Biostatistics week 12

- Linear regression and ANOVA analysis
- Linear regression with paired data
- Non-parametric tests for group comparison with >2 groups
- Questions and answer hour

Biostatistics looking back: any questions?

Tuesday, 11^{th} December, 10-11am, HG E 3
Exam is on these topics, MC questions, 60 minutes

Topics

- data visualization
- basic terms and summary statistics
- study types, confounding
- diagnostic tests
- models/distribution-types
- parameter estimation
- testing, confidence intervals, p-values
- linear regression
- reliability analysis
- outlook on more advanced or modern regression methods

Steps in linear modelling

0) Preprocessing

- learning the meaning of all variables, check for correlations
- give short and informative names
- check for impossible values, errors
- if they exist (missing, error): set them to NA
- be very careful with imputation methods, are missings systematic?

1) First-aid transformations

- bring all variables to a suitable scale (use also field knowledge)
- routinely apply the first-aid transformations

2) Find a good model

- start with a model including important confounders
- perform a residual analysis
- improve model by transformations or adding better predictors
- reduce step by step complexity and use anova for comparison
- use your specific knowledge to choose between variables

Limits of linear Regression

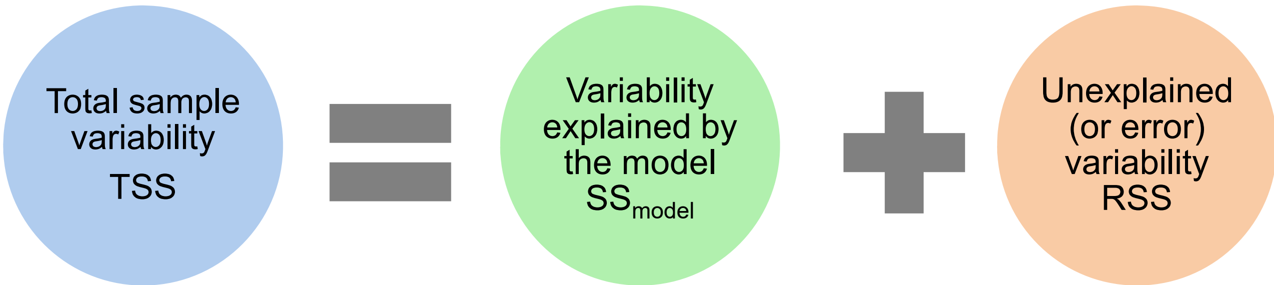
If your residuals do not follow a Normal distribution (even after transformations) use generalized linear modeling (glm – e.g. logisitic regression)

If your predictors show a strong correlation use shrinkage methods (e.g. lasso)

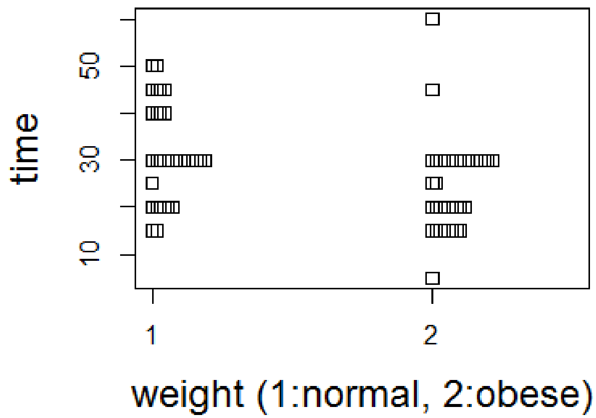
If your data are not independent use mixed models or methods for time-series.

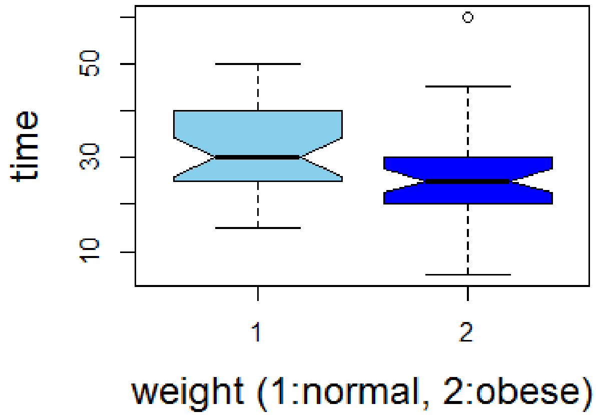
If you do not have a linear relation, use non-linear regression (e.g. nlm) or generalizes additive models (e.g. gam) or tree models

ANalysis Of Variance (ANOVA)
= linear regression with factor variables

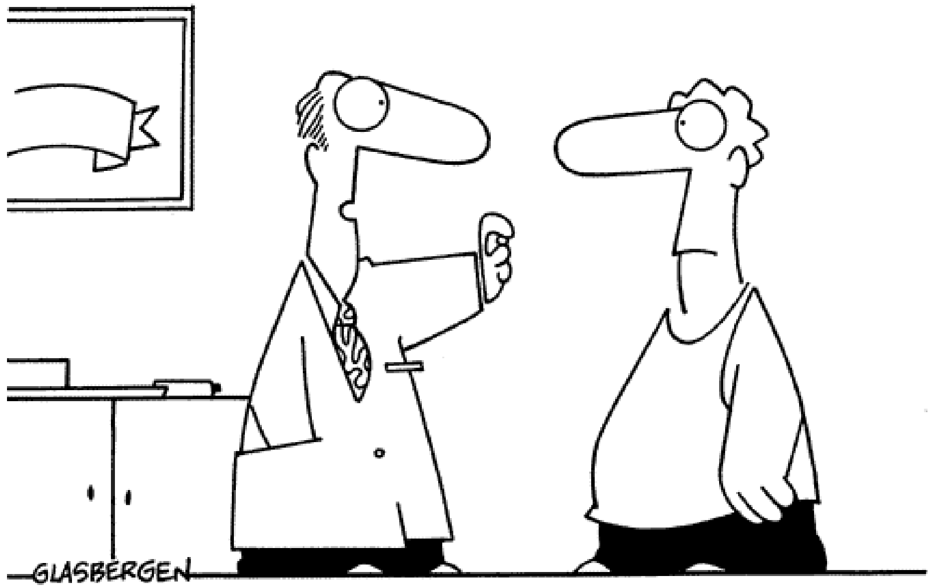


Example with one factorial predictor
Do medical doctors spend less time with obese patients?





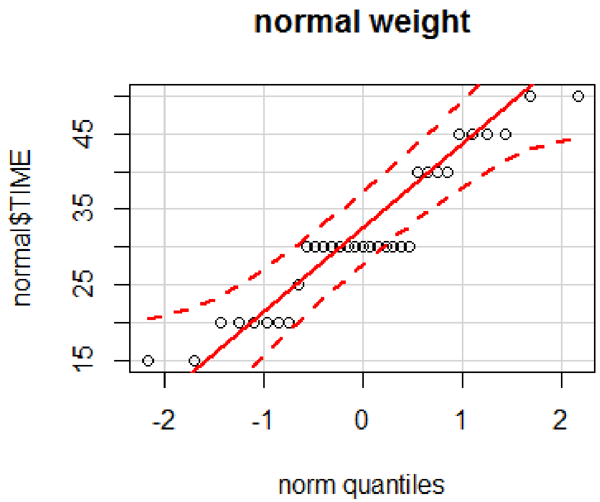
In an observational study it was measured how much time doctors spend with a patient.



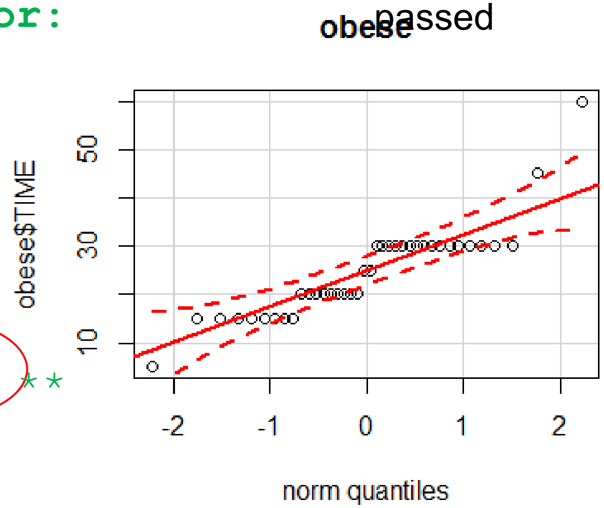
"To prevent a heart attack, take one aspirin every day. Take it out for a jog, then take it to the gym, then take it for a bike ride...."

Do medical doctors spend less time with obese patients? How can we test this with linear regression and ANOVA?

```
t.test(TIME~WEIGHT, data=dat)
# t = 2.9, df = 67, p-value = 0.0057
# alternative hypothesis: true difference in
# means is not equal to 0
# 95 percent confidence interval:
#   2   11
# sample estimates:
#   mean of x    mean of y
#     31            25

# do it by regression with one factorial predictor:

fit=lm(TIME~WEIGHT, data=dat)

anova(fit)
# get anova-table from lm-object
# Response: TIME
#              Df    Sum    Sq Mean   F value   Pr(>F)
# WEIGHT        1     776     776     8.16     0.0057 **
# Residuals    69    6561      95
```



Normality check passed



An ANOVA with 1 factor with 2 levels is equivalent to a two-sample t-test.

How to test for an effect between >2 groups? Applying 1-way ANOWA with >2 levels

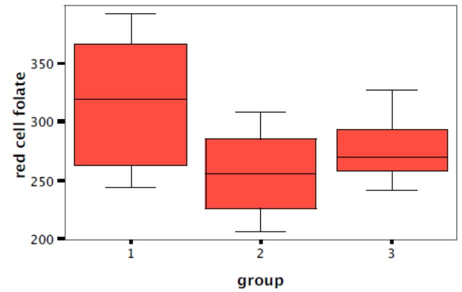
Here, we want to investigate, if three different treatments result in different levels of the output: folate in red blood cells

We can apply a regression with the group factor as predictor to investigate this question, given the folate values y in each group are i.i.d. normal distributed (check not shown).

```
fit=lm(folate~group, data=dat)
anova(fit)    # p=0.044
```

Since $p<5\%$, we can conclude that there are differences, i.e. the folate level is not the same in all groups.

group	red cell folate
1	243
1	251
1	275
1	291
1	347
1	354
1	380
1	392
2	206
2	210
2	226
2	249
2	255
2	273
2	285
2	295
2	309
3	241
3	258
3	270
3	293
3	328



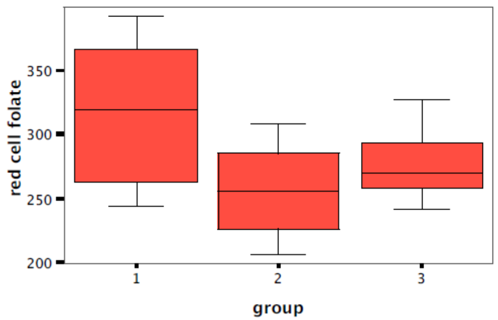
Remark: If there is only 1 factor as predictor, like treatment group, we talk about 1-way ANOVA regardless of the number of groups.

The ANOVA gets significant
Between which groups are the differences?

The significant ANOVA result, only tells us, that there are any differences. We need to perform post-hoc tests to investigate, between which groups we can really find differences.

We can perform three pair-wise t-tests. Only the t-test comparing group 1 versus 2 gets significant.

We need to correct for multiple testing, e.g. by Bonferroni-correction. Here, this correction leads to non-significance for all 3 tests.



Result of (uncorrected) pair-wise t-tests:

	Mean Diff.	DF	t-Value	P-Value
1 vs. 2	60.181	15	2.558	0.0218
1 vs. 3	38.625	11	1.327	0.2115
2 vs. 3	-21.556	12	-1.072	0.3046

List of post-hoc tests (from wiki)

- Fisher's least significant difference: LSD
- Bonferroni correction
- Duncan's new multiple range test
- Friedman test
- Newman–Keuls method
- Scheffé's method
- Tukey's range test
- Dunnett's test

The famous ANOVA table

H_0: all groups have the same population mean

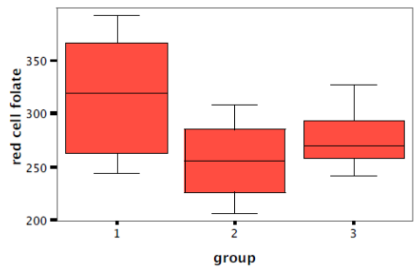
If this is true all group means are close to the overall mean and the ration
Of MSR and MSE follow a F-distribution

Source of Variation	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Residual error	n-2	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	n-1	$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

Non-parametric one-way ANOVA between >2 groups
in the case of independent data

If outcome-values given a certain predictor-value do not follow a Normal distribution, we use a non-parametric test.

independent data
All observation are independent



Dependent data
each line correspond to 1 person

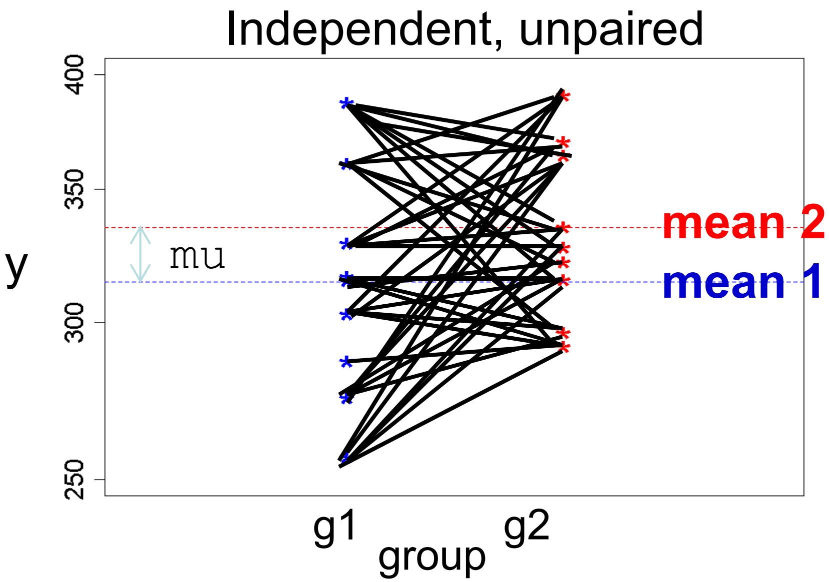
Data are independent, uncorrelated, un-paired

For the former example, it would look like:

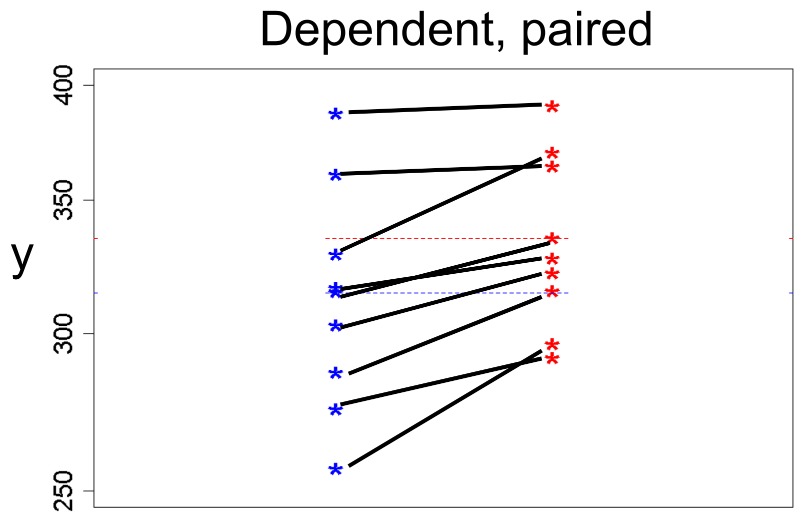
```
kruskal.test(folate~group, data=dat)
```

Remark: Paired post-hoc tests are needed in addition.

Unpaired and paired data with continuous outcome




```
t.test(g1,g2, mu=0,
       var.equal=T, paired=FALSE)

f.i=lm(y~group, data=dat)
```




```
t.test(g1,g2, mu=0,
       var.equal=T, paired=TRUE)

f.p=lm(y~group + pair.ID, data=dat)
```

Breaking the match results in a valid group/treat effect but invalid p-values.

Analyzing paired data with continuous outcome

Assumption: In each pair we assume to have the same treatment (x) effect size (`treat.effect`) meaning no interaction between pair and treatment.

Outcome is normal distributed in each treatment
~> **Appropriate analysis approaches**:

- paired t-test
- linear regression with fixed pair-effect (each pair has its own intercept)

```
lm( y ~ x + pair, data=dat)
```

Equivalent, yield same p-values and same **`treat.effect.fixMod`**

Alternative approach with valid treat.effect but problems with p-values:

- Mixed model with random pair-effect yields correct treatment effect, but p-values are only correct for no treatment effect and otherwise too small

`treat.effect.mixMod = treat.effect.fixMod`

lmer(y ~ x + (1|pair), data=dat, REML=T)

We assume that the intercepts (may vary across pairs) can be modeled as overall.intercept+random.intercept~N(0,s2)
predicted random pair effects are a shrinked version of fixed pair-effects
for the predicted random pair effects holds over all pairs: `mix.pair.effect/(fix.pair.effect)=const`

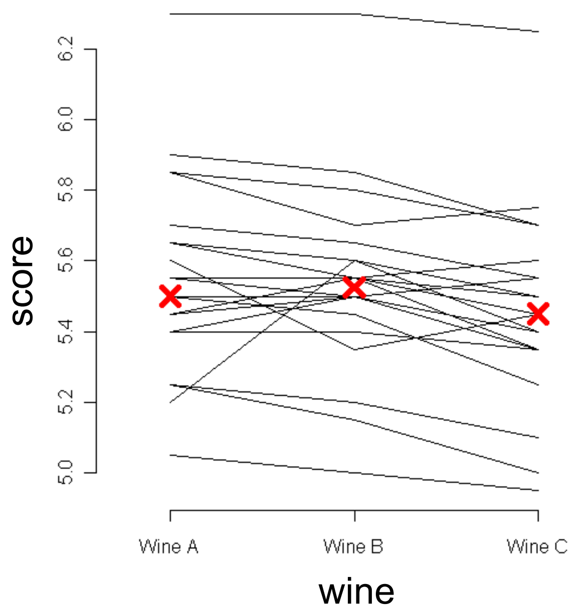
Non-parametric one-way ANOVA between >2 groups
in the case of independent data

Data are dependent, matched, grouped

Three different wines were tasted and scored by 22 people, where each person scored every wine. The data are not independent, since we have a person-grouping. To take account for individual differences in scoring, we perform the friedman-test:

```
friedman.test(Taste ~ Wine | Taster,
                 data=WineTasting)
```

Dependent data
each line correspond to 1 person



Remark: Paired post-hoc tests are needed in addition.

How to assess if there is an association between a numeric output variable and explanatory variables?

Outcome Variable	Parametric tests: The observations are normally distributed under fixed values of the input variables.		Non-parametric tests if the normality assumption is violated or the sample size is small
	un-paired independent	**paired**, dependent, correlated	
Continuous (e.g. pain scale, conc., cognitive function)	**Unpaired t-test=** **1-way ANOVA with 2 groups:** compares means between two independent groups	**Paired t-test:** compares means between two related groups (e.g., the same subjects before and after)	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the **paired** t-test for **2 groups** **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the **unpaired** t-test for **2 groups** **Kruskal-Wallis test:** non-parametric alternative to ANOVA for **>2 independent groups**. **Friedman test:** non-parametric alternative to ANOVA **>2 dependent groups**. **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient
	ANOVA: compares means between more than two independent groups: is there any difference between groups? **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Repeated-measures ANOVA:** compares changes over time in the means of ≥ 2 groups (repeated measurements) **Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	