Statistics for EES and others

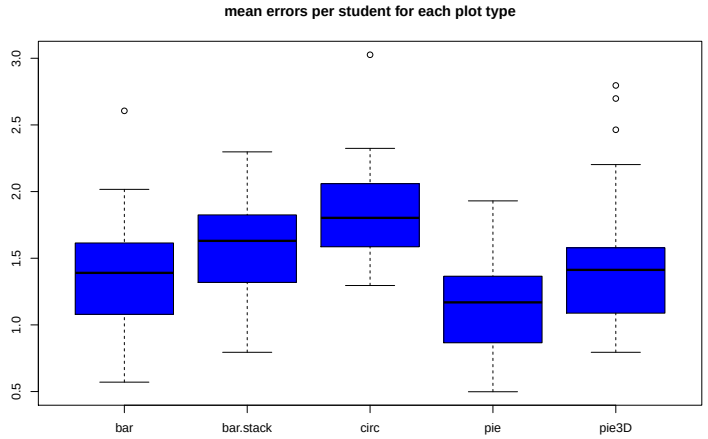
Comparing more than two groups: Multiple testing, ANOVA and Kruskal-Wallis

Dirk Metzler

June 1, 2021

Contents

1 Pairwise comparisons and multiple-testing corrections



Let's test whether errors (averaged for each student) differ significantly between any plot types.

```
> str(bam)
'data.frame': 120 obs. of  3 variables:
 $ plot.type: Factor w/ 5 levels "bar","bar.stack",..: 1 2 3 4 5 1 2 3 4 5 ...
 $ student  : Factor w/ 24 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ mean.err : num  1.07 1.39 1.68 1.35 1.42 ...


> pairwise.t.test(bam$mean.err,bam$plot.type,paired=TRUE,p.adjust="none")

Pairwise comparisons using paired t tests

data:  bam$mean.err and bam$plot.type

          bar     bar.stack circ    pie
bar.stack 0.06760 -         -       -
circ      4.8e-05 0.07526   -       -
pie       0.02918 0.00018   1.3e-06 -
pie3D     0.37389 0.33271   0.01560 0.01988

P value adjustment method: none
```

We see 10 pairwise comparisons. On the 5% level, some of them are significant.

Problem of multiple testing: If the null hypothesis ("only random deviations") is true, we falsly reject it with a probability of 5% in each test.If we then apply 10 or more test, the expectation value for the number of falsely rejected null hypotheses is $10 \cdot 0.05 = 0.5$.

There are stochastic dependencies in the test results as each data value is used in four different tests.If we neglect these dependencies, we assess the risk of falsely rejecting one or more null hypotheses to be ≈ 0.4.

```
> 1-dbinom(0,10,0.05)
[1] 0.4012631
```

To keep the overall error of falsely rejecting a null hypothesis at 5%, we have to apply a **multiple-testing correction** to the p-values.

Bonferroni Method: Multiply each p value with the number n of tests performed and reject only null hypotheses for which this product is below α (usually 0.05).Then, even if all null hypotheses are true, the probability of falsely rejecting one or more of them is $\leq \alpha$.

The Bonferroni method is very *conservative*, i.e. to be on the safe side, it tends to lose some of the significances.

An improvement is the **Bonferroni-Holm Method**: Let n be the number of test. Multiply the smallest p value with n, the second smallest with $n-1$, the third smallest with $n-2$ etc.

The Bonferroni-Holm Method is the default for the R command `pairwise.t.test`:

```
> pairwise.t.test(bam$mean.err,bam$plot.type,paired=TRUE)

Pairwise comparisons using paired t tests

data:  bam$mean.err and bam$plot.type

          bar     bar.stack circ    pie
bar.stack 0.27041 -         -       -
circ      0.00043 0.27041   -       -
pie       0.14592 0.00142   1.3e-05 -
pie3D     0.66542 0.66542   0.10922 0.11929

P value adjustment method: holm
```

A quite different approach of account for multiple testing:

The False Discovery Rate (FDR) and q-values.

References

[1] Y. Benjamini, Y. Hochberg (1995) Controlling the False Discovery Rate: a Practicle and Powerful Approach to Multiple Testing *J. R. Statisti. Soc. B* **57**:289–300

Of m test, S are significant. How many of them are false positives?

	H_0 not rejected	H_0 rejected	$\sum$
H_0 true	$m_0 - F$	F	m_0
H_1 true	$m_1 - T$	T	m_1
$\sum$	$m - S$	S	m

significance level (without multiple-testing adjustment): $\alpha = \Pr_{H_0}(\mathsf{significant}) = \mathbb{E}(F/m_0)$

$$Q = \begin{cases} F/S & \text{if} \quad S > 0 \\ 0 & \text{if} \quad S = 0 \end{cases}$$

False Discovery Rate FDR= $\mathbb{E}(Q) \approx \mathbb{E}(F)/\mathbb{E}(S)$

alternative concept: FDR= $\mathbb{E}(F/S \mid S > 0) \cdot \Pr(S > 0)$

Benjamini-Hochberg FDR

If f is the desired FDR, order hypothesis

$$H_{(1)}, H_{(2)}, \ldots, H_{(m)}$$

according to their respective p-values:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

Then reject $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$, where k is the largest i with

$$p_{(i)} \leq \frac{i}{m} f.$$

Benjamini and Hochberg (1995) proof that: FDR in rejecting $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ is then $\leq f$.

References

[1] J.D. Storey, R. Tibshirani (2003) Statistical significance for genomewide studies *PNAS* **100**: 9440–9445

Storey and Tibshirani (2003) propose calculation of "***q* values**".

If each hypothesis with a q value of e.g. below 0.05 is rejected (which counts as discovery), the FDR is 0.05.

Note that this does **not** imply that a "discovery" with e.g. $q = 0.03$ has a probability of 0.03 to be false.

(Just as the p value of a null hypothesis **not** being the probability of the null hypothesis to be true.)

FDR estimation by Storey and Tibshirani

The defining property of q values:

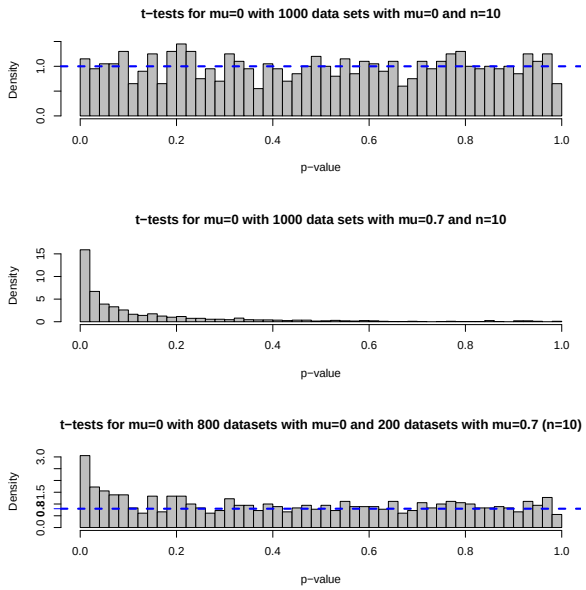
If you reject all null hypotheses with q value below f, your FDR will be f.

Compare to p values:

If you reject a null hypothesis with p value below α, your type-I error probability will be α.

If you reject all null hypothesis with multiple-testing corrected p value below α, your overall type-I error probability will be $\approx \alpha$.

Storey and Tibshirani (2003) estimate FDR and ***q* values** by decomposing distribution density of p values.



Several examples in Storey, Tibshirani (2003), e.g. data by Heldenfalk et al. (2001): search for differentially expressed genes between BRCA1- and BRCA2-mutation-positive tumors.

- 3226 candidate genes
- p-value $< 0.001 \Rightarrow$ 51 genes
- p-value $< 0.0001 \Rightarrow \sim$10 genes
- p-value $< 0.05/3226 \approx 0.0000155$ would correspond to Bonferroni-corrected α of 0.05.
- q-value $< 0.05 \Rightarrow$ 160 genes (expecting ≈ 8 false positives.)
- overall FDR: 2/3, implying that $\approx 1/3$ of the 3226 genes are differentially expressed.

q-values are not multiple-testing adjusted p-values

Assume you test many hypotheses simultaneouly. If you reject all hypothesis, in which your p-value or q-value is smaller than 0.05, then you get if you used...

...p-value (without multiple-testing adjustment): each true null hypothesis is falsely rejected with probability 5%.

...multiple-testing adjusted p-value (e.g. Bonferroni-Holm): even if all null hypotheses are true, the probability to falsely reject one or more of them is only 5%.

. . . q-value : among the rejected null hypothesis, the fraction of true (that is, falsely rejected) null hypotheses is $\lesssim 5\%$

(When reading "BH" in this context, check whether it refers to Bonferroni-Holm or Benjamini-Hochberg)

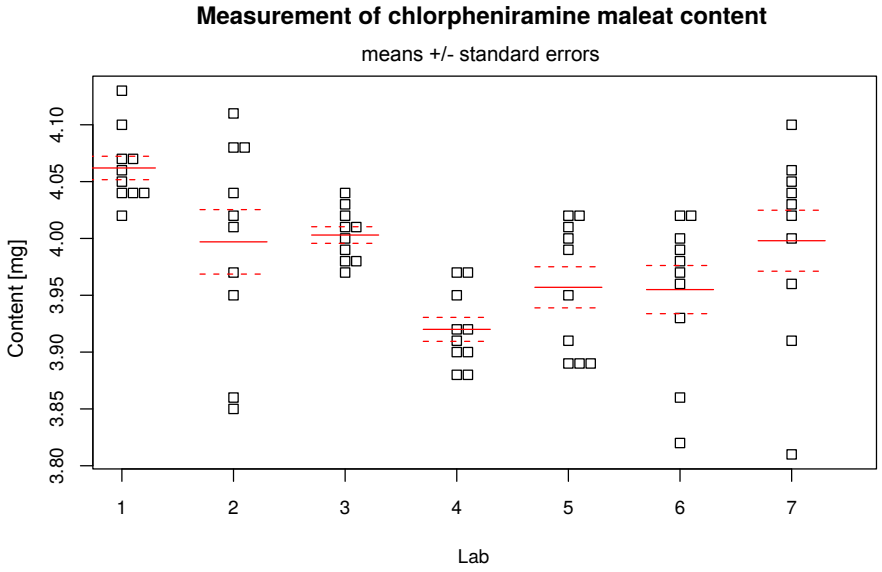
R packages for FDR / q values

qvalue by J.D. Storey: `http://www.bioconductor.org/packages/release/bioc/html/qvalue.html`

fdrtool by B. Klaus and K. Strimmer

References

[1] K. Strimmer (2008) A unified approach to false discovery rate estimation *BMC Bioinformatics* **9**:303



data from Kirchhoefer (1979) *J. Assoc. Offic. Anal. Chem.* 62

Caution: The labs are labeled with numbers. To tell R that these numbers are just names and have no numeric meaning, we have to convert them into a variable of type "factor":

```
> chlor <- read.table("chlorpheniraminmaleat.txt")
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ content: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4.04 ...
 $ Lab    : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
> chlor$Lab <- as.factor(chlor$Lab)
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ content: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4.04 ...
 $ Lab    : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
> attach(chlor)
```

Which labs give significantly different values?

p-values from pairwise comparisons with Welch t-tests:

```
> M <- matrix(NA,ncol=7,nrow=7,
+             dimnames=list(paste("Lab",1:7),paste("Lab",1:7)))
> for (i in 1:6) {
+     for (j in (i+1):7) {
+         M[i,j] <- round(t.test(content[Lab==as.character(i)],
+                                content[Lab==as.character(j)])$p.value,5)
+     }
+ }
> M
      Lab 1   Lab 2   Lab 3   Lab 4   Lab 5   Lab 6   Lab 7
Lab 1    NA 0.05357 0.00025 0.00000 0.00017 0.00055 0.04657
Lab 2    NA      NA 0.84173 0.02654 0.25251 0.25224 0.97985
Lab 3    NA      NA      NA 0.00001 0.03633 0.05532 0.86076
Lab 4    NA      NA      NA      NA 0.09808 0.16280 0.01944
Lab 5    NA      NA      NA      NA      NA 0.94358 0.22336
Lab 6    NA      NA      NA      NA      NA      NA 0.22543
Lab 7    NA      NA      NA      NA      NA      NA      NA
```

We see 21 pair-wise comparisons. On the 5% level, some of them are significant.

Problem of multiple testing: If the null hypothesis ("only random deviations") is true, we falsly reject it with a probability of 5% in each test.If we then apply 20 or more test, there will be on average one or more test in which we falsely reject the null hypothesis. Therefore, we should apply a correction to the p-values for multiple testing.

Bonferroni-Holm corrected p-values

```
> matrix(p.adjust(M,method="holm"),ncol=7,nrow=7,
+        dimnames=list(paste("Lab",1:7),paste("Lab",1:7)))
      Lab 1   Lab 2  Lab 3  Lab 4   Lab 5   Lab 6   Lab 7
Lab 1    NA 0.64284 0.0045 0.0000 0.00323 0.00935 0.60541
Lab 2    NA      NA 1.0000 0.3981 1.00000 1.00000 1.00000
Lab 3    NA      NA     NA 0.0002 0.50862 0.64284 1.00000
Lab 4    NA      NA     NA     NA 0.98080 1.00000 0.31104
Lab 5    NA      NA     NA     NA      NA 1.00000 1.00000
Lab 6    NA      NA     NA     NA      NA      NA 1.00000
Lab 7    NA      NA     NA     NA      NA      NA      NA
```

Again Bonferroni-Holm corrected p-values

```
> pairwise.t.test(content,Lab,paired=FALSE,pool.sd=FALSE)

Pairwise comparisons using t tests with non-pooled SD

data:  meas and flab

  1       2       3       4       5       6
2 0.64289 -       -       -       -       -
3 0.00446 1.00000 -       -       -       -
4 3.3e-07 0.39816 0.00015 -       -       -
5 0.00319 1.00000 0.50861 0.98082 -       -
6 0.00940 1.00000 0.64289 1.00000 1.00000 -
7 0.60542 1.00000 1.00000 0.31104 1.00000 1.00000

P value adjustment method: holm
```

Bonferroni corrected p-values

```
> matrix(p.adjust(M,method="bonferroni"),ncol=7,nrow=7,
+        dimnames=list(paste("Lab",1:7),paste("Lab",1:7)))
      Lab 1 Lab 2   Lab 3   Lab 4   Lab 5   Lab 6   Lab 7
Lab 1    NA     1 0.00525 0.00000 0.00357 0.01155 0.97797
Lab 2    NA    NA 1.00000 0.55734 1.00000 1.00000 1.00000
Lab 3    NA    NA      NA 0.00021 0.76293 1.00000 1.00000
Lab 4    NA    NA      NA      NA 1.00000 1.00000 0.40824
Lab 5    NA    NA      NA      NA      NA 1.00000 1.00000
Lab 6    NA    NA      NA      NA      NA      NA 1.00000
Lab 7    NA    NA      NA      NA      NA      NA      NA
```

q-values

```
> matrix(p.adjust(M,method="fdr"),ncol=7,nrow=7,
+        dimnames=list(paste("Lab",1:7),paste("Lab",1:7)))
      Lab 1     Lab 2     Lab 3    Lab 4      Lab 5     Lab 6     Lab 7
Lab 1    NA 0.1056109 0.0013125 0.000000 0.00119000 0.0023100 0.1056109
Lab 2    NA        NA 0.9513663 0.079620 0.31192412 0.3119241 0.9798500
Lab 3    NA        NA        NA 0.000105 0.09536625 0.1056109 0.9513663
Lab 4    NA        NA        NA       NA 0.17164000 0.2629846 0.0680400
Lab 5    NA        NA        NA       NA         NA 0.9798500 0.3119241
Lab 6    NA        NA        NA       NA         NA        NA 0.3119241
Lab 7    NA        NA        NA       NA         NA        NA        NA
```

One possibility in the anova case: Tukey's Honest Significant Differences (HSD).

For this, we first need to fit an anova model:

```
> chlor.aov <- aov(content~Lab)
```

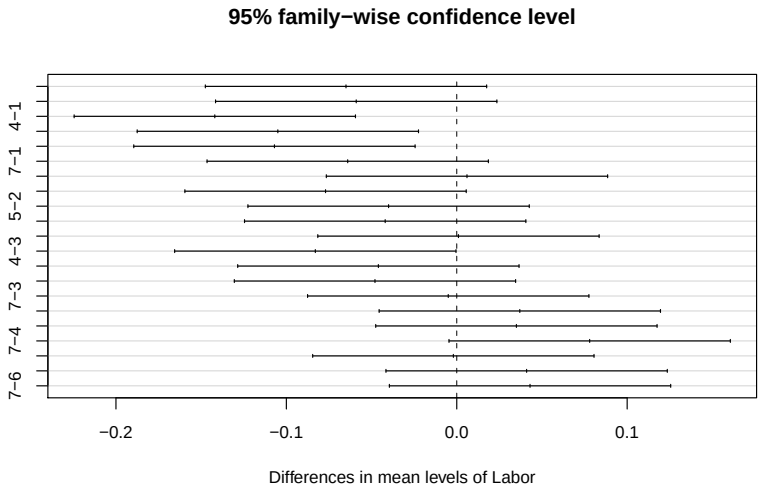
```
> TukeyHSD(chlor.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = content ~ Lab, data = chlor)

$Lab
      diff          lwr          upr     p adj
2-1 -0.065 -0.147546752  0.017546752 0.2165897
3-1 -0.059 -0.141546752  0.023546752 0.3226101
4-1 -0.142 -0.224546752 -0.059453248 0.0000396
5-1 -0.105 -0.187546752 -0.022453248 0.0045796
6-1 -0.107 -0.189546752 -0.024453248 0.0036211
7-1 -0.064 -0.146546752  0.018546752 0.2323813
3-2  0.006 -0.076546752  0.088546752 0.9999894
4-2 -0.077 -0.159546752  0.005546752 0.0830664
5-2 -0.040 -0.122546752  0.042546752 0.7578129
6-2 -0.042 -0.124546752  0.040546752 0.7140108
7-2  0.001 -0.081546752  0.083546752 1.0000000
4-3 -0.083 -0.165546752 -0.000453248 0.0478900
5-3 -0.046 -0.128546752  0.036546752 0.6204148
6-3 -0.048 -0.130546752  0.034546752 0.5720976
7-3 -0.005 -0.087546752  0.077546752 0.9999964
5-4  0.037 -0.045546752  0.119546752 0.8178759
6-4  0.035 -0.047546752  0.117546752 0.8533629
7-4  0.078 -0.004546752  0.160546752 0.0760155
6-5 -0.002 -0.084546752  0.080546752 1.0000000
7-5  0.041 -0.041546752  0.123546752 0.7362355
7-6  0.043 -0.039546752  0.125546752 0.6912252
```

We get confidence intervals [lwr,upr] for the differences between the labs and p-values for the null hypotheses that these differences are 0. Both are already corrected for multiple testig.

Visualization of HSD confidence intervals with `plot(TukeyHSD(chlor.aov))`:



Restriction: Tukeys HSD-Methode is only valid for *balanced designs*, i.e. the sample sizes must be equal across the groups. (And variances must be equal, which is already required in the anova.)

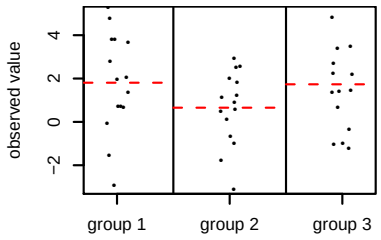
This is the case for the lab comparisons but – due to missing data – not for the blue area estimates.

2 ANOVA and F-Test

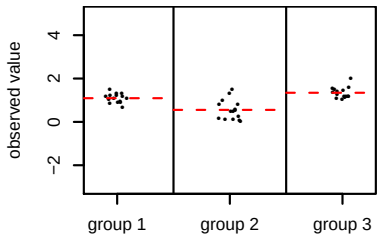
With the ANalysis Of VAriance (ANOVA) we test whether there are differences at all among the Labs:

```
> chlor.aov <- aov(content~Lab,data=chlor)
> summary(chlor.aov)
            Df  Sum Sq  Mean Sq F value    Pr(>F)
Lab          6 0.12474 0.020789  5.6601 9.453e-05 ***
Residuals 63 0.23140 0.003673
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Are the group means significantly different?



[-0.8cm] Large variability within groups



[-0.8cm] Small variability with groups

Or does this look like random deviations?

This depends on the ratio of the varibility of group means and the variability within the groups. The analysis of variance (ANOVA) quantifies this ratio and its significance.

Example

Blood-clotting time in rats under 4 different treatments

group	observation							
1	62	60	63	59				
2	63	67	71	64	65	66		
3	68	66	71	67	68	68		
4	56	62	60	61	63	64	63	59

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

Caution: In this simple example, the global mean is the mean of the group means. This is not always the case!

Example

Blood-clotting times in rats under 4 different treatments

gr.	$\overline{x}_{i\cdot}$	observations							
1	61	62	60	63	59				
		$(62-61)^2$	$(60-61)^2$	$(63-61)^2$	$(59-61)^2$				
2	66	63	67	71	64	65	66		
		$(63-66)^2$	$(67-66)^2$	$(71-66)^2$	$(64-66)^2$	$(65-66)^2$	$(66-66)^2$		
3	68	68	66	71	67	68	68		
		$(68-68)^2$	$(66-68)^2$	$(71-68)^2$	$(67-68)^2$	$(68-68)^2$	$(68-68)^2$		
4	61	56	62	60	61	63	64	63	59
		$(56-61)^2$	$(62-61)^2$	$(60-61)^2$	$(61-61)^2$	$(63-61)^2$	$(64-61)^2$	$(63-61)^2$	$(59-61)^2$

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

The *red* Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups: $ss_{\text{within}} = 112$, 20 degrees of freedom (df)

Sums of squares between groups: $ss_{\text{betw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$, 3 degrees of freedom (df)

$$F = \frac{ss_{\text{betw}}/3}{ss_{\text{within}}/20} = \frac{76}{5.6} = 13.57$$

Example: Blood-clotting times in rats under 4 different treatments.

ANOVA table (ANalysis Of VAriance)

	df	sum of squares (ss)	mean sum of squares (ss/df)	F value
groups	3	228	76	13.57
residuals	20	112	5.6	

Under the Null-Hypothesis H_0 "the group means are equal"
(and assuming independent, normally distributed observations)
is F Fisher-distributed with 3 and 20 degrees of freedom, and $p = \text{Fisher}_{3,20}([13.57, \infty)) \leq 5 \cdot 10^{-5}$.

Thus, we can reject H_0.


Sir Ronald Aylmer Fisher, 1890–1962

F-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in I groups,
$X_{ij} = j$-th observation in i-th group, $j = 1, \ldots, n_i$.
Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$, with independent $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.
($\mu_i = \mathbb{E}X_{ij}$ is the "true" mean within group i.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij}$ observed global mean
$\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ observed mean in group i

$$SS_{\text{within}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$$ sum of squares within the groups, $n - I$ degrees of freedom

$$SS_{\text{betw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$ sum of squares between the groups, $I - 1$ degrees of freedom

$$F = \frac{SS_{\text{betw}}/(I-1)}{SS_{\text{within}}/(n-I)}$$

F-Test

$X_{ij} = j$-th observation i-th group, $j = 1, \ldots, n_i$,
Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$, with independent $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

$$SS_{\text{within}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$$ sum of squares within groups, $n - I$ degrees of feedom

$$SS_{\text{betw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$ sum of squares between groups, $I - 1$ degrees of freedom

$$F = \frac{SS_{\text{betw}}/(I-1)}{SS_{\text{within}}/(n-I)}$$

Under the hypothesis $H_0 : \mu_1 = \cdots = \mu_I$ ("all μ_i are equal") F is Fisher-distributed with $I - 1$ and $n - I$ degrees of freedom
(no matter what the true joint value of μ_i is).

F-Test: We reject H_0 on the level of significance α if $F \geq q_\alpha$, whereas q_α is the $(1-\alpha)$-quantile of the Fisher-distribution with $I-1$ and $n-I$ degrees of freedom.

Computing the p-value with R

How to choose q such that $\Pr(F \leq q) = 0.95$ for Fisher(6,63)-distributed F?

```
> qf(0.95,df1=6,df2=63)
[1] 2.246408
```

computation of p-value: How probable is it that a Fisher(3,20)-distributed random variable takes a value ≥ 13.57?

```
> pf(13.57, df1=3, df2=20, lower.tail=FALSE)
[1] 4.66169e-05
```

Table of 95%-quantiles of the F distribution

This table shows the (rounded) 95%-quantiles of the Fisher distribution with k_1 and k_2 degrees of freedom (k_1: numerator, k_2: denominator)

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.7
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.6
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.1
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.57
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.34
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.31

Anova completely in R

The text file clotting.txt contains a column "time" with the clotting times and a column "treat" with the treatmetns A,B,C,D.

```
> rat<-read.table("clotting.txt",header=TRUE)
> rat.aov <- aov(time~treat,data=rat)
> summary(rat.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
treat        3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Some of what you should be able to explain

- Multiple Testing
  - What is the problem
  - Bonferroni and Bonferroni-Holm
  - FDRs and q values
  - Tukey's HSDs
- Model assumptions of anova/F-test
- How to calculate F statistic and degrees of freedom
- Carry out anova with R (make sure to have factor)

3 Anova with more than one factor

Hypothetical study: 100 LMU students were selected to participate in a 10km footrace.

The aim of the study was to assess how sportiveness dependeds on gender and smoking behavior. For this, the students were subdivided into four groups:

	male	female	$\sum$
smoker	18	9	27
non-smoker	30	43	73
$\sum$	48	52	100

(Smoking seems to be gender-specific, $p = 0.026$, Fisher's exact test)

```
> t.test(runtime[smoking=="s"],runtime[smoking=="n"])

Welch Two Sample t-test

data:  runtime[smoking == "s"] and runtime[smoking == "n"]
t = 0.1102, df = 60.611, p-value = 0.9126
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.522165  8.399714
sample estimates:
mean of x mean of y
 91.06888  90.63010
```

We need to correct for the gender effect when we test for the effect of smoking.

We do this by fitting a model in which both gender $g_i \in \{m, f\}$ and smoking behavior $s_i \in \{y, n\}$ of person i have an effects b_{g_i} ($= b_m$ or b_f) and c_{s_i} ($= c_n$ or c_y) on the run time r_i of i:

$$r_i = a + b_{g_i} + c_{s_i} + \varepsilon_i,$$

with $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.To avoid over-parametrization of the model we need some restriction on the parameter values, and the default one is that one case of each factor is used as a base line, e.g. $b_f = 0$ and $c_n = 0$.

This means that the value of

a is the average run time of non-smoking females,

$a + b_m$ is the average r_i of non-smoking males,

$a + c_s$ is the average r_i of smoking females, and

$a + b_m + c_s$ is the average r_i of smoking males.

Note in $a + b_m + c_s$ that the effects are assumed to be additive!

R fits the parameters a, b_m and c_s to the data:

```
> lm(runtime~sex+smoking)

Call:
lm(formula = runtime ~ sex + smoking)

Coefficients:
(Intercept)      sexmale     smokings
    102.134      -27.993        7.597
```

The estimated values are:

$$\widehat{a} \approx 102.1 \qquad \widehat{b}_m \approx -28 \qquad \widehat{c}_s \approx 7.6$$

We can now compare the model

$$r_i = a + b_{g_i} + c_{s_i} + \varepsilon_i,$$

to a model

$$r_i = a + b_{g_i} + \varepsilon_i,$$

in which smoking has no effect.

The models can be specified in R commands like `aov` and `lm` by the following R formula code:

`runtime ~ sex + smoking`

and

`runtime ~ sex`

Now we can test for the effect of smoking, while accounting for the gender effect:

```
> anova(lm(runtime~sex),lm(runtime~sex+smoking))
Analysis of Variance Table

Model 1: runtime ~ sex
Model 2: runtime ~ sex + smoking
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     98 26776
2     97 23691  1    3084.3 12.628 0.0005889 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The result of this anove is that the model accounting for gender and smoking behavior explains the data significantly better ($p < 0.0006$) than the model with gender alone.

If we want to test both factors, we can do it as follows:

```
> drop1(lm(runtime~smoking+sex),test="F")
Single term deletions

Model:
runtime ~ smoking + sex
        Df Sum of Sq   RSS    AIC F value     Pr(F)
<none>               20570 538.64
smoking  1    1078.7 21648 541.75   5.087   0.02635 *
sex      1   18548.6 39118 600.92  87.469 3.356e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note that the p-values are not multiple-testing corrected!

With $a + b_m + c_s$ being the average r_i of smoking males, we see that a model assumption is that the effects of gender and smoking add up.

If we do not want to assume this additivity, we can add an interaction effect $d_{\cdot,\cdot}$:

$$r_i = a + b_{g_i} + c_{s_i} + d_{g_i,s_i} + \varepsilon_i,$$

In R the formula code for this model is

```
runtime ~ sex + smoking + sex:smoking
```

or, for short:

```
runtime ~ sex * smoking
```

```
> lm(runtime~sex*smoking)

Call:
lm(formula = runtime ~ sex * smoking)

Coefficients:
     (Intercept)          sexmale          smokings  sexmale:smokings
        103.2930          -30.8130            0.9002           11.1265
```

$$\widehat{a} \approx 103.3 \qquad \widehat{b}_m \approx -30.8 \qquad \widehat{c}_s \approx 0.9 \qquad \widehat{d}_{m,s} \approx 11.13$$

We test whether the model with interaction term fits the data better than the additive model:

```
> anova(lm(runtime~sex+smoking),lm(runtime~sex*smoking))
Analysis of Variance Table

Model 1: runtime ~ sex + smoking
Model 2: runtime ~ sex * smoking
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1     97 23691
2     96 23691  1  0.020566 1e-04 0.9927
```

Obviously, the model with interaction term does not fit the data signifcantly better.

Indicator variables

In model equations like

$$r_i = a + b_{g_i} + c_{s_i} + d_{g_i,s_i} + \varepsilon_i,$$

it can by handy to use *indicator variables* I_A for events A, with

$$I_A = \begin{cases} 1 & \text{if } A \text{ is fulfilled} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we can write the model above as

$$r_i = a + b_m \cdot I_{g_i=m} + c_s \cdot I_{s_i=s} + d_{m,s} \cdot I_{g_i=m} \cdot I_{s_i=s} + \varepsilon_i.$$

Note that $I_{g_i=m} \cdot I_{s_i=s} = I_{\{g_i=m\} \cap \{s_i=s\}}$.

balanced design

In another (hypothetical) survey, a balanced design was used, that is, equal numbers of students were selected for the four groups:

	male	female	$\sum$
smoker	25	25	50
non-smoker	25	25	50
$\sum$	50	50	100

Balanced design, but no representative sampling!

```
> drop1(lm(runtime~smoking+sex),test="F")
Single term deletions

Model:
runtime ~ smoking + sex
        Df Sum of Sq   RSS    AIC F value     Pr(F)
<none>               23691 552.77
smoking  1    3084.3 26776 563.01  12.628 0.0005889 ***
sex      1   10648.1 34339 587.89  43.597 2.158e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> t.test(runtime[smoking=="s"],runtime[smoking=="n"])

Welch Two Sample t-test

data:  runtime[smoking == "s"] and runtime[smoking == "n"]
t = 2.9669, df = 94.736, p-value = 0.003808
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.674649 18.539956
sample estimates:
mean of x mean of y
 101.1723   90.0650
```

Note that the linear model commands

```
summary(lm(runtime~smoking+sex))
```

and

```
drop1(lm(runtime~smoking+sex),test="F")
```

are neither restricted to represenative sampling nor to balanced design.

But how to interpret the group means?

Representative sampling:

```
> mean(runtime[sex=="male"])
[1] 76.99001
> mean(runtime[sex=="female"])
[1] 103.4488
> mean(runtime[smoking=="s"])
[1] 91.06888
> mean(runtime[smoking=="n"])
[1] 90.6301
```

Balanced design:

```
> mean(runtime[sex=="male"])
[1] 85.29967
> mean(runtime[sex=="female"])
[1] 105.9376
> mean(runtime[smoking=="s"])
[1] 101.1723
> mean(runtime[smoking=="n"])
[1] 90.065
```

In the balanced design, smokers are overrepresented (compared to reality), and females are overrepresented among the smokers and underrepresented among the non-smokers.

Let i be the index for the row of a data table. The data are subdivided into groups and G_i is the group row i (or patient i) belongs to; e.g. G_i can be the treatment of patient i. Let Y_i be the response variable, e.g. the blood pressure of patient i. We can apply an anova to check whether Y depends on G, and the model behind it is:

$$Y_i = b_{G_i} + \varepsilon_i$$

with independent $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (note: same σ^2 for all). During the ANOVA we estimate the influence b_{G_i} of the group on Y_i by the group mean $\widehat{b_g}$. Thus, the residuals $r_i := Y_i - \widehat{b_{G_i}} \approx Y_i - b_{G_i} = \varepsilon_i$ should be approximately normally distributed.

More than one factor can play a role. For example we may take into account that the blood pressure Y_i of a patient may depend on the sex S_i of the patient. In this case the model behind the anova takes the form

$$Y_i = b_{G_i} + c_{S_i} + \varepsilon_i.$$

b_{G_i} depends only on the treatment group and c_{S_i} only on the sex of the person. If we also want allow in *interaction* between the treatment and the sex, we need another variable d_{G_i,S_i} that may depend on both:

$$Y_i = b_{G_i} + c_{S_i} + d_{G_i,S_i} + \varepsilon_i.$$

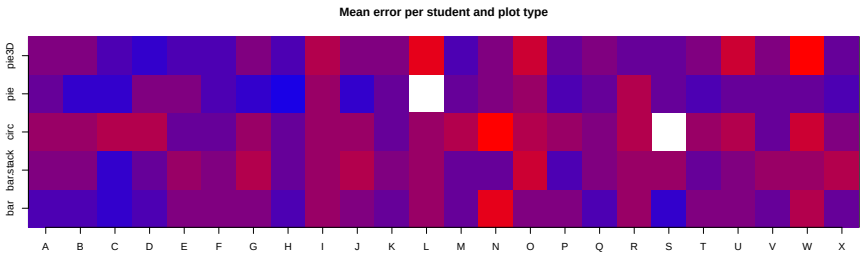
This makes possible, for example, that a certain treatment has a stronger effect for males than for females.

A *balanced design* means, that the sample size are the same for each combination of factors. E.g. 10 males and 10 females in each treatment group. Some ANOVA-based method will only work for balanced designs. Therefore, it is preferable to use a balanced design when planning an experiment. If the data, however, are observations from nature, the "design" is usually unbalanced and this has to be taken into account in the analysis.

One of the methods for which you need a balanced design is Tukey's HSD.

Note that even if the design was balanced, the balance can be destroyed if there are missing data.

Average errors of each student for each plot type



blue: small error,
red: large error,
white: missing data

```
> str(bam)
'data.frame': 120 obs. of  3 variables:
 $ plot.type: Factor w/ 5 levels "bar","bar.stack",..: 1 2 3 4 5 1 2 3 4 5 ...
 $ student  : Factor w/ 24 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ mean.err : num  1.07 1.39 1.68 1.35 1.42 ...
> (av <- aov(mean.err~plot.type,bam))
Call:
   aov(formula = mean.err ~ plot.type, data = bam)

Terms:
                plot.type Residuals
Sum of Squares   5.472053 21.382323
Deg. of Freedom         4       113

Residual standard error: 0.434999
Estimated effects may be unbalanced
2 observations deleted due to missingness

> summary(av)
             Df Sum Sq Mean Sq F value   Pr(>F)
plot.type     4  5.472  1.3680    7.23 3.21e-05 ***
Residuals   113 21.382  0.1892
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
2 observations deleted due to missingness
```

R code for anova withouth balanced design assumption:

```
> mod <- lm(mean.err~plot.type,bam)
> anova(mod)
Analysis of Variance Table

Response: mean.err
            Df  Sum Sq Mean Sq F value    Pr(>F)
plot.type    4  5.4721 1.36801  7.2296 3.207e-05 ***
Residuals  113 21.3823 0.18922
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Here, we get basically the same results as with aov, but this may be different if we have more than one factor, as we will see.

4 Mixed effects, that is, fixed effects and random effects

In the anova for the mean errors per student and visualization method, we have neglected that several values came from the same student.

Different values from the same student are in different groups (plot types), such that among-student variation will reduce between-group differences. Thus, our test is conservative, that is, the risk of false significance is reduced, and we are in a sense on the safe side.

But we can also explicitly allow for student effects, which allows us to use the full data instead of mean values per student-plot.type combination.

```
> mod <- lm(err~typ+student,ba)
> drop1(mod,test="F")
Single term deletions

Model:
err ~ typ + student
        Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>               786.09 -132.269
typ      4    39.973 826.07  -92.802 11.8100 2.317e-09 ***
student 23    81.832 867.93  -83.497  4.2047 2.193e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(mod)

Call:
lm(formula = err ~ typ + student, data = ba)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1437 -0.6859  0.0333  0.6128  3.3657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.280572   0.157108   8.151 1.16e-15 ***
typbar.stack  0.205367   0.093884   2.187  0.02896 *
typcirc       0.412197   0.094010   4.385 1.29e-05 ***
typpie       -0.201979   0.094144  -2.145  0.03218 *
typpie3D      0.098491   0.093884   1.049  0.29442
studentB     -0.069044   0.205690  -0.336  0.73719
studentC     -0.312302   0.205690  -1.518  0.12928
studentD     -0.059899   0.205690  -0.291  0.77096
studentE      0.007232   0.205690   0.035  0.97196
studentF     -0.132939   0.205690  -0.646  0.51824
studentG      0.094442   0.205690   0.459  0.64624
studentH     -0.371966   0.205690  -1.808  0.07087 .
studentI      0.359649   0.205690   1.748  0.08071 .
studentJ      0.094442   0.205690   0.459  0.64624
[...]
studentW      0.657800   0.205690   3.198  0.00143 **
studentX      0.054703   0.205690   0.266  0.79034
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9199 on 929 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1349,Adjusted R-squared:  0.1098
F-statistic: 5.366 on 27 and 929 DF,  p-value: < 2.2e-16
```

Model so far for error y_i in line $i = 1, 2, \ldots, 960$ of the data table:

$$y_i = a + b_{t_i} + c_{s_i} + \varepsilon_i$$

With "intercept" a

(estimating the average error of student A for bar plots),

plot-type effects $b_{\text{bar}}, b_{\text{circ}}, \ldots, b_{\text{pie}}$

(with $b_{\text{bar}} = 0$ and all others compared to plot type bar),

student effects $c_A, c_B, \ldots, c_X$

(with $c_A = 0$ and all others compared to student A) and

independent error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{960} \sim \mathcal{N}(0, \sigma^2)$.

Instead of estimating the 23 effect $c_B, \ldots, c_X$ of each student on the estimation error, we can use an alternative approach for which only the among-student variance is estimated.

This, however, requires the additional assumption that student-effects are normally distributed with some variance σ_s^2 that we have to estimate.

Thus we reduce the number of parameters to be estimated by 22.

We model the student effect as random effect and the plot-type effect as fixed (that is, non-random) effect. Thus, we obtain a **mixed-effects model**.

```
> library(lme4)
> mmod <- lmer(err~typ + (1 | student),ba)
> summary(mmod)
Linear mixed model fit by REML ['lmerMod']
Formula: err ~ typ + (1 | student)
   Data: ba

REML criterion at convergence: 2602

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1899 -0.7482  0.0303  0.6548  3.6248

Random effects:
 Groups   Name        Variance Std.Dev.
 student  (Intercept) 0.06817  0.2611
 Residual             0.84619  0.9199
Number of obs: 957, groups:  student, 24

Fixed effects:
            Estimate Std. Error t value
(Intercept)   1.39186    0.08513  16.349
typbar.stack  0.20537    0.09389   2.187
typcirc       0.41259    0.09401   4.389
typpie       -0.20321    0.09414  -2.159
typpie3D      0.09849    0.09389   1.049

Correlation of Fixed Effects:
            (Intr) typbr. typcrc typpie
typbar.stck -0.551
typcirc     -0.551  0.499
typpie      -0.550  0.499  0.498
typpie3D    -0.551  0.500  0.499  0.499
```

```
> drop1(mmod,test="Chisq")
Single term deletions

Model:
err ~ typ + (1 | student)
       Df    AIC    LRT   Pr(Chi)
<none>    2599.5
typ     4 2638.0 46.511 1.928e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

A test available in an additional package is supposed to be more accurate if number of groups is not very large:

```
> library(pbkrtest)
> KRmodcomp(mmod, lmer(err ~ (1 | student),ba) )
F-test with Kenward-Roger approximation; computing time: 0.33 sec.
large : err ~ typ + (1 | student)
small : err ~ (1 | student)
         stat     ndf     ddf F.scaling  p.value
Ftest  11.866   4.000 929.060         1 2.09e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From the same package: Test based on parametric bootstrap (takes longer but might be the most accurate approach)

```
> PBmodcomp(update(mmod,subset=!is.na(err),REML=FALSE),
+           lmer(err ~ (1 | student),ba,subset=!is.na(err),
+                REML=FALSE))
Parametric bootstrap test; time: 15.56 sec;
samples: 1000 extremes: 0;
large : err ~ typ + (1 | student)
small : err ~ (1 | student)
         stat df   p.value
LRT    46.511  4 1.928e-09 ***
PBtest 46.511     0.000999 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

5 Type I and Type II ANOVA

Be careful with the interpretation of ANOVA tables! The R command anova, applied to a single model gives a so-called "Type I Anova", where each line take only the variables in the lines above into account. Example: Chill coma recovery times measured by different persons on different days for different fly lines.

```
> anova(model4)
Analysis of Variance Table

Response: log(ccrt)
           Df  Sum Sq Mean Sq F value    Pr(>F)
line        1  1.2224 1.22238 13.1486 0.0003812 ***
day        11  2.8471 0.25883  2.7841 0.0023769 **
person      1  0.0850 0.08504  0.9147 0.3402393
[...]
```

For example, the p-value 0.0023769 tells how much better the model with line and day can explain the data compared to a model that only takes line into account. Thus, the values assigned to variables depend on the input order.

If you use the R command drop1 with the option test="F", you get a so-called "Type II Anova", in which each line shows the influence of one variable, given the estimates of *all* other variables.

```
> drop1(model4,test="F")
[...]
      Df Sum of Sq    RSS     AIC F value    Pr(F)
<none>              15.618 -418.91
line    1   0.05860 15.677 -420.23  0.6304 0.428338
day    11   2.47080 18.089 -414.18  2.4161 0.008177 **
person  1   0.08504 15.703 -419.92  0.9147 0.340239
```

For example, the p-value 0.008177 says that a model that takes line, day and person into account explains the data significantly better than a model that uses only line and person.

Back to the footrace example with non-balanced design:

```
> summary(aov(runtime~sex+smoking))
            Df  Sum Sq Mean Sq F value    Pr(>F)
sex          1 17473.7 17473.7  82.400 1.316e-14 ***
smoking      1  1078.7  1078.7   5.087   0.02635 *
Residuals   97 20569.7   212.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(aov(runtime~smoking+sex))
            Df  Sum Sq Mean Sq F value    Pr(>F)
smoking      1     3.8     3.8  0.0179    0.8939
sex          1 18548.6 18548.6 87.4693 3.356e-15 ***
Residuals   97 20569.7   212.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

But for the dataset with *balanced design* (for which aov is more appropriate) the input order does not matter even for Type I anova:

```
> summary(aov(runtime~sex+smoking))
            Df  Sum Sq Mean Sq F value    Pr(>F)
sex          1 10648.1 10648.1  43.597 2.158e-09 ***
smoking      1  3084.3  3084.3  12.628 0.0005889 ***
Residuals   97 23691.2   244.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(aov(runtime~smoking+sex))
```

```
            Df  Sum Sq Mean Sq F value    Pr(>F)
smoking      1  3084.3  3084.3  12.628 0.0005889 ***
sex          1 10648.1 10648.1  43.597 2.158e-09 ***
Residuals   97 23691.2   244.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Some of what you should be able to explain

- balanced design: pros and cons
- Type I vs. type II anova
- anova with mixed effects
  - model assumptions (in precise mathematical terms)
  - when to apply it
  - how to assess significance
- Things to explain listed on page 14.

6 Non-parametric: The Kruskal-Wallis Test

The one-factor anova is based on the assumption that all values are independent and normally distributed. The group means $\mu_1, \mu_2, \ldots, \mu_m$ may be different. (It is the aim of the test to find this out.) The variances within the groups must be equal.

In formulae: If Y_{ij} is the jth value in group i, it is assumed that

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

whereas all ε_{ij} are independently $\mathcal{N}(0, \sigma^2)$ distributed with the same σ^2 for all groups!

The null hypothesis is $\mu_1 = \mu_2 = \cdots = \mu_m$.
Not all deviations from normality cause problems.

However, anovas are not robust against outliers or distributions that may sometimes produce exceptionally huge values.

In such cases we can apply the **Kruskal-Wallis Test**. Like the Wilcoxon-Test it uses the *Ranks* instead of the actual values. It is a *non-parameteric test*, no particular probability distribution is assumed.

Null hypothesis of the Kruskal-Wallis test: all values of Y_{ij} come from the same distribution, independent of their group.

Like in the anova, we also have to assume for the Kruskal-Wallis test that the data are independent of each other.

- Let R_{ij} be the rank of Y_{ij} within the total sample.
- Let
$$\overline{R}_{i.} = \frac{1}{J_i} \sum_{j=1}^{J_i} R_{ij}$$
be the average rank in group i, with J_i the sample size in group i.
- The mean rank in the total sample is
$$\overline{R}_{..} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2},$$
with I being the number of groups and N being the total sample size.
- Under the null hypothesis all ranks have expectation value $R_{..}$.
- We measure the deviation from this expectation by
$$S = \sum_{i=1}^{I} J_i \cdot (\overline{R}_{i.} - \overline{R}_{..})^2.$$
- To get a p value from S we have to know the probability distribution of S under the null hypothesis. For small values of I and J_I, the latter can be found in tables.
- For $I \geq 3$ und $J_i \geq 5$, or $I > 3$ and $J_i \geq 4$ we can use that the following scaling K of S is approximately χ^2 distributed with $I - 1$ degrees of freedom:
$$K \;=\; \frac{12}{N \cdot (N+1)} S \;=\; \frac{12}{N \cdot (N+1)} \cdot \left(\sum_{i=1}^{I} J_i \cdot \overline{R}_{i.}^2 \right) - 3 \cdot (N+1)$$

Kruskal-Wallis-Test with R

```
> kruskal.test(time~treat,data=rat)

Kruskal-Wallis rank sum test

data:  time by treat
Kruskal-Wallis chi-squared = 17.0154, df = 3,
                                p-value = 0.0007016
```

```
> kruskal.test(content~lab,data=chlor)

Kruskal-Wallis rank sum test

data:  content by lab
Kruskal-Wallis chi-squared = 29.606, df = 6,
                                p-value = 4.67e-05
```

Some of what you should be able to explain...

...about the Kruskal-Wallis test

- test statistic
- rescaled test statistic and χ^2 approximation