Basic Statistics

Katharina J. Hoff[1,2]

[1]Institute for Mathematics and Computer Science
Bioinformatics Group
University of Greifswald [2]Center for Functional Genomics of Microbes
University of Greifswald

July 2024

Data Types

Quantitative Data

- continuous
  - interval-scaled *[^oC: $10^oC \neq 2*5^oC$]*
  - ratio-scaled *[absolute scale, e.g. Kelvin]*
  - **approx. gaussian distributed**
  - not approx. gaussian distributed
- discrete
  - countable *[e.g. number or kids 1, 2, 3, ...]*
  - not countable

Distributions...

Examples:

- Cauchy distribution
- Binomial distribution
- Discrete distribution
- Poisson distribution
- Exponential distribution
- ...

The kind of statistics covered in this lecture today usually imply a **normal distribution** (also called Gaussian distribution).

Normal distribution

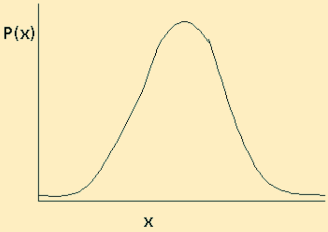
Standard normal distribution

$$P(x)dx = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Normal distribution with μ and σ^2

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma)^2}$$

Probability function



Does my data come from a normal distribution?

Ideas:

- **Statistical test** with H_0: data comes from a normal distribution.
  Example: Komogorov-Smirnov Test
  Problem: result is either
  - to a confidence of a fixed α, this data does not come from a normal distribution, or
  - with an **unknown error** β, the data comes from a normal distribution.
- **Visual inspection**
  Examples: Histogram, Boxplot
  Problem: unknown reliability.

Type I and type II errors

	H_0 is true	H_0 is false
Reject H_0	type I error (α)	correct
Fail to reject H_0	correct	type II error (β)

It is possible to calculate the required case number to achieve a small β-error. But often, β is unknown!

Visualization of 1D data

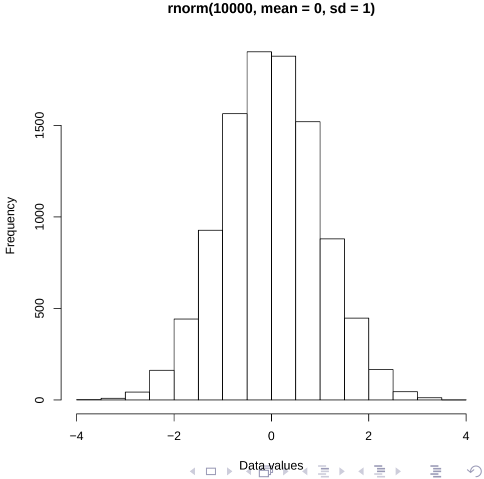
Histogram

Advantage

- intuitive interpretation

Disadvantage

- for n samples, n plots are needed



Visualization of 1D data

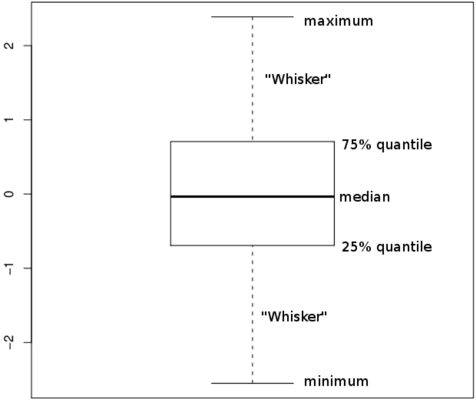
Boxplot

Advantages

- visualization of multiple samples is possible → comparing samples
- quantiles, even for small sample sizes

Disadvantage

- only quantiles are shown



Descriptive statistics

$$Min(x) = \text{smallest value in a sample} x$$

$$Max(x) = \text{largest value in a sample} x$$

$$Median(x) = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ even} \end{cases}$$

Descriptive statistics

Quartiles

0%
25%
50%
75%
100%

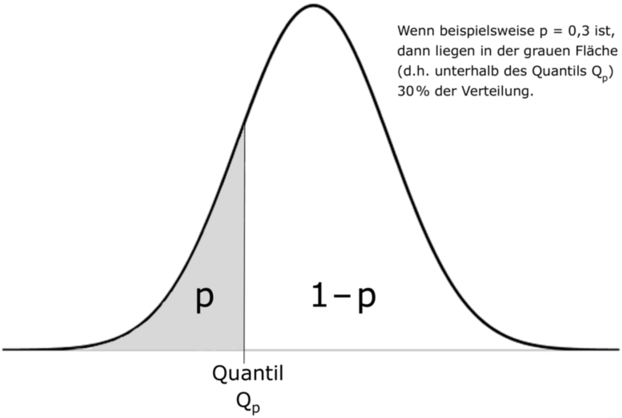


Image: Marcus Glöder, 2006

Descriptive statistics

And while we are at it...

$$Mean(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Sd(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - Mean(x))^2}{n-1}}$$

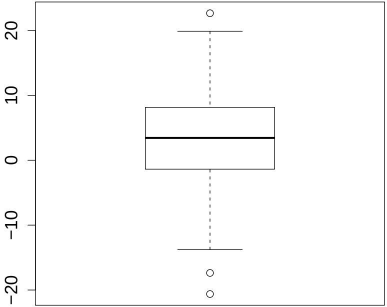
$$Var(x) = (Sd(x))^2$$

$$IQR(x) = Q75\% - Q25\%$$

Outliers

What is an outlier?

A data point that doesn't match your expectations.
(In R: values more extreme than 1.5*IQR)



Discard outliers?

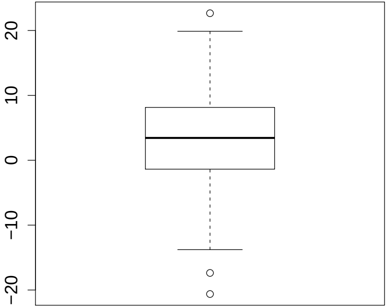
If you have a **good reason**!
Examples:

- measuring machine broke exactly at this data point
- it is kown that the treatment went wrong for this data point
- ...

Outliers

What is an outlier?

A data point that doesn't match your expectations.
(In R: values more extreme than 1.5*IQR)

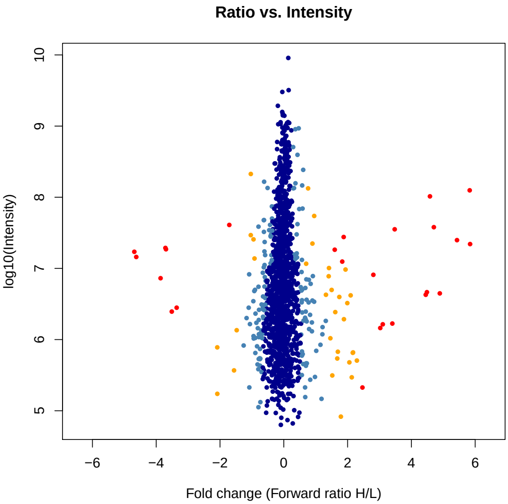


Discard outliers?

Never discard outliers without a logical explannation!

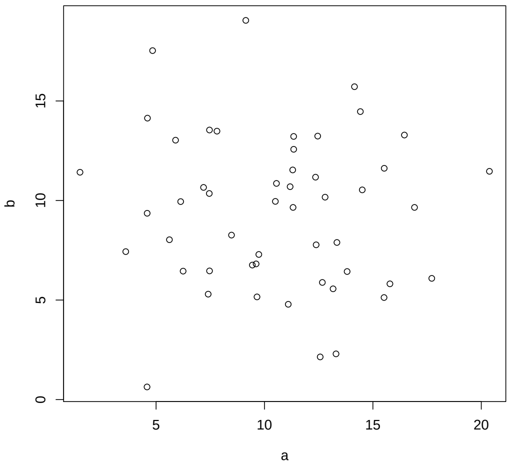
Visualization of 2D data

Scatterplot



Visualization of 2D data

Scatterplot



Correlation

Question

Are 2 variables A and B dependent on each other (linearly)?
$\Longrightarrow$ Are they **correlated**?

Pearson's correlation coefficient

$$\rho = \frac{Cov(A,B)}{\sqrt{Var(A)Var(B)}}$$

$$Cov(A,B) = \sum_{i=1}^{N} \frac{(a_i - mean(A))(b_i - mean(B))}{N}$$

Prerequisites

- normal distribution
- observations must be independent from each other

Is a correlation coefficient meaningful?

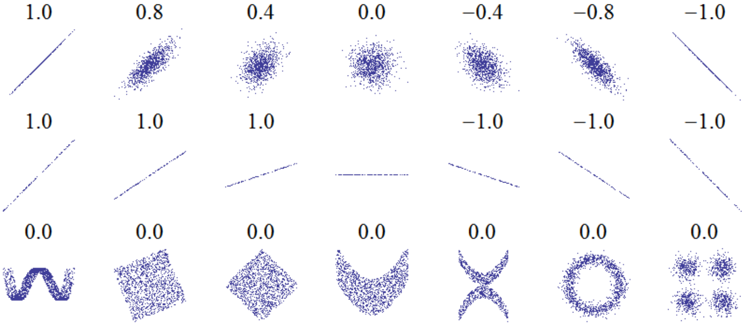


Image:
http://upload.wikimedia.org/wikipedia/commons/0/02/Correlation_examples.png

Testing for significance of a correlation coefficient

Step 1: Formulate hypothesis

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Testing for significance of a correlation coefficient

Step 2: Calculate test statistic

$$t = \rho \frac{\sqrt{N-2}}{\sqrt{1-\rho^2}}$$

Testing for significance of a correlation coefficient

Step 3: Compare to value for α and df in table

- α-error is usually 5%
- degrees of freedom: $N - 2$
- If the test statistic is more extreme than the table value, then H_0 is rejected.
- If the test statistic is not more extreme than the table value, then H_0 cannot be rejected.

Testing for significance of a correlation coefficient

Output from statistics software: **p-value**

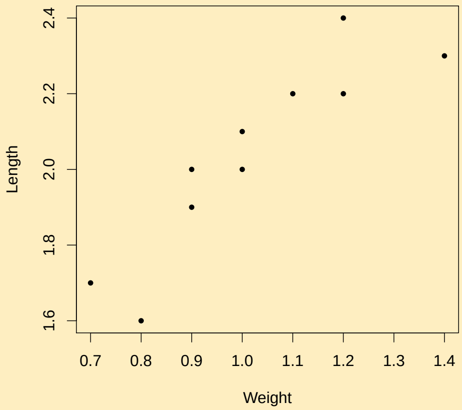
Definition
Assuming that the null hypothesis is true, the p-value is the probability to obtain a test statistic at least as extreme as the one that was actually observed.

Interpretation

- set an a priori α-error, e.g. 0.05
- compare p-value to α
- if p-value $\leq \alpha \Rightarrow$ reject H_0
- if p-value $> \alpha \Rightarrow$ do not reject H_0

Correlation example

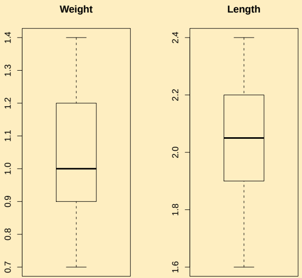
Length and weight of beans was measured:



Correlation example

Prerequisites

- independent observations? → we hope that!
- normal distribution? → hard to say, few data points.
  Boxplot:



→ we assume normal distribution!

Correlation example

Hypotheses

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$\alpha = 0.05$$

Correlation example

Calculate Pearson's correlation coefficient

$$\rho = \frac{Cov(weight, length)}{\sqrt{Var(weight)Var(length)}} = \frac{0.048}{\sqrt{0.044 * 0.06489}} = 0.8983$$

Calculate test statistic

$$t = \rho\frac{\sqrt{N-2}}{\sqrt{1-\rho^2}} = 0.8983\frac{\sqrt{10-2}}{\sqrt{1-0.8983^2}} = 5.7832$$

Check table

For $df = N - 2 = 8$, $\alpha = 0.05$, the "critical value" is **2.306**. 5.7832 is more extreme.

Therefore, we reject the H_0! To an α-error of 5%, ρ of 0.8983 is significantly different from zero!

Correlation example

Statistic software output (example R):

```
t = 5.7832, df = 8, p-value = 0.000413
sample estimates:
 cor
0.8983172
```

→ The p-value is smaller than 0.05
→ to an α-error of 5%, ρ of 0.8983 is significantly different from zero!

Spearman's rank correlation

Prerequisites

- observations must be independent from each other

Spearman's correlation coefficient

- given are n raw scores X_i, Y_i
- convert to ranks x_i, y_i
- calculate differences $d_i = x_i - y_i$
- $$\rho^* = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

*) for data without ties

Spearman's rank correlation

Rank conversion

Raw data:

Index	1	2	3	4	5	6	7	8	9	10
Weight	0.7	1.2	0.95	1.4	1.25	1.1	1.05	0.9	1.0	0.8
Length	1.7	2.2	2.05	2.3	2.4	2.25	2.0	1.9	2.1	1.6

Sort values

Index	10	1	8	7	3	9	2	6	4	5
Length	1.6	1.7	1.9	2.0	2.05	2.1	2.2	2.25	2.3	2.4

Index	1	10	8	3	9	7	6	2	5	4
Weight	0.7	0.8	0.9	0.95	1.0	1.05	1.1	1.2	1.25	1.4

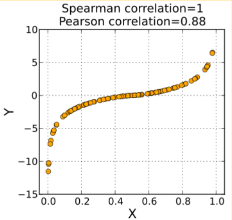
Rank	1	2	3	4	5	6	7	8	9	10

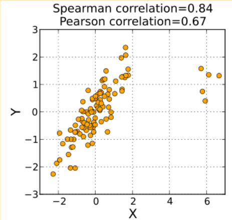
Rank data for calculations

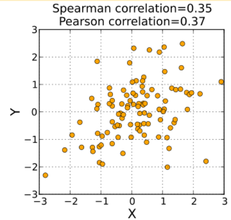
Weight	1	8	4	10	9	7	6	3	5	2
Length	2	7	5	9	10	8	4	3	6	1

Spearman's rank correlation

When is Spearman's coefficient more suitable?





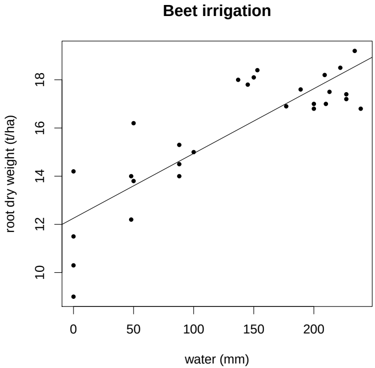


Images: by Skbkekas at http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Linear Regression

Question

Is there a functional dependence (linear) between a target variable and a quantitative influencing variable?



$$y_i = \alpha + \beta x + \epsilon_i$$

Idea: minimize squares to get optimal model!

Linear Regression

Prerequisites

- normal distribution of residuals
- (homogeneity of variances of residuals)

Equations...

- Sum of squares:

$$SQ(x) = \sum_{j=1}^{m} n_j x_j^2 - \frac{(\sum_{j=1}^{m} n_j x_j)^2}{\sum n_j}$$

- Sum of products:

$$SP(xy) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_j y_{ij} - \frac{\sum_{j=1}^{m} n_j x_j \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij}}{n}$$

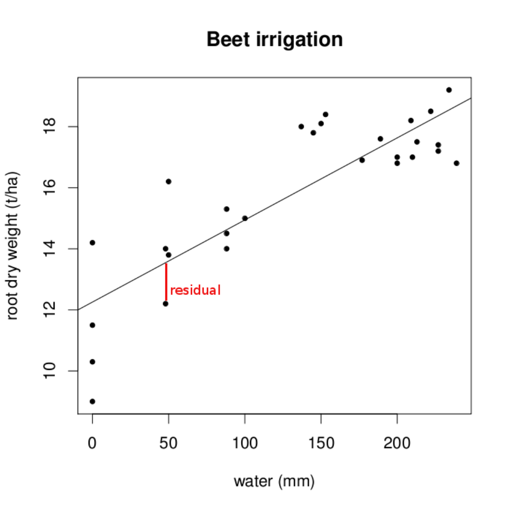
- Estimators:

$$b = \frac{SP(xy)}{SQ(x)}$$

$$a = \overline{y}_{..} - b\overline{x}_{.} \text{ with } \overline{y}_{..} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{i} y_{ij} \text{ and } \overline{x}_{.} = \frac{1}{n} \sum_{j=1}^{m} n_j x_j$$

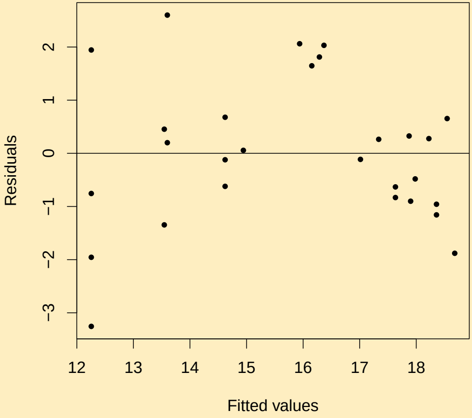
Is my regression model good?

What are residuals?



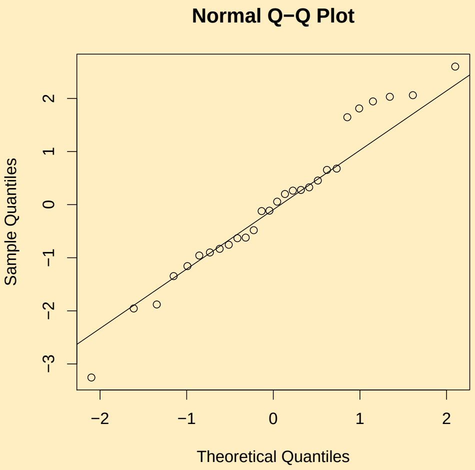
Is my regression model good?

Residual plot



Is my regression model good?

QQ-Plot



Significance of a regression model

- t-tests for parameters
- similar to significance test for correlation coefficient: $H_0^1 : \alpha = 0$ and $H_0^1 : \beta = 0$
- Example software output (R):

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.255531 0.505482 24.245  < 2e-16 ***
water          0.026881 0.003261  8.244 1.00e-08 ***
```

Statistical Tests

- two-sample tests
  - t-test
  - wilcoxon test
- tests for more than two samples
  - ANOVA
  - multiple t-tests
  - non-parametric alternative `nparcomp`

From hypotheses to interpretation

1. state hypotheses (one-sided/two-sided)
2. set α-error
3. calculate test-statistic
4. interpretation:
    - compare to critical value (df/α)
    - compare p-value to α
    - interpret confidence intervals

From hypotheses to interpretation

two-sided hypotheses

$$H_0 : x = y$$

$$H_1 : x \neq y$$

one-sided hypotheses (greater)

$$H_0 : x \leq y$$

$$H_1 : x > y$$

one-sided hypotheses (less)

$$H_0 : x \geq y$$

$$H_1 : x < y$$

t-Test

There are several t-tests:

- one-sample t-test
- **two-sample t-test**
- paired t-test
- two-sample t-Welch-test

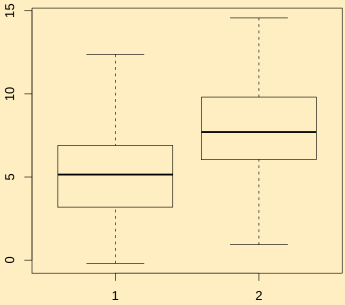
two-sample t-test

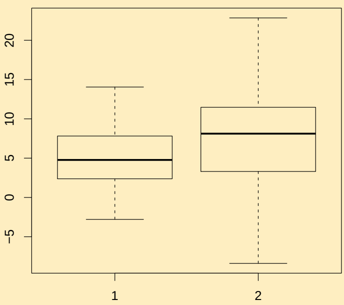
Prerequisites

- approx. normal distribution of both samples
- homogeneity of variances
- independence of data

two-sample t-test

Are the variances homogeneous?





and/or F-test (in R: `var.test()`)

two-sample t-test

t-statistic

Given 2 samples $x : x_1, ..., x_n$ and $y : y_1, ..., y_m$:

$$T = \sqrt{\frac{nm}{n+m}} \frac{mean(x) - mean(y)}{\sqrt{\frac{(n-1)var(x) + (m-1)var(y)}{n+m-2}}}$$

$$df = n + m - 2$$

two-sample t-test

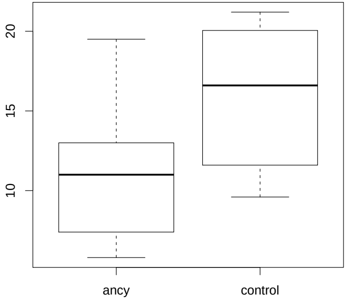
What if the variances are heterogeneous???

$\Longrightarrow$ two-sample t-Welch-test is more robust!

Example: t-Test

Brassica campestris plants were treated with Ancymidol. Control group was treated with water. Target was the height of plants in cm.

Control	Treatment
10.0	13.2
13.2	19.5
19.8	11.0
19.3	5.8
21.2	12.8
13.9	7.1
20.3	7.7
9.6	



Example: t-Test

Brassica campestris plants were treated with Ancymidol. Control group was treated with water. Target was the height of plants in cm.

Control	Treatment
10.0	13.2
13.2	19.5
19.8	11.0
19.3	5.8
21.2	12.8
13.9	7.1
20.3	7.7
9.6	

$$mean(Control) = 15.91$$

$$mean(Treatment) = 11.01$$

$$var(Control) = 22.85$$

$$var(Treatment) = 22.24$$

$$n_{Control} = 8$$

$$m_{Treatment} = 7$$

Example: t-Test

$$H_0 : \mu_{Control} = \mu_{Treatment}$$
$$H_1 : \mu_{Control} \neq \mu_{Treatment}$$

Example: t-Test

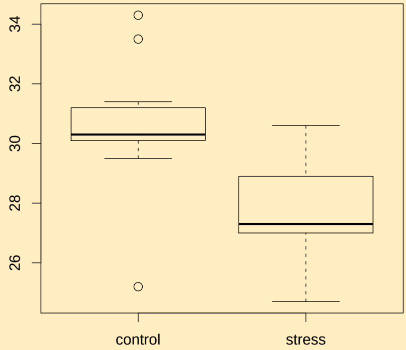
$$T = \sqrt{\frac{7 * 8}{7}} \frac{11.01 - 15.91}{\sqrt{\frac{(7-1) * 22.85 + (8-1) * 11.85}{8+7-2}}} = -1.99$$

Critical value for $\alpha = 0.05$ and $df = 8 + 7 - 2 = 13$ is -1.77.

T is not more extreme than critical value $\Longrightarrow$ H_0 cannot be rejected.

Wilcoxon-Test

My data doesn't look like it's normal?!



Wilcoxon-Test

Prerequisites

- test for two samples
- homogeneity of variances
- at least ordinal data
- independence of data

Test procedure

similar to t-test:

- one-sided or two-sided hypotheses
- compare test statistic to critical value
- interprete test statistic or p-value

Wilcoxon-Test

W-statistic

- pool samples
- order values, give ranks
- $$W_{m,n} = \sum_{i=1}^{m} R(X_i)$$
  where $R(X_i)$ is the rank of X_i

Example: How to calculate W-statistic

Raw data:

A	25.2	29.5	30.1	30.2	31.1	34.3
B	24.7	25.7	27	30	28.9	

Ranks:

1	2	3	4	5	6	7	8	9	10	11
24.7	25.2	25.7	27	28.9	29.5	30	30.1	30.2	31.1	34.3
B	A	B	B	B	A	B	A	A	A	A

W-statistic:

$$W_{5,6} = \sum_{i=1}^{5} R(B_i) = 1 + 3 + 4 + 5 + 7 = 20$$

Analysis of variances

- univariate **an**alysis **o**f **v**ariances (ANOVA, one target variable)
- **m**ultivariate **an**alysis **o**f **va**riances (MANOVA, several target variables)
- one-way ANOVA (one treatment variable)
- factorial ANOVA (several treatment variables)
- ...

ANOVA

Is there a difference between (more than two) samples due to the treatment?

Use variances to check! If variances overlap: no difference.

Example for two-factorial ANOVA model (including interaction term):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

ANOVA

Prerequisites

- independence of cases
- normality of residuals
- homogeneity of sample (and residual) variances

ANOVA

Cause of var.	Sum Sq.	df	Mean Sq.	F	p-value
Between factor A	SQ_A	df_A	MQ_A	F_A	p_A
Between factor B	SQ_B	df_B	MQ_B	F_B	p_B
...	...	...	...	...	...
Within groups	SQ_E	df_E	MQ_E		
Total	SQ_T	df_T			

ANOVA

Calculating one-way ANOVA manually: Data

Rep.	Treatm. 1	Treatm. 2	Treatm. 3	
1	3	4	8	
2	2	2	4	
3	3	7	7	
Y_i	8	13	19	$Y_{..} = 40$
n_i	3	3	3	N = 9
$\overline{Y}_{i.}$	2.66	4.33	6.33	$\overline{Y}_{..} = 4.44$

Y_i is the sum of the ith sample
$Y_{..}$ is the sum of all sample sums
n_i is the number of repetitions in the ith sample
N is the number of all repetitions
$\overline{Y}_{i.}$ is the ith sample mean
$\overline{Y}_{..}$ is the mean of all samples

ANOVA

Calculating one-way ANOVA manually: SQ_A

$$SQ_A = \sum_{i=1}^{a} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

$$= 3(2.66 - 4.44)^2 + 3(4.33 - 4.44)^2 + 3(6.33 - 4.44)^2 = 20.26$$

Calculating one-way ANOVA manually: MQ_A

$$MQ_A = \frac{SQ_A}{(a-1)} = \frac{20.26}{(3-1)} = 10.13$$

ANOVA

Calculating one-way ANOVA manually: SQ_E

$$SQ_E = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \overline{Y}_{i.})^2$$

$$= (3-2.66)^2 + (2-2.66)^2 + (3-2.66)^2 + (4-4.33)^2 + (2-4.33)^2$$

$$+(7-4.33)^2 + (8-6.33)^2 + (4-6.33)^2 + (7-6.33)^2 = 22$$

Calculating one-way ANOVA manually: MQ_E

$$MQ_E = \frac{SQ_E}{(N-a)} = \frac{22}{(9-3)} = 3.66$$

ANOVA

Calculating one-way ANOVA manually: F

$$F = \frac{MQ_A}{MQ_E} = \frac{10.13}{3.66} = 2.76$$

$$df_A = 3 - 1 = 2$$

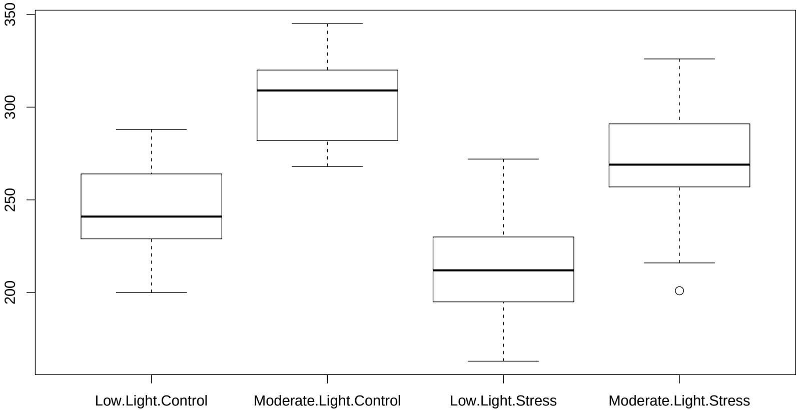
$$df_E = 9 - 3 = 6$$

$\Longrightarrow$ compare to critical value...

Example: two-factorial ANOVA (with software R)

Control Low Light	Stress Low Light	Control Moderate Light	Stress Moderate Light
264	235	314	283
200	188	320	312
225	195	310	291
268	205	340	259
215	212	299	216
241	214	268	201
232	182	345	267
256	215	271	326
229	272	285	241
288	163	309	291
253	230	337	269
288	255	282	282
230	202	273	257

Example: two-factorial ANOVA (with software R)



Example: two-factorial ANOVA (with software R)

Hypotheses

$H_0^A : \mu_{stress} = \mu_{nostress}$ $H_0^B : \mu_{light} = \mu_{dark}$

$H_1^A : \mu_{stress} \neq \mu_{nostress}$ $H_1^B : \mu_{light} \neq \mu_{dark}$

$H_0^{A'B}$: $\mu_{stressfactor,lightfactor} = \mu_{stressfactor} + \mu_{lightfactor} - \mu$

$H_1^{A'B}$: $\mu_{stressfactor,lightfactor} \neq \mu_{stressfactor} + \mu_{lightfactor} - \mu$

Example: two-factorial ANOVA (with software R)

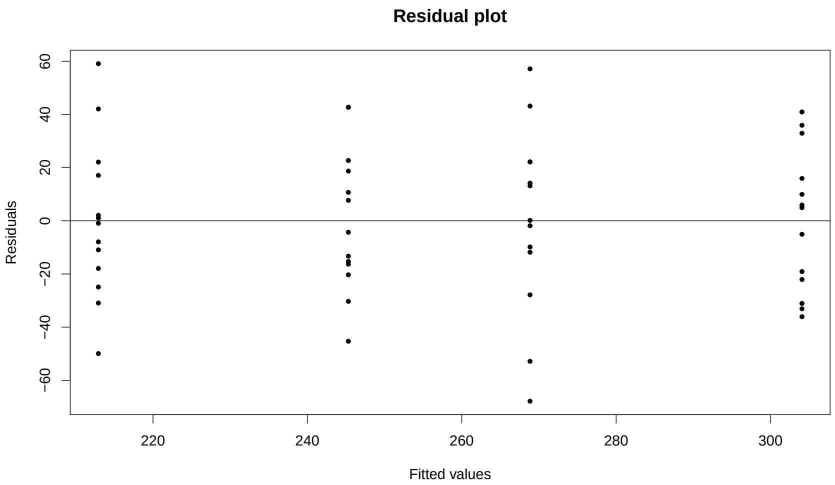
Model

$$Y_{ijk} = \mu + stressfactor_i + lightfactor_j + (stressfactor\ lightfactor)_{ij} + \epsilon_{ijk}$$

Computed table of variances

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Light	1	42752	42752	47.7490	1.010e-08
Stress	1	14858	14858	16.5954	0.0001725
Light:Stress	1	26	26	0.0294	0.8645695
Residuals	48	42976	895		

Example: two-factorial ANOVA (with software R)



Example: two-factorial ANOVA (with software R)

- observations are (hopefully) independent
- variances of groups are equal
- distribution of residuals is normal
- p-value of interaction is >0.05 $\Rightarrow$ not significant
- p-values of Light and Stress are <0.05 $\Rightarrow$ significant
- easy example because of two influencing levels, only!
  More levels $\rightarrow$ don't know location of differences!

But I have to know where exactly is the difference...

$\rightarrow$ multiple two-sample tests!

Image: http://de.dreamstime.com/lizenzfreies-stockbild-smiley-image6259996

Multiple tests: what's the problem?

Example

- k = 5 groups
- r = 3 variables
- t = 4 time points

$$\text{Number of tests} = \frac{k(k-1)}{2} * r * t = \frac{5(5-1)}{2} * 3 * 4 = 120$$

- α = 0.05 means 5 out of 100 tests are significant although there is no effect
- with 120 tests, we have to expect 6 significant results although there is no effect! (6/120 = 0.05)

$\Longrightarrow$ we should not conduct multiple tests to local α!

Solutions for multiple comparisons

- Bonferroni
- Holm-Bonferroni
- Benjamini-Hochberg
- Westfall-Young
- ...

Solutions for multiple comparisons

Bonferroni

Multiply p-values with the total number of comparisons;
compare to $\alpha = 0.05$.

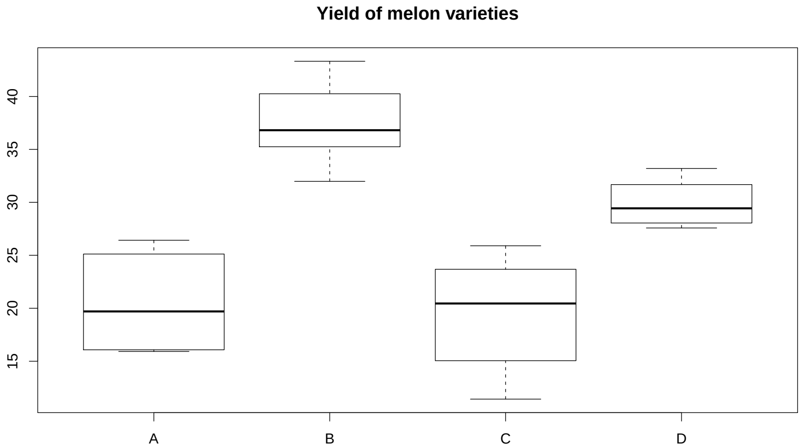
Holm-Bonferroni

- sort p-values
- multiply the smallest p-value with number of comparisons N
- if significant $\rightarrow$ multiply second smallest p-value with $N - 1$
- if significant $\rightarrow$ multiply third smallest p-value with $N - 2$
- ...
- stop at the first non-significant p-value!

What is the test for multiple comparisons?

- choose a test that is suitable for your data, e.g. t-test, t-Welch-test, Wilcoxon-test, npar.t.test, etc.
- if possible, try to minimize the number of tests, e.g. many-to-one instead of all-pairs-comparison

Example: Multiple comparisons (t-test)



Example: Multiple comparisons (t-test)

Hypotheses

H$_0$: $\mu_A = \mu_B = \mu_C = \mu_D$

H$_1$: $\mu_A \neq \mu_B$

$\mu_A \neq \mu_C$

$\mu_A \neq \mu_D$

$\mu_B \neq \mu_C$

$\mu_B \neq \mu_D$

$\mu_C \neq \mu_D$

Example: Multiple comparisons (t-test)

two-sample tests

→ use t-Welch-test (robust towards heterogeneity of variances)

Test	t-statistic	p-value
A vs. B	-6.75	0.0000578
A vs. C	0.34	0.744
A vs. D	-4.43	0.002881
B vs. C	6.43	0.0001192
B vs. D	4.05	0.003772
C vs. D	-4.25	0.004361

Example: Multiple comparisons (t-test)

Bonferroni & Holm-Bonferroni correction

Test	p-value	$p_{Bonferroni}$	p_{Holm}
A vs. B	0.0000578	0.0003468 *	0.0003468 *
B vs. C	0.0001192	0.0007152 *	0.000596 *
A vs. D	0.002881	0.017286 *	0.011524 *
B vs. D	0.003772	0.022632 *	0.011316 *
C vs. D	0.004361	0.026166 *	0.008722 *
A vs. C	0.744	1	0.744

Correlation

Are 2 variables A and B dependent on each other?

- Pearson: data is normal
- Spearman: data is not normal

Regression

Is there a functional dependence (linear) between a target variable and an influencing variable?

- normal distribution of residuals
- (homogeneity of variances of residuals)

2-sample tests

Is there a difference in (parameter x) between two samples?

- samples are normal, variances are homogeneous → t-test
- samples are normal, variances are heterogeneous → t-Welch-test
- samples are not normal, variances are homogeneous → Wilcoxon-test
- samples are not normal, variances are heterogeneous → npar.t.test (R package nparcomp)

Possible hypotheses:

- two-sided
- less
- greater

ANOVA

Is there a difference between several samples due to one (or more) influencing variables?

- normality of residuals
- homogeneity of sample variances

Multiple comparisons

The easiest way:

- two-sample tests for local differences
- correct p-values for multiplicity (e.g. Bonferroni)

Useful R-packages for MCPs:

- multcomp
- multtest
- nparcomp

Acknowledgements

- Prof. Dr. Ludwig Hothorn, Hannover
- Dr. Frank Bretz, Zürich
- Prof. Dr. Edgar Brunner, Göttingen
- Dr. Frank Konietschke, Göttingen
- Dr. Jan-Erik Englund, Alnarp
- Prof. Dr. Mario Stanke, Greifswald
- Anne-Kathrin Schultz, Göttingen

Acknowledgements

Thank you for your attention!