Bio144 Datenanalyse in der Biologie

4. Semester Bachelorkurs Biologie / Biomedizin, FS 2017

Overarching goals

- Provide a solid foundation for answering biological questions with quantitative data.
- **Help students to *understand* the language of a statistician**
- Ability to understand and interpret results in research articles.
- Give the students a challenging, engaging, and enjoyable learning experience.

Logistics

5 CP, 150 hours
14-week semester (12 - 13 available)
4 hours per week "contact" (56 total)
5 hours per week self-study (70 total)
24 hours exam preparation
One large lecture theatre (447 pl.) and multiple smaller rooms available Thursday and Friday afternoon 13-17h.
Assessment during module examination period, admission to exam if exercises completed successfully.

Four-hour afternoon session

Two hours of lecture, two hours of practical.
Perhaps we have the last 30-mins of each afternoon class for "any questions", because we can't reply to questions outside class (e.g., by email or 1:1) as there are too many students.
Or: we have 1-2 assistants that are available during 1h/week for questions („question hour")

Ensuring appropriate course contents

Survey existing courses at other Universities and look at possible course text books?
Ask professors what they think we should teach and what datasets they might recommend / provide?

Possible course texts

http://www.roberts-publishers.com/biology/the-analysis-of-biological-data-44.html
(SM: Nice book to find some examples, does however mainly cover topics from 2nd sememster; compare to the book by Werner Stahel "Statistische Datenanalyse", which is what Luchsinger does already in Mat183)

he book Statistics – An Introduction Using R (M.J. Crawley)

https://global.oup.com/academic/product/the-new-statistics-with-r-9780198729068?cc=ch&lang=en&#

https://global.oup.com/academic/product/getting-started-with-r-9780199601622?cc=ch&lang=en&# (though wait for second edition)

Handouts

- Beamer presentations
- Course notes: Script from Stahel
- Others?

Themes running through the course:

- Importance of how to specify a question
- Importance of model checking, visual inspection and interpretation
- Scepticism about p-values / proper interpretation of p-values
- Causality vs. correlation
- One or a few datasets that can be used throughout the course
  - Note: in the R library datasets there are plenty of nice data examples. Remember!
- A general workflow for quantitative problem solving
- Importance of communication, including reproduction of analysis

Can we construct a general workflow for the modelling process
(Owen takes care of this in 1st week, I can breifly refer to it)

1. Identify a biological, medical, ... question (Give examples of types of question).
2. Collect the data, e.g. by experiment, observations etc. (!in this course usually already done!)
3. Tidy and clean the data (!in this course usually already done!)
4. Give a graphical representation of the data.
5. Choose an appropriate model.
6. Fit the model to obtain parameters and their uncertainty (confidence intervals, p-values etc)
7. Check if modelling assumptions are met, e.g. by graphical inspection.
8. If necessary, adapt the model (e.g., use another model, include or remove covariates etc)

9. Interpret the results and compare to step 1.
10. Communicate your answer with impact (clarity, simplicity, meaning).

Schedule (expecting 12 weeks)

No	Contents
L1	**Introduction and Outlook** - Introduction, importance, outlook, expectations (both ways). - Why do Biologists need statistics? - Recap of semester 2nd (test their recall and fill in the gaps?) - Start with examples: bodyfat, Hg-study, one logit, one poisson What are the questions? What does one want to know from the data? Start with visual inspection. => **Aim of course**: we want to be able to answer these questions objectively and quantitatively! Very important to do any research with data. - Introduce the workflow for quantitative problem solving. - Stress importance of **visual data exploration**, the world map of graphs (what kind of graphs exist) (Q: are students familiar with plotting histograms, barplots, boxplots etc?) - questions to be answered in this course - What is a „model“ (e.g. model for the heart, statistical model; mention uncertainty!) - Einschub: Quizz. *Exercise session:* Re-introduction to R; ev use something of Owen's „Introduction to R“ course
L2	**Introduction into simple linear Regression (one covariate):** - Are two continuous variables associated? - Graphing two continuous variables - Guess the slope, intercept, linearity - interpretation (why linear?) - parameter estimation (LS) - Modelling assumptions, check if they are met (Tukey-Anscombe, histogram of residuals) - correlation and the meaning of R^2 - confidence intervals, p-values (- confidence and prediction ranges) Ev leave some of it for week 3

	Exercise session: Doing it all in R, working examples
L3	**Multiple linear regression (several covariates):** - Do values of a continuous variable depend on multiple continuous variables? - Checking modelling assumptions (Tukey-Anscombe diagram, qq-plot) - Interpretation of p-values, t-test, F-test (for factor covariates with at least 3 levels or overall fit) - R^2 in the multiple linear framework - Residual analysis and model checking as for simple linear reg (week 3) - Binary and factor covariates, dummy encoding *Exercise session:* **Doing it all in R, working examples**
L4	**Residual analysis / checking assumptions** - Interactions - multiple vs. many simple regressions - Dive deeper into model checking and residual analysis - Introduce full list of model checking tools: - Tukey-Anscombe diagram - QQ-plot - variance plot - leverage plot - What to do when things go wrong? - transformation of the outcome, of covariates - handling of outliers *Exercise session:* **Doing it all in R, working examples**
L5	**ANOVA and ANCOVA** - Introduce ANOVA and ANCOVA - Explain that this is a special cases of linear regression, see e.g. chapter 9 of B. Bolker („Ecological Models and Data in R“) - History of ANOVA and ANCOVA (and why people still do it) - Introduce (orthogonal) contrasts and explain how they can be extracted from the regression framework - Post-hoc tests (which groups are different?) and adjusting for multiple comparison => multiple testing problem. Bonferroni correction (harsh) or Tukey HSD correction. *Exercise session:*

	Doing it all in R, explaining that aov() is a wrapper for the lm() function in R. Extract analogous results from an AN(C)OVA and linear regression
L6	**Matrix notation, matrix algebra, linAlg** - The very basics of vector and matrix notation, matrix algebra (i.e. matrix multiplication) - Why is matrix notation useful? a) Often used to compact equations, i.e in multiple regression b) Necessary to understand the language of a statistician, e.g. in books where you look up some theory - Ev show how matrix algebra is used also in other biological fields (e.g. for population dynamics etc) *Exercise session:* Do some simple matrix examples Ev. time to recapture some of previous weeks
L7	**Model selection** • Model selection is very much related to model checking (i.e., checking assumptions, residual analysis etc) • Selection criteria: Cp, AIC, BIC • Automatic model selection (stepwise backward/forward, all subsets etc) and caveats of it, warnings • Problems with collinearity of covariates • Relative importance of individual terms (i.e., which % that are explained by a covariate => explain that this is much more relevant than any p-value) • Occam's Razor principle: principle of parsimony: Systematic effects should be included in a model only if there is convincing evidence for the need of them. *Exercise session:* - Model selection procedures in R. - Calculation of AIC, BIC - Ev. introduce the relaimpo R-package (Grömping 2006) to calculate the relative importance of covariates (very useful as a supplement to the information provided by the p-value)
L8	**Interpretation of the results, causality and cautionary notes** • Interpretation and misuse of p-values, reproducibility of results (in Biomedicine much is not reproducible) • Model selection bias (i.e., bias that can emerge when „blind“ model selection is done), see paper by Freedman 1983 • Causality vs. Correlation

	• Bradford Hill criteria for causal inference • Explanation v.s. prediction (for instance: body fat example for prediction, find an example where explanation is more important) • Experiments v.s. observational studies • Model selection under causality considerations/ causal graphs *Exercise session:* - Could give an example where model selection bias emerges (too many covariates), maybe simulated example with only few observations and many covariates. Illustrate how covariates can become spuriously significant. - Maybe introduce the pcalg R-package by Kalisch, Mächler, Colombo (2012) to make causality considerations - This exercise ca be not purely with R, but also about interpretation etc.
L9	**Non-normal data I : Binary response** - What to do when outcome is binary or binomial? - When association between categorical variables required: contingency analysis (odds ratios); however, when adjusting for covariates: Logistic regression; illustrate that this is a generalization of linear regression - link functions - interpretation of parameters (odd ratios) - Avoid too mathematical formulation *Exercise session:* Simple logistic regression example(s) in R, including interpretation of the results.
L10	**Non-normal data II : Count response** - What to do when outcome is a count? - Poisson regression as another generalization of linear regression - link functions - interpretation of parameters (odd ratios) - Avoid too mathematical formulation *Exercise session:* Simple Poisson regression example(s) in R, including interpretation of the results.
L11	**L11.1 Measurement error in regression models** - Effects: ME can bias the regression parameters, mainly attenuation (underestimation) of the true effect. - When do I have to start to worry? - Simple methods to correct for ME (attenuation factor in lin. Reg, SIMEX in

	some more general cases, Bayesian approach (only mention it)) **L 11.2 Repeated measurements / random effects** - Introduction to dealing with repeated measurements and the idea of including random effects. Perhaps use example from mercury study (family effect) or child movement study (childcare effect). - Random effects capture dependency structure of similar (e.g. grouped) observations *Exercise session:* Give a linear regression example, once with true and once with error-prone covariate. Difference? How to correct for it? SIMEX. Some exercises using LMM
L12	**Miscellanea, repetitions, outlook** - „How to“ / good practice of research: • Reproducibility • Posthoc tests • How to communicate my results - Tidying and cleaning data. - Repetitions and outlook (mixed models, time series, survival models). Idea: use some data sets (e.g. from surival) to illustrate caveats of regression techniques presented in this lecture *Exercise session*: Analysis of full data set, do entire workflow in a real data example (model fitting, model selection, interpreation etc) Maybe only 1h lecture and 3h of exercises for practicing and answering questions?