Statistics 2

Course Overview

Casper Albers & Jorge Tendeiro

Lecture 14, 2019 – 2020

university of groningen

Overview

Decision tree

One slide per lecture overview

Preview future courses

Literature for this lecture

No new literature

Practical requirement

Practical requirement: Having attended at least 10 practicals.

Check your progress on 'my grades' (Nestor). Not agreeing? Contact Karin Siebenga, as soon as possible.

- Passed requirement: Access to exam and to follow-up courses Statistics III and Research Methods Practical.
- Failed requirement: No access to exam. Also not 'for practice purposes'. No access to follow-up courses. Next opportunity to pass: Next year.

Exam

- **Second partial exam:**
  - When: Monday 20 January, 12:15 – 13:15
  - Where: Aletta Jacobshal
- **Resit exam:**
  - When: Monday 6 April, 12:15 – 14:15
  - Where: Not yet known
  - Note: Resit takes place in Semester 2. Want to write your BSc-thesis in semester 2? Then pass regular exam.

Please check rooster.rug.nl: Locations could've changed after making these slides!

Exam: Study tips

- Learning the slides is insufficient. Read the book and papers.
- (Re)do practical exercises.
- Study together in small study groups. Discuss the exercises.
- Practice, practice, practice:
  - Examples in the book.
  - Sample exam on Nestor.
  - Additional resources mentioned in the reader.

Ways to classify methods

- The number of dependent variables.
- The number of independent variables.
- Whether the variables are categorical, continuous, or a combination of both.
- Depending on the assumptions that can be made.

Note that by summarizing statistical modelling to a decision tree, many nuances will get lost. Please keep that in mind.

Number of dependent variables

- 'Zero':
  When using contingency tables (Stats 1), there is no distinction IV/DV.
- One:
  For all other methods in Statistics 1a, 1b, 2, and 3.
- More than one:
  Use methods like MANOVA

This rule is a poor classifier for this course.

Dependent variable

Categorical Dependent Variable? Use logistic regression (Stats 3).

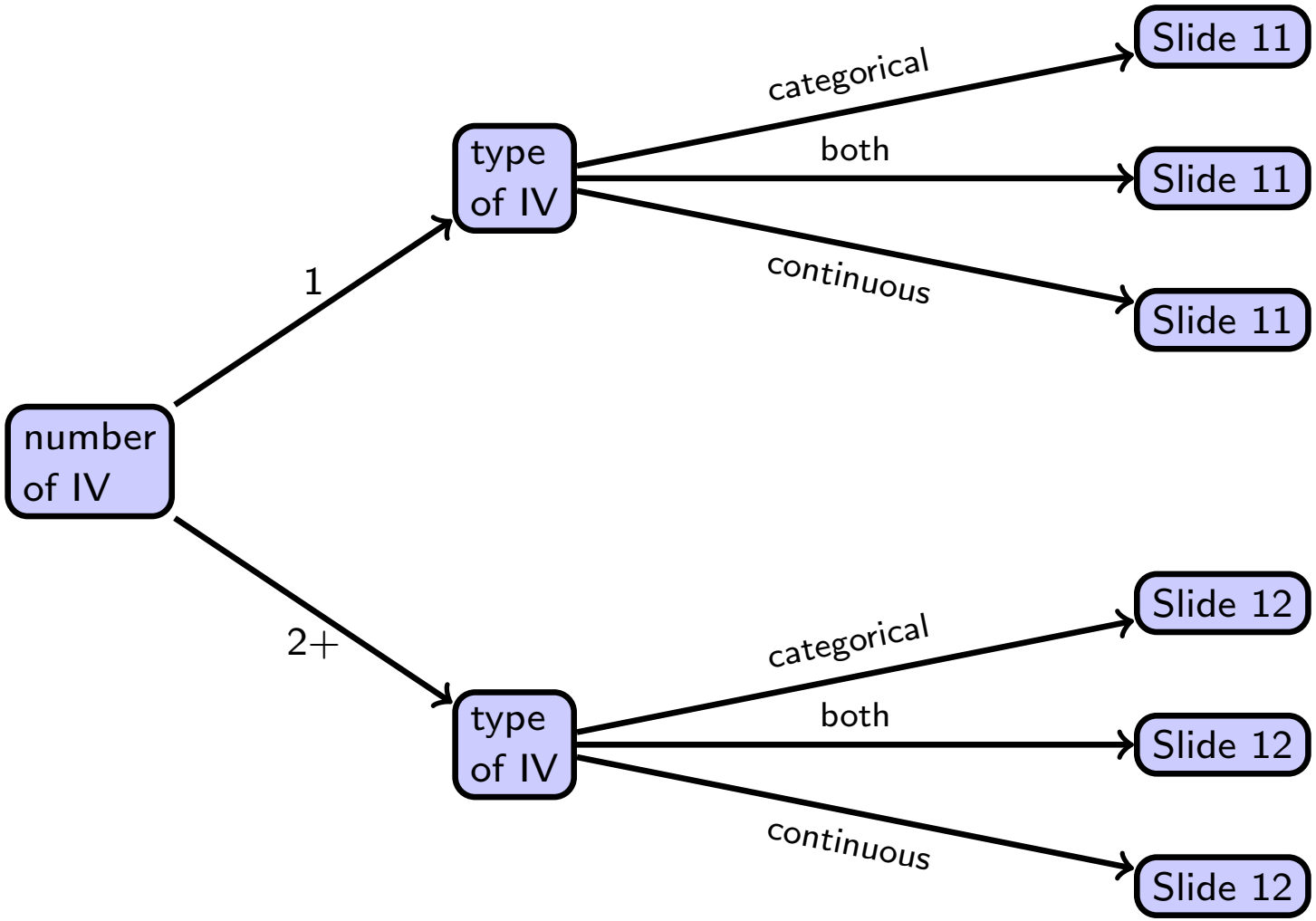
All methods in this course work with continuous Y.

Independent Variable(s)

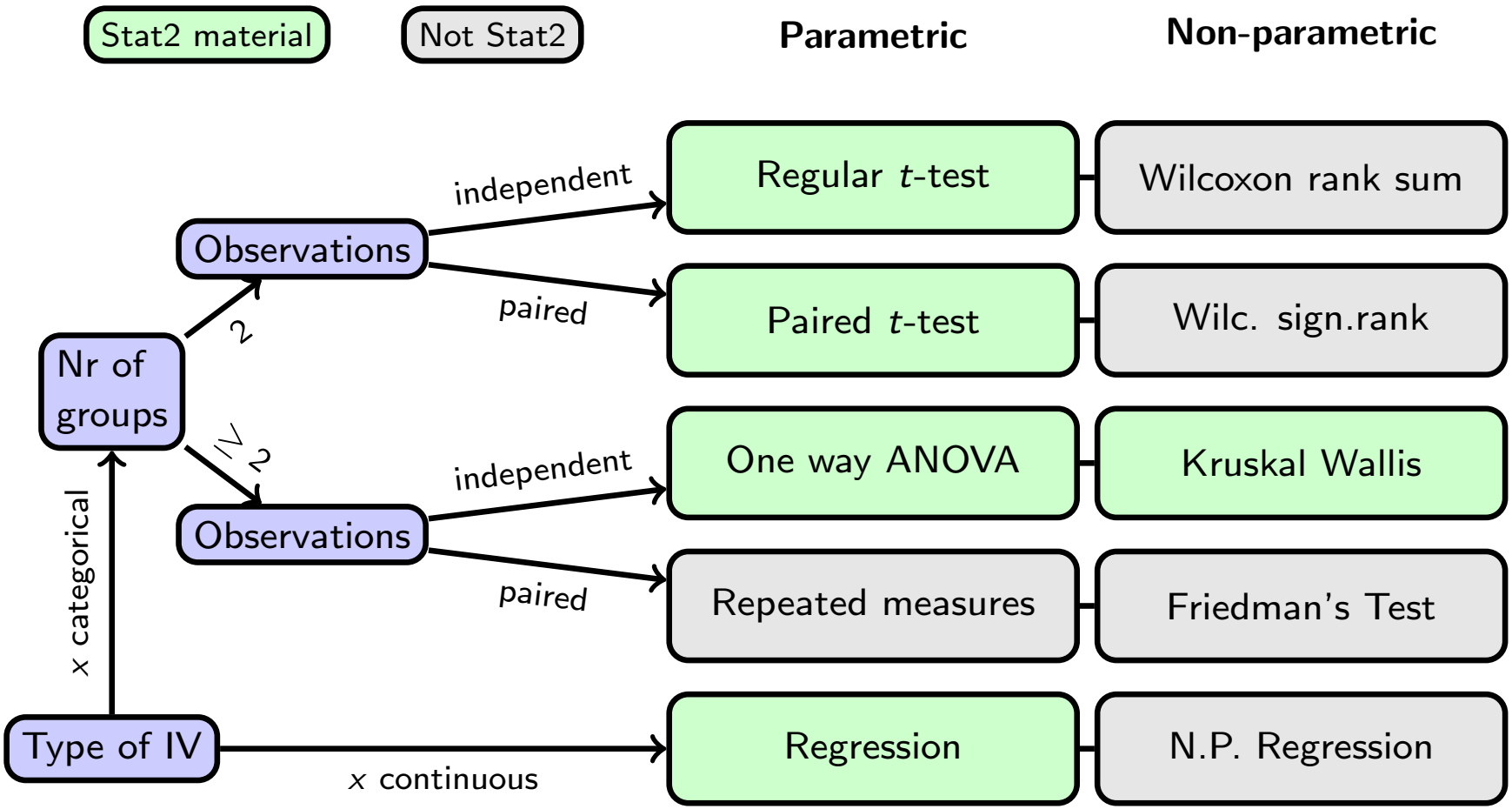
Subclassify according to the independent variables:

- How many are there?
- Are they continuous, categorical, or some of both?

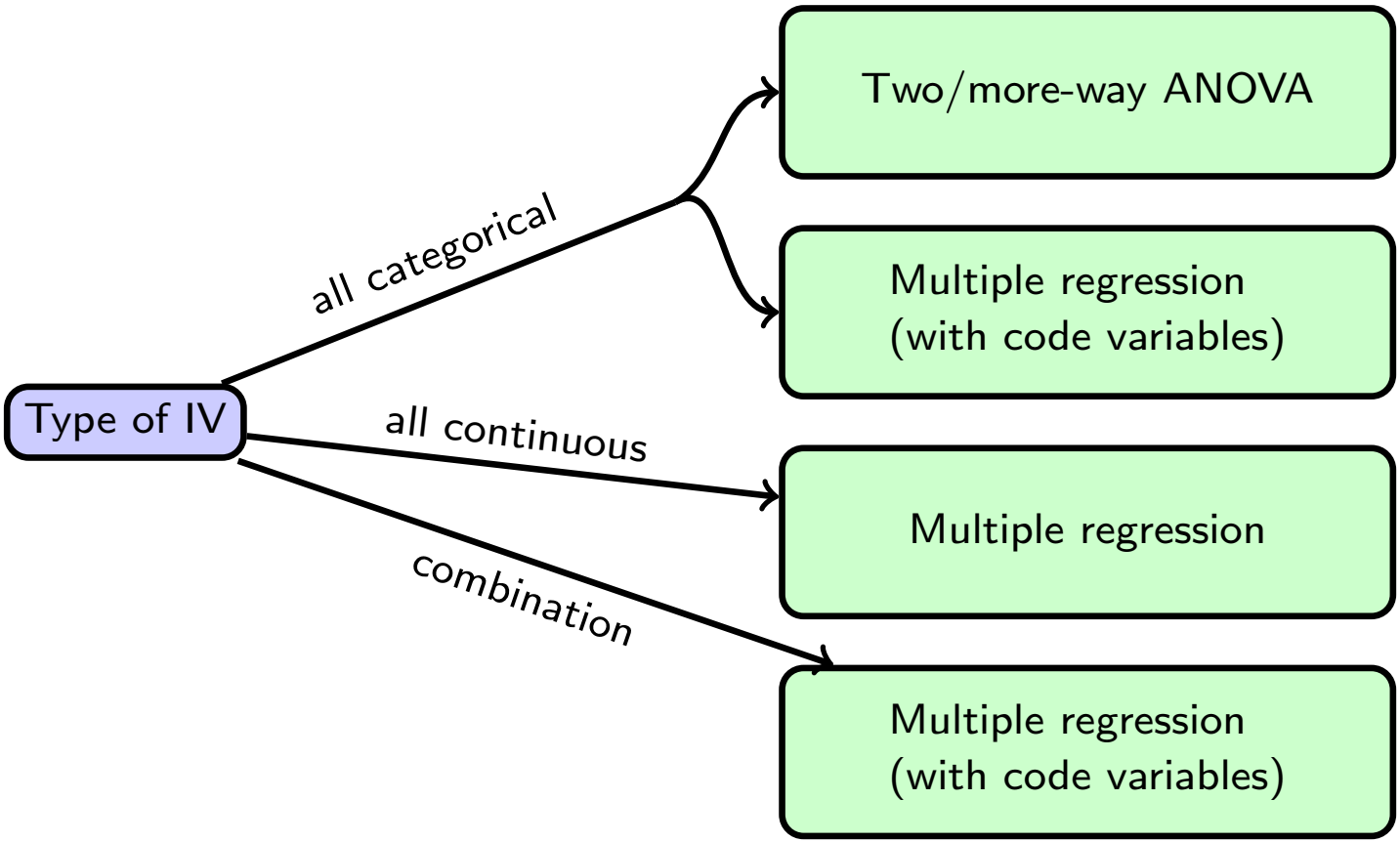
Dependent variable is continuous



One independent variable



More than one independent variable



- In the > 1 IV case, we only looked at parametric methods.
- In the > 1 IV case, we only looked at unpaired measurements .

Overview: One slide per lecture

- 15 slides.
- Key message of each week is highlighted.
- Caution: Things not on these 15 slides are important too and will be examined!

Lecture 0: Overview Statistics I

Don't forget what you learned in Stats 1a and 1b, it's still important.

- Data = Model + Error.
- Test: $\dfrac{(\text{estimate statistic}) - (\text{expected value if } \mathcal{H}_0 \text{ is true})}{\text{standard error}}$.
- Confidence interval: Estimate $\pm$ margin of error.
- Margin of error = critical value sampling distribution $\times$ SE.
- Remember from Statistics I:
  - p-value.
  - Hypothesis testing, confidence intervals.
  - Population/sample.
  - Normal and t distributions.
  - t-test.
  - Mean, sd, pooled sd, ...

- SLR models a linear relation between x and μ_y.
- Population regression line: $\mu_y = \beta_0 + \beta_1 x$.
- Statistical model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- Population parameters:
  - β_0: Intercept. Value of μ_y if $x = 0$.
  - β_1: Slope. Change in μ_y if x increases by 1.
  - σ: Spread of measurements around μ_y.
- Estimators: $b_0 = \bar{y} - b_1 \bar{x}$, $b_1 = r_{xy} s_y / s_x$.
- R^2 is percentage explained variance.

- Two types of test:
  - Is the model significant? $\Rightarrow \mathcal{H}_0$: $R^2 = 0$.
    ANOVA F-test.
  - Are the parameters significant? $\Rightarrow \mathcal{H}_0$: $\beta_i = 0$.
    t test.
- In simple linear regression, tests on β_1, R^2, and $\rho = 0$ coincide.
- Tests and CI for ρ using Fisher's Z-transformation:

$$r_z = \frac{1}{2} \log \frac{1+r}{1-r} \overset{\mathcal{H}_0}{\sim} \mathcal{N}\left(\rho_z, \frac{1}{\sqrt{n-3}}\right).$$

- Prediction intervals are wider than confidence intervals.

Lecture 3: Model validity. Causality & association

It is important to think about reasons why your model might be invalid.

- Multicollinearity, outliers, and influential points can be a problem.
- Know how to compute and interpret the VIF.
- Know how to interpret Cook's distance.
- Types of association and causality.

- Predict y from multiple x's:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

- Residuals $\varepsilon \sim \mathcal{N}(0, \sigma)$, with σ constant.
- Model significance: Measured through R^2.
- Parameter significance: Measured through β_i.
- It can be useful to adjust the multiple correlation coefficient R^2.

Source	SS	df	MS	F
Model	$\sum_i (\hat{y}_i - \bar{y})^2$	p	SS/df	Model MS/Res.MS
Residual	$\sum_i (y_i - \hat{y}_i)^2$	$n - p - 1$	s_p^2	
Total	$\sum_i (y_i - \bar{y})^2$	$n - 1$	var(y)	

Lecture 5: Interaction effects in multiple regression

Moderator analysis:

- Regression of y on x_1 and x_2:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

- Interaction: The combined impact of x_1 and x_2 on y is different from the sum of the separated effects.
- Centering is useful to improve interpretability.
- Simple regression equations, simple slopes.
- Useful: Plot SRE for $x_2 \in \{M - 1SD, M, M + 1SD\}$.

Lecture 6: (Semi)partial correlation; standardized regression

- Because of multicollinearity the importance of x_1 in predicting y can change when other predictors are included in the model.
- Drawing Ballentine Venn diagrams is useful.
- Semi-partial correlation: How much of the total variance of y is uniquely explained by this IV?
- Partial correlation: What proportion of the variance of y not explained by the other IVs, is uniquely explained by this IV?

Lecture 7: Assumptions

Four main assumptions:

1. Normality
   - Check: QQ-plot, skewness, kurtosis.
   - Solution: Non-parametrics, increase n, transformations.
2. Linearity
   - Check: Residual plot.
   - Solution: Transformation.
3. Homoscedasticity
   - Check: Boxplots (ANOVA), residual plots (regression).
   - Solution: Transformations.
4. Independence
   - Check: Design.
   - Solution: Other techniques.

- $d_i = \begin{cases} 0 & \text{if person } i \text{ not in group 1} \\ 1 & \text{if person } i \text{ in group 1.} \end{cases}$
- ANOVA is a special case of linear regression.
- Using code variables (dummies) is useful when:
  - Combining continuous and categorical independent variables.
  - Testing contrasts.

1. Write down the general model and dummy codes.
2. Fill in the dummy codes: One model per group.
3. Rewrite the equation $\mu_i = \ldots$ into $\beta_i = \ldots$
4. Estimate the parameters.
5. Perform the desired test(s).

Lecture 9: Multiple comparisons and contrasts

- Confidence interval for group means $\bar{y}_i \pm t^* \frac{s}{\sqrt{n}}$.
  (Based on s_i versus based on s_p.)
- ANOVA $\mathcal{H}_0$ rejected? There is a difference, but where?
- Multiple t tests lead to chance capitalization.
- Then, overall error rate $\approx 1 - (1 - \alpha)^k$.
- Contrasts: Differences in terms of the μ_i.
- Multiple comparisons: Testing *all* pairwise differences.
- Post Hoc tests to compensate for chance capitalization:
  - LSD (adjusts df), Bonferroni (adjusts α), and many more.

Lecture 10: One way ANOVA

- Extension of independent-samples t-test to more than two groups.
- Tests for differences in population group means.
- Each group: $y_i \sim \mathcal{N}(\mu_i, \sigma)$. $\mathcal{H}_0$: $\mu_1 = \ldots = \mu_I$.
- Split the variance: SS total = SS group + SS error.
- Compare within and between group variances.
- F = Group MS/Residual MS.
- Kruskal-Wallis: Nonparametric ANOVA.

Source	SS	df	MS	F
Group	$\sum_{i,j}(\bar{y}_i - \bar{y})^2$	$g-1$	SS/df	GMS/RMS
Error	$\sum_{i,j}(y_{ij} - \bar{y}_i)^2$	$n-g$	s_p^2	
Total	$\sum_{i,j}(y_{ij} - \bar{y})^2$	$n-1$	var(y)	

Lecture 11: Two way ANOVA

- Group membership defined through two factors, A and B.
- For both factors, you can test the main effect:

$$H_0 : \mu_1 = \ldots = \mu_g.$$

- Interaction: The difference between differences of means.
- Main effects df's: $g_A - 1$ and $g_B - 1$.
  Interaction df: $(g_A - 1)(g_B - 1)$.
- Use means plot to quickly visualise situations.
- Effect sizes for ANOVA:
  Various alternatives for 'proportion explained variance'.

Lecture 12: Bayesian statistics

- Recall Bayes' rule.
- Understand each of its components:
  - Prior;
  - Likelihood;
  - Marginal likelihood;
  - Posterior.
- Effect of prior, sample mean, sample size on posterior.
- Cromwell's rule.
- Reallocation of credibility.
- Summarize posterior.

Lecture 13: Good statistics, bad statistics.

It is important to think about reasons why your model might be invalid.

- Scientists are human. The make mistakes. Intentional and unintentional.
- Questionable Research Practices.
- The Reproducibility Crisis.
- Methods to tackle the problem:
  - Open science
  - Preregistration
  - Replication
  - Meta-analyses

Lecture 14: Overview

No new material.

Preview to Statistics III

- Still a lot of open questions after Statistics 2.
- Regression:
  - How to select which IVs to include in a model?
  - Nonlinear regression.
  - Logistic regression: When DV is categorical.
- ANOVA:
  - Contrasts in two way ANOVA.
  - Repeated measures ANOVA ('paired ANOVA').
- Important part of the course: Applying statistics to data and writing a report about it.
- And much more!

Other statistics courses in your future

Optional Modules in third year

- PSB3E-M16: Statistical Solutions to Research Problems in Psychology.
  - Four case studies driven by recent psychological research questions.
  - Aimed at those that enjoy statistics.
- PSB3E-M11: Programming for Psychologists.
  - Learn how to program.
- Possibilities for individual literature studies or thesis.
  - A lot is possible. Contact us *in time* when interested.

Other statistics courses in your future

(Regular) Master – Elective modules available:

- PSMM-2: Repeated Measures.
  - Multiple measurements (in time) per subject.
  - Multilevel model: Dependence between observations.
- PSMM-6: Test Construction.
  - Studies principles of test and questionnaire construction.
  - Studies construction, evaluation, and interpretation of tests and questionnaires.

Other statistics courses in your future

Research Master – *Many* opportunities, many courses:

- Advanced Statistics
- Applied Statistics
- Strutural Equation Modelling
- Multilevel Analysis
- Statistical Analysis of Social Networks
- Statistical Modelling of Single Cases
- Traineeships, literature studies, thesis projects, ...