

Kurs Bio144:

Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 5: ANOVA

23./24. March 2017

Overview

- One-way ANOVA
- Post-hoc tests and contrasts
- Two-way ANOVA
- ANOVA as special cases of a linear model

Note:

ANOVA = ANalysis Of VAriance (Varianzanalyse)

Course material covered today

The lecture material of today is based on the following literature:

- Chapter 12 from Stahel book “Statistische Datenanalyse”
- “Getting Started with R” chapters 5.6 and 6.2

ANOVA and ANCOVA

ANOVA = Varianzanalyse

ANCOVA = Kovarianzanalyse

Introduction by Sir R. A. Fisher (1890-1962). He worked at the agricultural research station in Rothamstead (England). AN(C)OVA are/were therefore traditionally used to analyze agricultural experiments.

Questions of AN(C)OVA:

- Generally: Are the means of two or more groups different?
- Example: Are different plant breeds different in important aspects (e.g., yields / Ertrag)?
- Example: What is the influence of different treatments on plants (Biology) or patients (Medicine)?

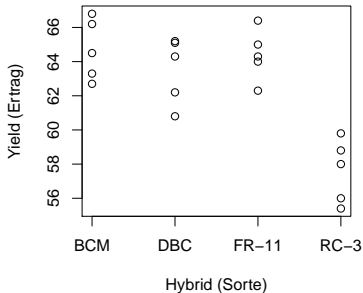
Beispiel: Ertragspotential bei Hybrid-Mais mit erhöhter Pilzbrand-Resistenz

(Source: W. Blanckenhorn, UZH)

Es wurden 4 Hybrid-Mais-Sorten angebaut und ihr Körnerertrag ermittelt. Jede Sorte wurde an 5 Orten angepflanzt.

Frage: Unterscheiden sich die Hybrid-Mais-Sorten im Ertrag?

Achtung: Die Frage bezieht sich auf *irgendeinen* Unterschied. Präziser könnte man fragen, ob sich irgendeine der Sorten von den anderen unterscheidet?



Naive idea: To carry out pairwise t -tests between any two groups.

- ① How many tests would this imply?
- ② Why is this not a very clever idea?

Naive idea: To carry out pairwise t -tests between any two groups.

- ① How many tests would this imply?
- ② Why is this not a very clever idea?

(Generally, the number of pairwise tests can be calculated by $g(g - 1)/2$, where g =number of groups.)

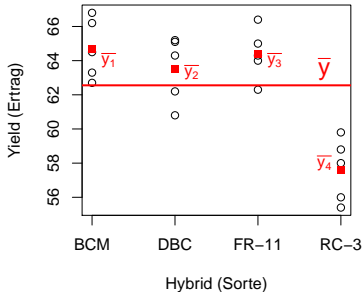
Better idea

Formulate a model that is able to **test simultaneously** whether there is an **overall difference between the groups**. That is, ask only **one question!**

This leads us to the

Idea of the ANOVA analysis:

Compare the variability within groups (MS_E) to the variability between the group means (MS_G).



We formulate a model as follows:

$$y_{ij} = \mu_i + e_{ij} ,$$

where

- y_{ij} = "Ertrag der j -ten Pflanze der Sorte i "
- μ_i = "Mittlerer Ertrag der Sorte i "
- $e_{ij} \sim N(0, \sigma_e^2)$ is an independent error term.

Typically, this is rewritten as

$$y_{ij} = \mu + \beta_i + e_{ij} ,$$

where $\mu + \beta_i = \mu_i$ from above, thus the **group mean** of group i .

Single factor ANOVA (Einfaktorielle Varianzanalyse)

More generally, this leads us to the **single factor ANOVA**:

Assume we have g groups and in each group i there are n_i measurements of some variable of interest, denoted as y . The model is then given as

$$\begin{aligned} y_{ij} &= \mu + \beta_i + e_{ij} \quad \text{for} \quad i = 1, \dots, g, \\ j &= 1, \dots, n_i, \\ e_{ij} &\sim N(0, \sigma_e^2) \quad i.i.d. \end{aligned} \tag{1}$$

- μ plays the role of the **intercept** β_0 in standard regression models.
- The estimation of $\mu, \beta_2, \dots, \beta_g$ is again done by **least squares minimization**.
- The $e_{ij} \sim N(0, \sigma_e^2)$ *i.i.d.* assumption is again crucial, so **model checking** will be needed again.

Attention: Model (1) is overparameterized, thus an additional constraint is needed! Most popular:

- $\beta_1 = 0$ (**treatment contrast**; default in R).

Interpretation: Group 1 is usually chosen such that it is some sort of **reference group** or **reference level**, for example a standard diet, while groups 2, 3, etc. corresponde to novel diets whose effect is tested in an experiment.

- $\sum_i \beta_i = 0$ (**sum-to-zero contrast**).

Interpretation: The effects β_1, β_2 etc give the deviation from the population averaged effect.

ANOVA as a special case of a linear model

The clue is: Model (1) is identical to the regression model with a factor covariate, see slides 36/37 from week 3.

Interpretation: The levels of the factor are now the different group memberships.

Thus (assuming $\beta_1 = 0$):

$$y_{ij} = \begin{cases} \mu + e_{ij}, & \text{for group 1} \\ \mu + \beta_2 + e_{ij}, & \text{for group 2} \\ \dots & \\ \mu + \beta_g + e_{ij}, & \text{for group } g, \end{cases}$$

and $\hat{y}_{ij} = \bar{y}_{.i} = \mu + \beta_i$ can be interpreted as the predicted value.

The ANOVA test: The F -test

Aim of ANOVA: to test *globally* if the groups differ. That is:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g \quad \text{or, equivalently} \quad \beta_1 = \beta_2 = \dots = \beta_g$$

$$H_1 : \text{At least two groups are different}$$

Remember from slide 44 of week 3 (or Stahel script p.33): We have already used the F -test for categorical variables (see F -test for the earthworks, slide 45 of week 3). This was equivalent to testing if all β s that belong to a categorical variable are $=0$ at the same time.

This is **the very same problem here**, thus we need the F -test again!

To derive the ingredients of the F -test, we again look at the decomposition of variance:

Variance decomposition

(Remember this idea from week 3, slide 24, and replace \hat{y}_{ij} by $\bar{y}_{.i}$)

$$\begin{aligned} SS_{total} &= SS_{\text{between groups}} + SS_{\text{within groups}} \\ \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^g n_i (\bar{y}_{.i} - \bar{y})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{.i})^2 \end{aligned}$$

Degrees of freedom:

$$n - 1 = (g - 1) + (n - g)$$

From this:

$$\left. \begin{aligned} MS_G &= \frac{SS_{\text{between}}}{g-1} \\ MS_E &= \frac{SS_{\text{within}}}{n-g} \end{aligned} \right\} \Rightarrow F = \frac{MS_G}{MS_E} \text{ is } \sim F_{g-1, n-g} \text{ distributed.}$$

Interpretation of the F statistic

- MS_G : Quantifies the variability **between** groups.
- MS_E : Quantifies the variability **within** groups.

Thus $F = \frac{MS_G}{MS_E}$ is large when MS_G is “large” with respect to MS_E . The larger F , the more likely that H_0 is false.

- F increases
 - when the group means become more different, or
 - when the variability within groups decreases.
- On the other hand, F decreases
 - when the group means become more similar, or
 - when the variability within groups increases.

► ANOVA App

The ANOVA table

An overview of the results is typically given in an ANOVA table (Varianzanalyse-Tabelle):

Variation	df	SS	MS = SS/df	F	p
Between groups	$g - 1$	SS_G	MS_G	$\frac{MS_G}{MS_E}$	$\Pr(F_{g-1, n-g} > F)$
Within groups	$n - g$	SS_E	MS_E		
Total	$n - 1$	SS_{total}			

Our first ANOVA: Hybrid-Mais example

HYBRID	LOCATION	YIELD	HYBRID	LOCATION	YIELD
FR-11	NW	62	DBC	NW	61
FR-11	NE	64	DBC	NE	64
FR-11	C	64	DBC	C	65
FR-11	SE	65	DBC	SE	62
FR-11	SW	66	DBC	SW	65
BCM	NW	63	RC-3	NW	55
BCM	NE	63	RC-3	NE	56
BCM	C	66	RC-3	C	60
BCM	SE	67	RC-3	SE	58
BCM	SW	64	RC-3	SW	59

```
> str(d.mais)
```

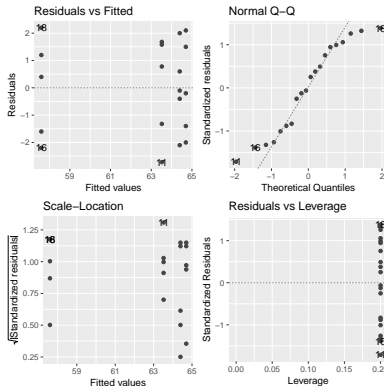
```
'data.frame':    20 obs. of  3 variables:
 $ HYBRID  : Factor w/ 4 levels "BCM","DBC","FR-11",...: 3 3 3 3 3 1 1 1 1 1 ...
 $ LOCATION: Factor w/ 5 levels "C","NE","NW",...: 3 2 1 4 5 3 2 1 4 5 ...
 $ YIELD   : num  62.3 64 64.3 65 66.4 63.3 62.7 66.2 66.8 64.5 ...
```

Hybrid-Mais example – Estimation

Using the `lm()` function in R and then look at the ANOVA table:

```
> r.mais <- lm(YIELD ~ HYBRID, d.mais)
```

As always, before looking at the results, let's do some model checking. This is identical to all we did so far, because we are **still working with linear models!**



Always when we needed to do an F -test and when categorical covariates were involved, the `anova()` table was required:

```
> anova(r.mais)

Analysis of Variance Table

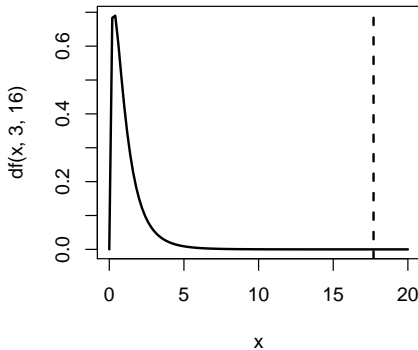
Response: YIELD
          Df Sum Sq Mean Sq F value    Pr(>F)
HYBRID      3  167.441    55.814   17.681 2.474e-05 ***
Residuals   16   50.508     3.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can see that the value of $F = 17.68$ is F -distributed with 3 and 16 degrees of freedom, and the p -value of the test " $\beta_1 = \beta_2 = \beta_3 = \beta_4$?" is < 0.0001 .

Conclusion: the four groups are not all the same!

Exercise: Look at the table a bit closer. How are Df, Sum Sq, Mean Sq, F value and $\Pr(<F)$ related?

The F -distribution with 3 and 16 degrees of freedom, as well as the estimated value $F=17.68$:



You may also want to look at the summary table of the `lm()` output:

```
> summary(r.mais)

Call:
lm(formula = YIELD ~ HYBRID, data = d.mais)

Residuals:
    Min       1Q   Median       3Q      Max
-2.72  -1.45   0.15   1.52   2.20

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.7000    0.7946   81.427 < 2e-16 ***
HYBRIDDBC     -1.1800    1.1237  -1.050  0.309
HYBRIDFR-11   -0.3000    1.1237  -0.267  0.793
HYBRIDRC-3    -7.1000    1.1237  -6.318 1.02e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.777 on 16 degrees of freedom
Multiple R-squared:  0.7683,    Adjusted R-squared:  0.7248
F-statistic: 17.68 on 3 and 16 DF,  p-value: 2.474e-05
```

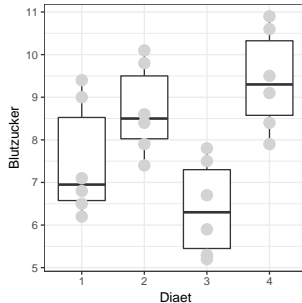
The table contains the estimates of the intercept 64.70 (μ in ANOVA notation, β_0 in regression notation), and estimates for β_2 , β_3 , β_4 .

Exercise: Ernährung und Blutzucker

Remember example 3 from the first week:

24 Personen werden in 4 Gruppen unterteilt. Jede Gruppe erhält eine andere Diät (DIAET). Es werden zu Beginn und am Ende (nach 2 Wochen) die Blutzuckerwerte gemessen. Die Differenz wird gespeichert (BLUTZUCK).

Frage: Unterscheiden sich die Gruppen in der Veränderung der Blutzuckerwerte?



Interpret the results and the residual plots (why is mutate() needed first?):

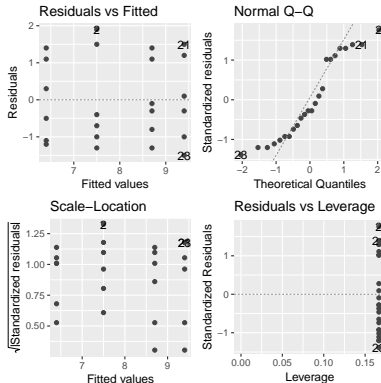
```
> d.blz <- mutate(d.blz,DIAET=as.factor(DIAET))  
> anova(lm(BLUTZUCK ~ DIAET,d.blz))
```

Analysis of Variance Table

Response: BLUTZUCK

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DIAET	3	31.56	10.52	7.5143	0.001476 **
Residuals	20	28.00	1.40		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Multiple comparisons, multiple tests

To remember:

- The F -Test is used to check whether **any two group means** differ.
- Using pairwise tests is not a very good idea (see slide 6), because this leads to a **multiple testing problem**:

When many tests are carried out, the probability to find a “significant” result **by chance** increases.

For instance, for four groups there are $4 \cdot 3/2 = 6$ pairwise combinations that could be tested. The probability to find *at least one result by chance* is much higher than the 5% error level!!!

Post-hoc tests

Still: If the test $\beta_1 = \beta_2 = \dots = \beta_g = 0$ is rejected, one is often interested

- 1 in finding the actual group(s) that deviate(s) from the others.
- 2 in estimates of the pairwise differences.

Several methods to circumvent the problem of too many “significant” test results (type-I error) have been proposed. The most prominent ones are:

- Bonferroni correction
- Tukey honest significant differences (HSD) approach
- Fisher least significant differences (LSD) approach

Bonferroni correction

Idea: If a total of m tests are carried out, simply divide the type-I error level α_0 (often 5%) such that

$$\alpha = \alpha_0 / m .$$

Tukey HSD approach

Idea: Take into account the distribution of *ranges* (max-min) and design a new test.

Fisher's LSD approach

Idea: Adjust the idea of a two-sample test, but use a larger variance (namely the pooled variance of all groups).

Calculate the pairwise differences and tests with adjustments for the “Blutzucker” example:

Differences:

	1	2	3
2	1.2		
3	-1.1	-2.3	
4	1.9	0.7	3.0

Tukey HSD p -values:

	1	2	3
2	0.32		
3	0.40	0.01	
4	0.05	0.74	0.001

Bonferroni p -values:

	1	2	3
2	0.57		
3	0.74	0.02	
4	0.07	1.00	0.002

Fisher p -values:

	1	2	3
2	0.09		
3	0.12	0.003	
4	0.01	0.32	3e-04

- Bonferroni p -values are the most conservative (largest p).
- Fisher p -values are the least conservative (smallest p).

Other contrasts

Sometimes additional comparisons are of interest. For example, a new diet is to be compared to other, existing diets.

In the “Blutzucker” example, this could be, for instance: “Is diet 1 different from diets 2-4?”

(Check also chapter 5.6.5 in GSWR, 2nd edition)

Two-way ANOVA (Zweiweg-Varianzanalyse)

Example (from Hand et al. 1994 / Hothorn/Everitt "A Handbook of Statistical Analyses Using R"):

Experiment to study the weight gain of rats, depending on four diets.

Protein amounts were either high or low, and the protein source was either beef or cereal. 10 rats for each diet were selected.

Question: How does diet affect weightgain?

Complication: This is a factorial design (gekreuzte Faktoren), because each combination of protein source (beef/cereal) \times level (high/low) is present (2×2 groups).

Design:

	beef	cereal
high	group 1	group 2
low	group 3	group 4

Start by looking at means and standard deviations in the groups, as well at a graphical description of the means:

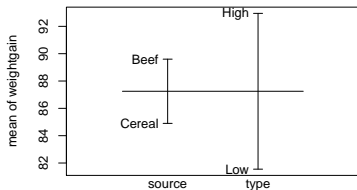
```
> summarise(group_by(weightgain,source,type),meanW = mean(weightgain),sdW = sd(weightgain))
```

```
Source: local data frame [4 x 4]
```

```
Groups: source [?]
```

	source	type	meanW	sdW
	<fctr>	<fctr>	<dbl>	<dbl>
1	Beef	High	100.0	15.13642
2	Beef	Low	79.2	13.88684
3	Cereal	High	85.9	15.02184
4	Cereal	Low	83.9	15.70881

```
> plot.design(weightgain)
```



Factors

- Protein source (beef/cereal) seems less important than the amount (high/low).
- Variances seem to be equal in the four groups.

Two-way ANOVA – The model

In the presence of a **factorial design**, the idea is to add separate effects β_i (here $i = 1, 2$) and γ_j (here $j = 1, 2$) for the i th level of the first factor and the j th level of the second factor:

Assume we have a factorial design with two factors β_i and γ_j , then the k th outcome in the group of i and j , y_{ijk} is modelled as

$$y_{ijk} = \mu + \beta_i + \gamma_j + e_{ijk} \quad \text{with} \quad e_{ijk} \sim N(0, \sigma_e^2) \quad i.i.d.$$

Again, additional constraints are needed!

- $\beta_1 = \gamma_1 = 0$ (**treatment contrast**; default in R).
- $\sum_i \beta_i = \sum_i \gamma_i = 0$ (**sum-to-zero contrast**).

Two-way ANOVA in R

In R, a two-way ANOVA is as simple as one-way ANOVA, just add another variable:

```
> r.weight <- lm(weightgain ~ source + type, weightgain)
> anova(r.weight)
```

Analysis of Variance Table

Response: weightgain

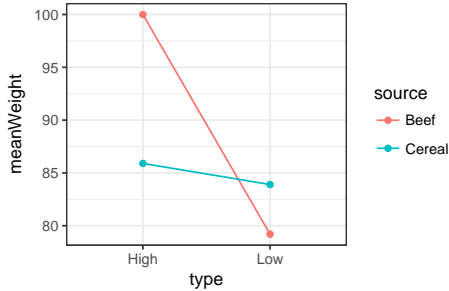
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
source	1	220.9	220.90	0.9150	0.34501
type	1	1299.6	1299.60	5.3829	0.02596 *
Residuals	37	8933.0	241.43		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Interpretation: There seems to be a difference between low and high amounts of protein, but the source (beef/cereal) seems less relevant.

However: what if the additive model does not hold?

A so-called **interaction plot** helps to understand if the additive model is reasonable:



The lines are **not parallel**, indicating that **there is an interaction** between type and source!

Note: if the additive model $\beta_i + \gamma_j$ holds, the lines would be parallel.

Two-way ANOVA with interaction

- If the purely additive model is not correct, a more general model with an interaction term $(\beta\gamma)_{ij}$ may be used:

$$y_{ijk} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + e_{ijk} \quad \text{with} \quad e_{ijk} \sim N(0, \sigma_e^2) \quad i.i.d.$$

- As in linear regression, interactions allow for an **interplay between the variables**.
- In the rats experiment, increasing the amount from low to high has a different effect in the beef than in the cereal diet.
- Moreover: The plot on the previous slide shows that for the low amount of proteins case, the cereal diet leads to a larger average weight gain!

Two-way ANOVA in R – Including an interaction

Again the rats example, this time including the interaction term:

```
> r.weight2 <- lm(weightgain ~ source * type, weightgain)
> anova(r.weight2)
```

Analysis of Variance Table

Response: weightgain

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
source	1	220.9	220.90	0.9879	0.32688
type	1	1299.6	1299.60	5.8123	0.02114 *
source:type	1	883.6	883.60	3.9518	0.05447 .
Residuals	36	8049.4	223.59		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The coefficient estimates can be obtained as follows:

```
> summary(r.weight2)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	100.0	4.728577	21.148009	6.842420e-22
sourceCereal	-14.1	6.687218	-2.108500	4.201233e-02
typeLow	-20.8	6.687218	-3.110411	3.644273e-03
sourceCereal:typeLow	18.8	9.457155	1.987913	5.446757e-02

Interpretation of the coefficients

This works in the same way as for categorical covariates in regression! To see this, let us estimate the means from the model. From the above output, we have [because of using treatment contrasts]:

$$\hat{\beta}_{beef} = 0, \hat{\beta}_{cereal} = -14.1,$$

$$\hat{\gamma}_{high} = 0, \hat{\gamma}_{low} = -20.8,$$

$$(\hat{\beta}\hat{\gamma})_{cereal/low} = 18.8, (\hat{\beta}\hat{\gamma})_{beef/high} = (\hat{\beta}\hat{\gamma})_{beef/low} = (\hat{\beta}\hat{\gamma})_{cereal/high} = 0.$$

Therefore:

$$\text{Group 1: beef / high} \quad \hat{y}_{beef,high} = 100 + 0 + 0 + 0 = 100$$

$$\text{Group 2: cereal / high} \quad \hat{y}_{cereal,high} = 100 + (-14.1) + 0 + 0 = 85.9$$

$$\text{Group 3: beef / low} \quad \hat{y}_{beef,low} = 100 + 0 + (-20.8) + 0 = 79.2$$

$$\text{Group 4: cereal / low} \quad \hat{y}_{cereal,low} = 100 + (-14.1) + (-20.8) + 18.8 = 83.9$$

Compare these values to slide 29!

A cautionary note

Be careful: In the presence of interactions, the p -values of the main effects should no longer be interpreted as before!

It is then required that separate analyses are carried out. For example for "Beef" and "Cereal" protein sources:

```
> anova(lm(weightgain ~ type,subset(weightgain,source=="Beef")))
```

Analysis of Variance Table

Response: weightgain

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	2163.2	2163.20	10.253	0.00494 **
Residuals	18	3797.6	210.98		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

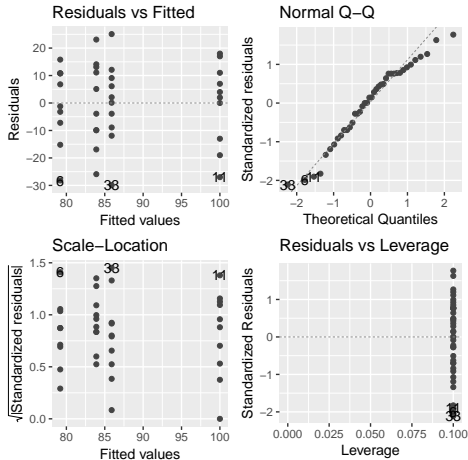
```
> anova(lm(weightgain ~ type,subset(weightgain,source=="Cereal")))
```

Analysis of Variance Table

Response: weightgain

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	20.0	20.00	0.0847	0.7744
Residuals	18	4251.8	236.21		

And finally (what we should have done before checking the results!), the model diagnostics:



Exercise:

In an experiment the influence of four levels of fertilizer (DUENGER) on the yield (ERTRAG) on 5 species (SORTE) of crops was investigated. The data contains the following columns:

DUENGER (4 levels)

SORTE (5 levels)

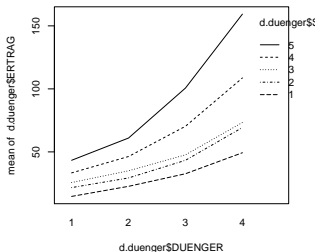
ERTRAG

The first 10 rows of the data:

```
> head(d.duenger, 10)
```

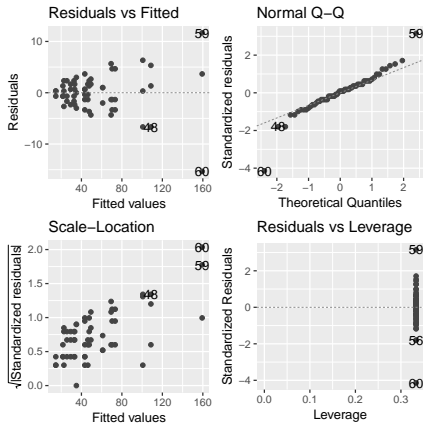
	DUENGER	SORTE	ERTRAG
1	1	1	14
2	1	1	15
3	1	1	15
4	2	1	20
5	2	1	25
6	2	1	23
7	3	1	35
8	3	1	31
9	3	1	32
10	4	1	52

And the interaction plot:



The interaction plot indicates that an interaction between SORTE and DUENGER is needed in the analysis. The results and residual plots are given here and on the next slide:

```
> d.duenger <- mutate(d.duenger, SORTE=as.factor(SORTE), DUENGER=as.factor(DUENGER))  
> r.duenger <- lm(ERTRAG ~ DUENGER*SORTE, d.duenger)
```




```
> anova(r.duenger)
```

```
Analysis of Variance Table
```

```
Response: ERTRAG
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DUENGER	3	35801	11933.5	580.707	< 2.2e-16 ***
SORTE	4	27805	6951.3	338.262	< 2.2e-16 ***
DUENGER:SORTE	12	7674	639.5	31.121	< 2.2e-16 ***
Residuals	40	822	20.6		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is here the problem (look at the TA and the scale-location plots)?
Ideas? Interpretation?

Some summary remarks

- The t -test to compare the mean of two groups is a special case of ANOVA.
- ANOVA is a special cases of the linear regression model.
- ANOVA is often taught in separate lectures, although it could be integrated in a lecture on linear regression.
- ANOVA is traditionally most used to analyze experimental data.