

Course Bio144: Data Analysis in Biology

Lecture 1: Introduction and Outlook

Owen Petchey

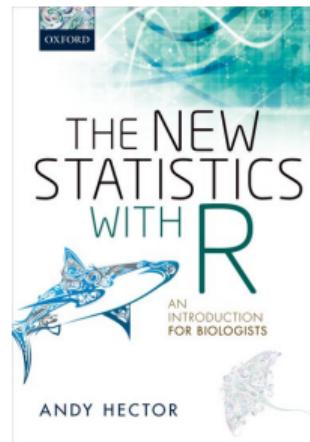
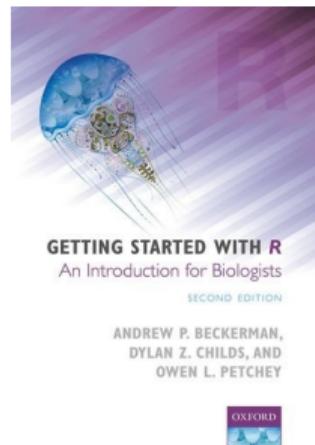
University of Zurich

24 November, 2020

Literature

Compulsory literature (books available as ebooks from uzh):

1. *Lineare Regression* by W. Stahel (pdf on course webpage)
 2. *Getting Started with R, An introduction for biologists* (**Second Edition**)
Beckerman, Childs & Petchey, Oxford University Press (DO NOT USE THE FIRST EDITION!).
 3. *The New Statistics With R* by A. Hector, Oxford University Press;
- See “Course texts/material” on course website.



Schedule (12 lecture units + 2 self-study weeks)

Unit 1 Introduction and outlook

Unit 2 No lecture

Unit 3 Simple linear regression

Unit 4 Residual analysis, model validation

Unit 5 Multiple linear regression

Unit 6 ANOVA

Unit 7 ANCOVA; Matrix Algebra

Unit 8 Model selection

Unit 9 Interpretation of results, causality

Unit 10 Count data (Poisson regression)

Unit 11 Binary Data (logistic regression)

Unit 12 Measurement error, random effects, selected topics

Overarching goals of the course

- ▶ 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**.

My belief: A solid foundation in statistics makes you independent!

Why is statistical data analysis so relevant for the biological and medical sciences?

Awareness that, without a profound knowledge in statistical data analysis, it will be hard to analyze your data from Bachelor, Master or PhD theses. . . .

- ▶ **Medicine:** What is the effect of a drug? Which factors cause cancer?
- ▶ **Ecology:** What is a suitable habitat for a certain animal? Which resources does it need or prefer?
- ▶ **Evoloutionary biology:** Do highly inbred animals have decreased chances to survive or reproduce?

Learning by doing??

→ **Not advisable** in statistics. Experience is essential, there are many pitfalls.

A good foundation in statistics **makes you more independent** from consultants or the goodwill of colleagues. Without such a knowledge, you will always need help from others.

Data analysis/statistics is itself an exciting part of research!

Data analysis is at the **interface between mathematics and biology/medicine** (and many other applied research fields).

Own examples

Otter (*lutra lutra*)

Research questions: What is the preferred habitat by otters? How do otters adapt to human altered landscapes?

Method: Study in Austria, 9 Otter were radio-tracked and monitored during 2-3 years.



Flexible habitat selection paves the way for a recovery of otter populations in the European Alps



Irene C. Weinberger ^{a,*}, Stefanie Muff ^{a,b}, Addy de Jongh ^c, Andreas Kranz ^d, Fabio Bontadina ^{e,f}

^a Institute of Ecology and Evolutionary Biology, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland

^b Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

^c Dutch Otterstation Foundation, Spanjaardslaan 136, 8917 AX Leeuwarden, Netherlands

^d alka-kranz Ingenieurbüro für Wildökologie und Naturschutz, Am Waldgrund 25, 8044 Graz, Austria

^e SWILD – Urban Ecology & Wildlife Research, Wührstr. 12, 8003 Zurich, Switzerland

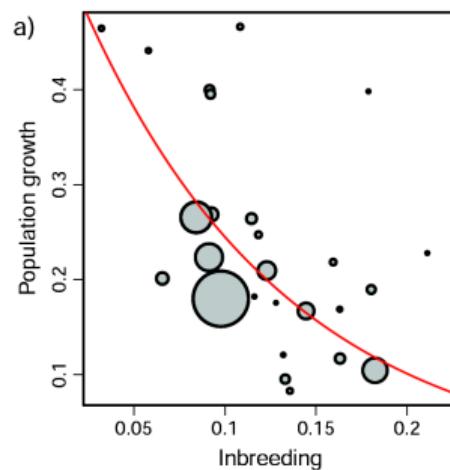
^f Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology, 8903 Birmensdorf, Switzerland

{ <http://www.prolutra.ch/> }

Inbreeding in Alpine ibex

Research question: Does inbreeding in Alpine ibex populations have a negative effect on long-term population growth? Inbreeding depression!

Methods: Genetic information from blood samples allow to quantify the level of inbreeding in each ibex population. In addition, long-term monitoring of population sizes and harvest rates.



Mercury (Hg) in the soil

Wohnzone im Wallis von Quecksilber vergiftet

Vor über vierzig Jahren hatten 3,1 Tonnen Quecksilber einen Abflusskanal nahe der Walliser Gemeinde Visp verschmutzt. Noch heute müssen die Einwohner mit den Folgen leben.



Artikel zum Thema

Konvention gegen Quecksilber verabschiedet

Ein neues internationales Abkommen schränkt die Verwendung von Quecksilber in der Industrie ein. Massgeblich daran beteiligt war die Schweiz. [Mehr...](#)

19.01.2013

Research question: Is the Hg level in the environment (soil) of people's homes associated to the Hg levels in their bodies (urin, hair)?

Method: Measurements of Hg concentrations on people's properties, as well as measurements and survey of children and their mothers living in these properties.

Highly delicate, emotionally charged, political question! ▶ [Schweiz Aktuell, 20. Juni 2016](#)

Physical activity in children (Splashy study)



splashy.ch

Research question: Which factors influence physical activity patterns in children aged 2-6 years?

Method: The children had to wear accelerometers for several days. In addition, their parents had to fill in a detailed questionnaire.

Observed variables were, e.g., media consumption, behavior of the parents, age, weight, social structure, . . .

Statistics in the news (April 2016)

Wissen 61

Überschätzte Statistiken

Daten-Analysen entscheiden heute darüber, ob ein Medikament als wirksam gilt. Doch verstehen viele Forscher die Bedeutung dieser Rechnungen gar nicht. Von Patrick Imhasly

Karies machen kleinen nicht nur Menschen, sondern auch statistische Größen. Das gilt sowohl für den Zahnarzt als auch für den Statistiker. Karies kommt, vor allem aber jeder, der im weiteren Sinn etwas mit Daten zu tun hat. Ein *p*-Wert kann auf die schiefen Zähne geraten. Wenn das Vater der modernen Statistik, Carl Pearson, 1900 einen *p*-Wert als eine Art Informations-Sensitivität für die Aussagekraft von Daten erneuterte, ist in der Praxis kaum zu einem simples Lächeln gekommen.

Engt die statistische Analyse von Daten einen Wissenschaftler so sehr ein, dass er besser (<0.01 Prozent), gehen diese als *signifikant* – den Daten wird dann zentralisch attestiert. Ein *p*-Wert kann darüber darüber, ob ein neues Medikament als wirksam eingestuft wird oder ob ein Frosch seine Partnerin kann, eben ein *p*-Wert weiß nie dazu gedacht, wissenschaftliches Denken anzuregen. John Noga, der Direktor der Amerikanischen Statistischen Vereinigung (ASA), fügt offenbar hinzu:

„Statistiken werden durch den *p*-Wert immer häufiger, ohne zu verstehen, was er bedeutet, das Risiko einer Fehlentscheidung und die Unschärfe der Wissenschaft. Der Mediziner und Pädiatologe John Kadlecik von der Universität Stanford spricht in einem Interview darüber: „Der falsche Gebrauch des *p*-Werts ist dermaßen einfach und erfolgt so ununterbrochen, dass manche Fach-

wissen daran – vor allem wenn sie mit Forschungsgeldern und Publikationsbedürfnis konkurrieren – nicht daran denken, dass es sich um eine schiefen Zahntyp handelt.“

Die ASA geht voran, um ersten Mal in fast 180-jähriger Geschichte Empfehlungen für die Statistik zu erlassen. Eine der wichtigsten ist, dass die statistischen Größen verzweigt eingesetzt werden.

Wieder die Null-Hypothese

Der *p*-Wert sagt nicht das aus, was man gewünscht vor ihm erreicht, erklärt der Statistiker Leonhard Held, Professor für Statistik am Applied Health Research Center der Universität Berne mit. Das ist kein Fehler, sondern ein Fehler, der die Wissenschaftlichkeit, ob eine bestimmte Hypothese zutrifft, und auch nicht, ob sie besser ist, als andere. Der *p*-Wert ist, wie die ASA festhält, ein Maß für die Wahrscheinlichkeit, dass die Null-Hypothese stimmt. Diese Wahrscheinlichkeit ist genau genommen, dass der Test, der auf der Null-Hypothese basiert, eine falsche Hypothese ablehnt. Ein *p*-Wert von 0,05 bedeutet, dass der Test, der auf der Null-Hypothese basiert, unter den Bedingungen der Null-Hypothese mit einer Wahrscheinlichkeit von 95 Prozent eine falsche Hypothese ablehnen kann – und nicht, dass eine bestimzte Hypothese mit einer Sicherheit von 95 Prozent stimmt.

Über die eigentlich interessante Hypothese kann der *p*-Wert nur reden, wenn sie bestätigt ist. Der *p*-Wert liefert also keine ultimative Beweise für einen positiven Effekt, sondern nur eine Aussage über die Wahrscheinlichkeit, dass eine falsche Hypothese abgelehnt wird, und es interessiert sie auch nicht.

Haben Sie sich schon einmal gefragt, warum ein *p*-Wert nicht direkt die Wahrscheinlichkeit ist? Praktisch historisch entstand und herauswuchs klar definierte Regeln, die die *p*-Werte erlaubten. Und da die *p*-Werte, überall wo sie eingesetzt werden, die Interpretation des einzelnen Forschers, ob wirklich etwas passiert ist, nicht beeinflussen können, haben sie sich gewöhnt. Werden sie die wissenschaftlich gewünschten Signifikanzgrenzen im Gefangen zu halten, sagt Leonhard Held vom Institut für Epidemiologie, Biostatistik und Prevention der Universität Zürich.

Das Problem ist, dass die *p*-Werte nicht nur die Wahrscheinlichkeit bestimmen, sondern auch die Aussagekraft. Das ist ein Fehler, den George Evelyn Smith schon vor 15 Jahren im *British Medical Journal* erkannt hat:

„Statistiken führen heute zu übermessen vielen Daten, dass man allen Untergang feststellt.“

Die *p*-Werte sind also ein Instrument, um Vertrauensintervalle zu versichern, die spezifische Aussagen über die Unsicherheit einer Schätzung machen. Statistisch gesehen ist es ein Messinstrument für statistische Evidence – was Beispiel Bayes-Paketen. Mit ihrer Hilfe kann sich die Wahrscheinlichkeit einer Hypothese erhöhen, aber sie kann sie auch abschwächen, statt dass diese wie beim *p*-Wert nach einem Schwarz-Weiss-Schema angezettelt oder abgelehnt wird.



Producing nonsense with statistics...

... is too easy ...

A profound knowledge of data analysis and statistics protects you from producing nonsense – and helps to detect it. See for example:

[The risks of alcohol \(by David Spiegelhalter, 23. August 2018\)](#)

[“Calling bullshit” course \(University of Washington\)](#)

Data example 1: Prognostic factors for body fat

(From Theo Gasser & Burkhardt Seifert *Grundbegriffe der Biostatistik*)

Body fat is an important indicator for overweight, but difficult to measure.

Question: Which factors allow for precise estimation (prediction) of body fat?

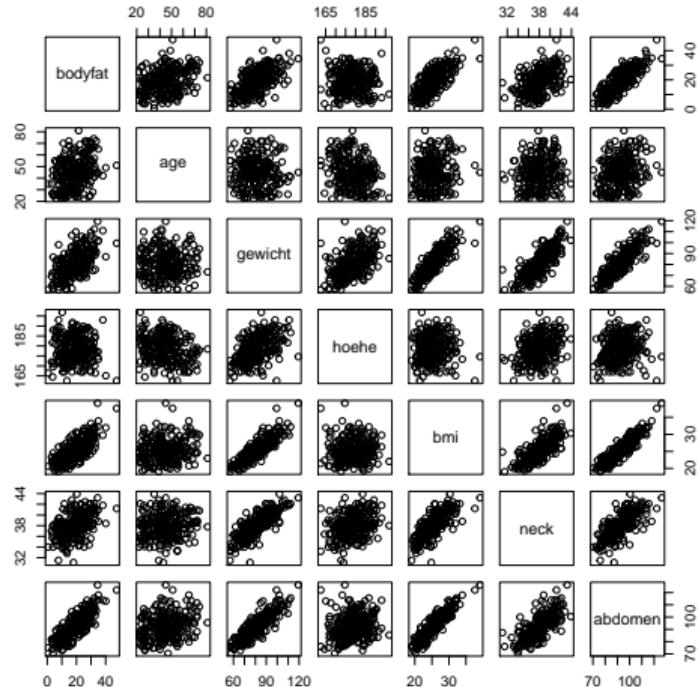
Study with 241 male participants. Measured variable were, among others, body fat (%), age, weight, body size, BMI, neck thickness and abdominal girth.

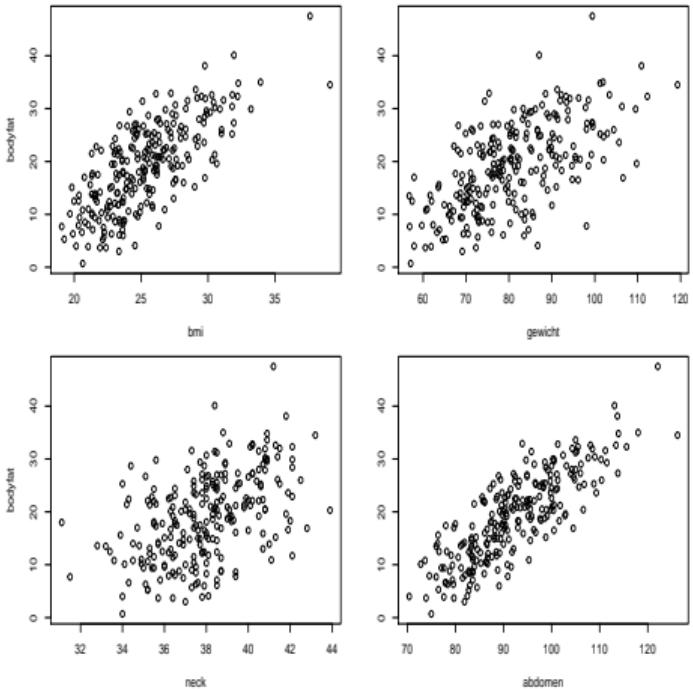
```
glimpse(d.bodyfat)
```

```
## Rows: 243
## Columns: 7
## $ bodyfat <dbl> 12.3, 6.1, 25.3, 10.4, 28.7, 20.9, 19.2, 12.4, 4.1, 11.7, 7...
## $ age      <int> 23, 22, 22, 26, 24, 24, 26, 25, 25, 23, 26, 27, 32, 30, 35, ...
## $ gewicht   <dbl> 70.03, 78.66, 69.92, 83.88, 83.65, 95.45, 82.17, 79.90, 86....
## $ hoehe     <dbl> 172.09, 183.52, 168.28, 183.52, 180.98, 189.87, 177.17, 184...
## $ bmi       <dbl> 23.65, 23.36, 24.69, 24.91, 25.54, 26.48, 26.18, 23.56, 24....
## $ neck      <dbl> 36.2, 38.5, 34.0, 37.4, 34.4, 39.0, 36.4, 37.8, 38.1, 42.1, ...
## $ abdomen   <dbl> 85.2, 83.0, 87.9, 86.4, 100.0, 94.4, 90.7, 88.5, 82.5, 88.6...
```

Scatterplots

```
pairs(d.bodyfat)
```





We are looking for a *model* that **predicts** body fat as precisely as possible from variables that are easy to measure.

Data example 2: Mercury (Hg) in Valais (Switzerland)

Question: Association between Hg concentrations in the soil and in urine of the people living in the respective properties. We use a slightly modified data set here.

```
#glimpse(d.hg)
```

A first visual inspection is not very informative. There is not much that is visible by eye:

Which other factors might be responsible for high Hg concentrations in urine?

From these plots it is hard to tell which factors exactly influence the Hg pollution in humans.

It is always useful to look at the distribution of the variables in the model. Let us plot the histogram of Hg concentrations:

All Hg values seem to “stick” at 0.

The scatterplot does also look much more reasonable with log-transformed values:

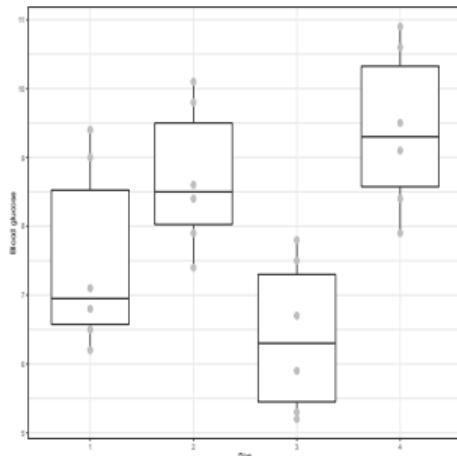
Remember: The idea to log-transform the variables was mainly obvious thanks to **visual inspection!**

Data example 3: Diet and blood glucose level

24 persons were split into 4 groups. Each group followed another diet (DIAET). The blood glucose concentrations were measured at the beginning and at the end (after 2 weeks). The difference of these values was stored (BLUTZUCK).

Question: Are there differences among the groups with respect to changes in blood glucose concentrations?

Let's look at the raw data (points and boxplots):



Data example 4: Blood-screening

Is a high ESR (erythrocyte sedimentation rate) an indicator for certain diseases (rheumatic disease, chronic inflammations)?

Specifically: Is there an association between the concentrations of the plasma proteins Fibrinogen and Globulin and ESR level $ESR < 20\text{mm/hr}$?

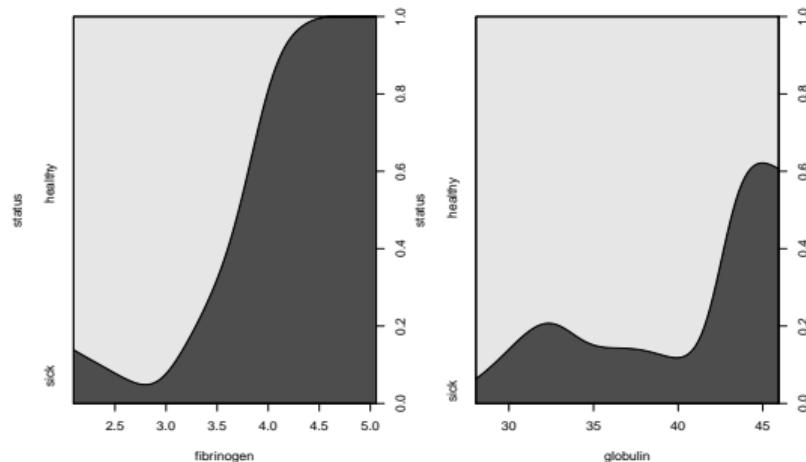
The plasma data come with the package HSAUR3

```
##      fibrinogen globulin      ESR status
## 1        2.52      38 ESR < 20 healthy
## 5        3.41      37 ESR < 20 healthy
## 9        3.15      39 ESR < 20 healthy
## 10       2.60      41 ESR < 20 healthy
## 19       2.60      38 ESR < 20 healthy
## 15       2.38      37 ESR > 20    sick
```

The distinction $\text{ESR} < 20 \text{mm/hr}$ (healthy) vs. $\text{ESR} \geq 20 \text{mm/hr}$ (sick) leads to a **binary** response variable.

→ *conditional density plot*

```
par(mfrow=c(1, 2))
cdplot(status ~ fibrinogen, plasma)
cdplot(status ~ globulin, plasma)
```



What is a model?

A model is an approximation of the reality. **Understanding how the real world works** is usually only possible thanks to simplifying assumptions.

→ This is exactly **the purpose of statistical data analysis**.

In 2014, David Hand wrote:

In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality – a point well made by George Box in his often-cited remark that "all models are wrong, but some are useful."

Steps in a modelling process (“work flow”)

The scopes of statistical data analysis

- a. **Prediction (extrapolation), interpolation.** Example body fat: use substitute measurements to predict body fat of a person.
- b. **Explanation; determination of important variables.** Example physical activity of children: The study aims to find factors that (positively or negatively) influence the movement behavior of children.
- c. **Estimation of parameters and quantify the uncertainty.** Example: Effect size of a novel drug.
- d. Optimization.
- e. Calibration.

In this course we are concerned with a-c.

Goals of the course (part 2)

By the end of the course you will be able

- ▶ to **analyze** all data examples introduced here using R (and of course many more),
- ▶ to **report and interpret** the results,
- ▶ to **draw conclusions** from them,
- ▶ to **give graphical descriptions** of the data and the results,
- ▶ to **be critical** about what you see.

Graphical representation of data

You should remember the following options for graphical data descriptions. Several of them appeared already in previous examples.

Representation	Useful for
Scatterplots	Pairwise dependency of continuous variables.
Histograms	Distribution of continuous variables.
Boxplots	Distribution of continuous variables, ev. conditionally on categories.
Conditional density plots	Dependency of a binary variable from a continuous variable.
Coplots	Dependencies among multiple variables.

Coplots

Ideal to graphically display dependencies when more than two variables are involved.
Very useful for categorical variables. Example: Mercury in Valais.

```
#coplot(log(Hg_urin) ~ age | mother * migration,  
# d.hg, panel = panel.smooth)
```

There are many “fancy” ways to graphically display data (**nice-to-know**):

- ▶ 3D-plots
- ▶ Spatial representations (using geodata)
- ▶ Interactive graphs and animations

Many R packages are available for various purposes. Interactive apps can, for example, be generated with Shiny. Check out the shiny gallery:

<http://shiny.rstudio.com/gallery/>

Next week: Simple linear regression

It will be partially a repetition of what you heard in Mat183, chapter 10.2.

Refrences

MISSING REFERENCES