Table

Response: ShootDryMass
             Df Sum Sq Mean Sq F value    Pr(>F)
fert          2   7019  3509.4  9.4299 0.0002499 ***
variety       1  22684 22684.5 60.9546 5.858e-11 ***
fert:variety  2  38622 19311.2 51.8903 2.875e-14 ***
Residuals    66  24562   372.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Two-way ANOVA; coefficients

```
➢ summary.lm(model2)
aov(formula = ShootDryMass ~ fert + variety + fert * variety,
    data = rice)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                108.333      5.569  19.453  < 2e-16 ***
fertNH4Cl                  -58.083      7.876  -7.375 3.47e-10 ***
fertNH4NO3                 -35.000      7.876  -4.444 3.45e-05 ***
varietyANU843             -101.000      7.876 -12.824  < 2e-16 ***
fertNH4Cl:varietyANU843     97.333     11.138   8.739 1.27e-12 ***
fertNH4NO3:varietyANU843    99.167     11.138   8.904 6.45e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 19.29 on 66 degrees of freedom
Multiple R-squared: 0.7356,     Adjusted R-squared: 0.7155
F-statistic: 36.72 on 5 and 66 DF,  p-value: < 2.2e-16
```

Presentation issues

- So far we have treated all comparisons as of equal interest. Often they are not. There are several possibilites:
  - Interest may be in comparing treatments with a control, with comparisons between treatments of lesser interest.
  - Interest may be in comparing treatments with one another.
  - There may be several groups of treatments, with the main interst in comparing the different groups., etc.

Presentation issues (2)

- Any of the previous situations should lead to specifying in advance the comparisons of interest.
- When we present our data, we should be careful not to mislead the reader and to give them enough information to understand what has been presented.
- Next we'll see few instructions of presenting data that are useful.

Presentation issues (3)

- For graphical presentations, use a layout that reflects the data structure, i.e., a one-way layout for a one-way data, and a two-way layout for a two way data.
- Explaing clearly how error bars should be interpreted - ? SE, ? 95 % confidence interval, ? SED limits or whatever.
- When there is more than one source of variation, explain what source of ”error” is/are represented.
  - Analyst should try to find the error what is relevant and interesting to be presented in the graphs.

Nested variance structure

- Some experiments have a data structure where the variation is nested within another variable. This kind of structure requires special attention in the model formula.
- Example: Ten apples are taken from a box. 5 are assigned to one tester, 5 to another tester randomly. Both testers make two firmness tests on each of their five fruit.
- Here we have a nested structure, where the variance of the fruit is nested within the tester.

Nested variance structure (2)

- Easy mistake here would be to analyze this as a two parallel group desing, i.e. comparing ten observations against ten observations.
  - This would be wrong as we only have 5 fruits / group.
    - We would end up with too accurate error estimate, i.e. Underestimation of the variation.
- How these kind of models can be handled in R, will be handled on next lecture.