

# Course Bio144: Data Analysis in Biology

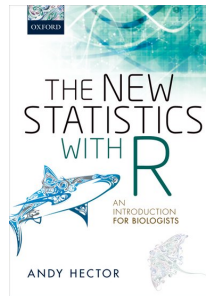
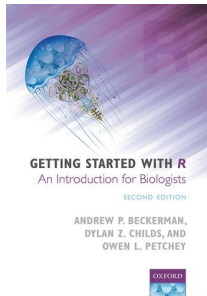
Owen L. Petchey (Practicals) & Stefanie Muff (Lectures)

Lecture 1: Introduction and Outlook  
22./23. February 2018

# 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 weeks + 2 self-study weeks)

**Week 1** Introduction and outlook

**Week 2** Simple linear regression

**Week 3** Residual analysis, model validation

**Week 4** Multiple linear regression

**Week 5** ANOVA

**Self-study week**

**Week 6** ANCOVA; Matrix Algebra

**Week 7** Model selection

**Week 8** Interpretation of results, causality

**Week 9** Count data (Poisson regression)

**Self-study week**

**Week 10** Binary Data (logistic regression)

**Week 11** Measurement error, random effects

**Week 12** Selected topics, repetition and outlook

# 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....

Examples:

- **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?
- **Evolutionary biology:** Do highly inbred animals have decreased chances to survive or reproduce?

## !! Warning !!

"Learning by doing" is often **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.

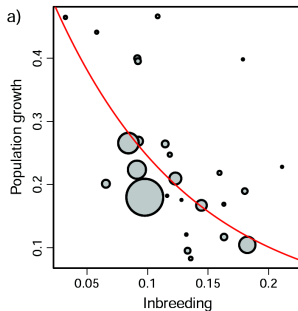


<http://www.prolutra.ch/>

## 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.





### 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,...



# 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:

Finding dodgy statistics (The Guardian, July 17, 2016)

“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.

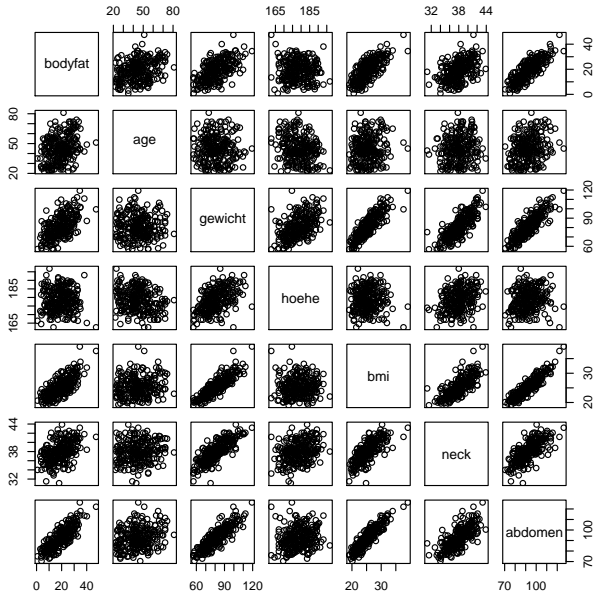
```
> glimpse(d.bodyfat)
```

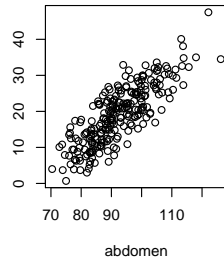
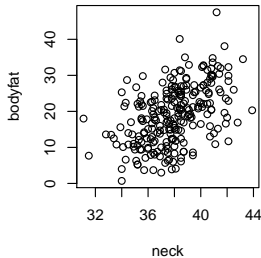
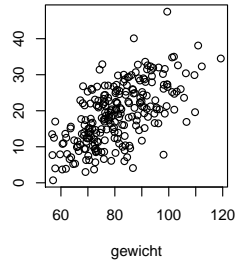
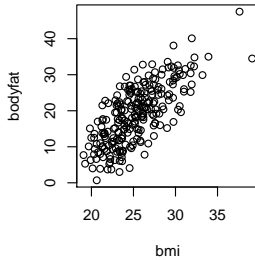
```
Observations: 243
```

```
Variables: 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...
```

```
> 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)
```

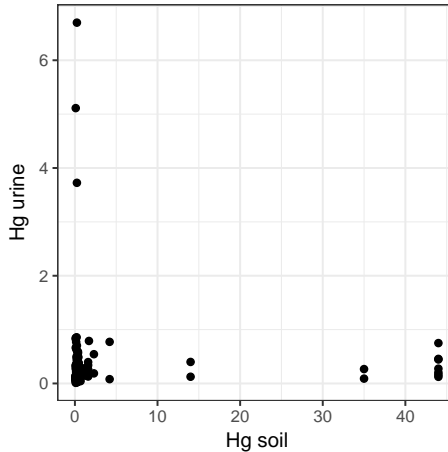
```
Observations: 156
```

```
Variables: 10
```

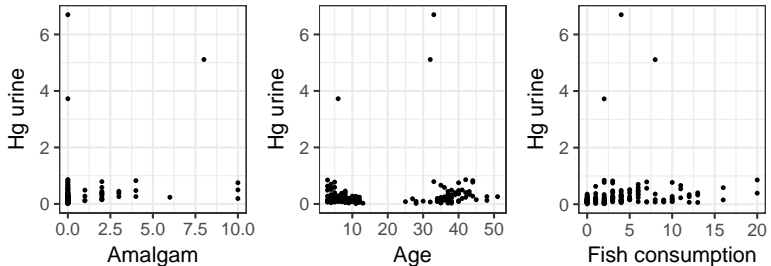
```
$ Hg_urin      <dbl> 0.25806452, 0.03597122, 0.16025641, 0.31428571, 0.28...  
$ Hg_soil      <dbl> 0.49, 0.42, 0.18, 0.49, 0.24, 0.20, 0.10, 14.00, 0.1...  
$ veg_garden   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...  
$ migration    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
$ smoking      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
$ amalgam      <int> 3, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0...  
$ age          <int> 51, 11, 34, 8, 6, 40, 7, 48, 11, 38, 7, 5, 35, 4, 39...  
$ fish         <int> 3, 2, 5, 4, 4, 2, 2, 4, 0, 7, 2, 4, 0, 0, 4, 0...  
$ last_time_fish <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
$ mother       <fctr> 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, ...
```



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

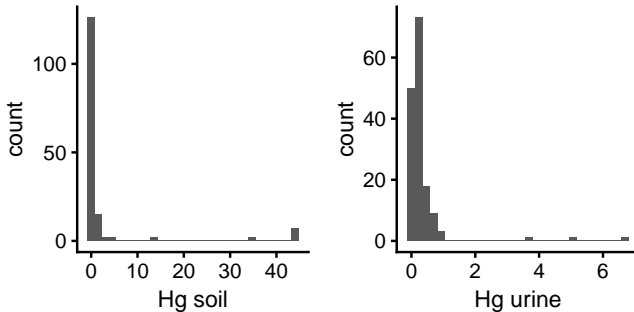


Which other factors might be responsible for high Hg concentrations in urine?



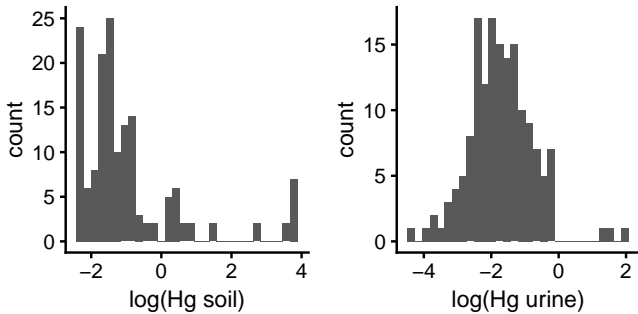
From these plots it is hard to tell which factors exactly influence the Hg pollution in humans.

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

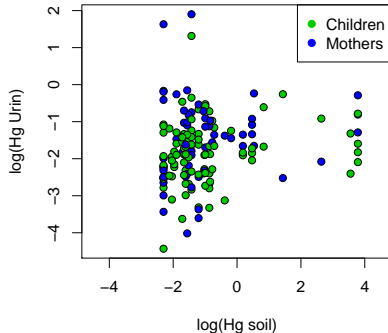


All Hg values seem to “stick” at 0.

In such cases it can help to *log-transform* the respective variables.



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**!

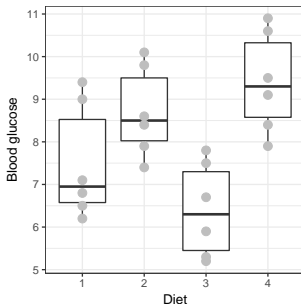
## Data example 3: Diet and blood glucose level

(Elpelt and Hartung, 1987, p. 190)

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):



Does this question seem familiar to you? (What would you do for two groups?)

For more than 2 groups we need the *ANOVA* (=ANalysis Of VAriance) approach (see chapter 10.1 in the Mat183 script).

We will see in lecture 5 that there are in fact differences between the diets.

## Data example 4: Blood-screening

(From Hothorn and Everitt, 2014, Chapter 7.1)

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

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

The plasma data come with the package HSAUR3 (Hothorn and Everitt, 2014).

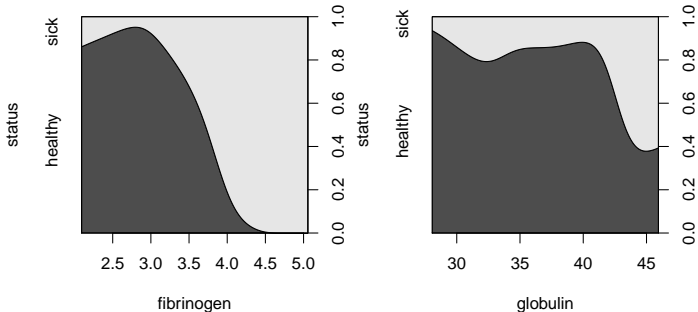
```
> plasma[c(1,5,9,10,15,29),]  
      fibrinogen globulin      ESR status  
1         2.52       38 ESR < 20 healthy  
5         3.41       37 ESR < 20 healthy  
9         3.15       39 ESR < 20 healthy  
10        2.60       41 ESR < 20 healthy  
19        2.60       38 ESR < 20 healthy  
15        2.38       37 ESR > 20   sick
```



The distinction  $ESR < 20mm/hr$  (healthy) vs.  $ESR \geq 20mm/hr$  (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” (Box, 1979).*

# 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.

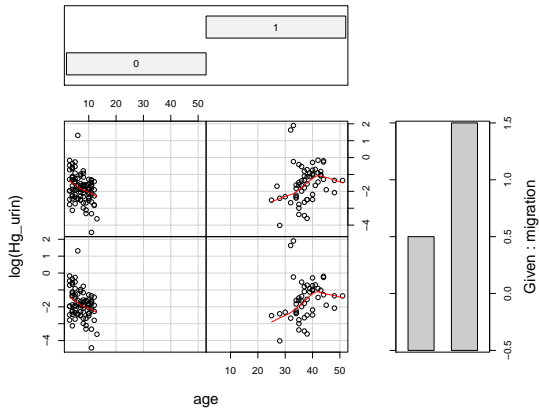
| Representation            | Useful for  |
|---------------------------|---|
| Scatterplots              | Pairwise dependency of continuous variables.                          |
| Histograms                | Distribution of numerical variables.                                  |
| Boxplots                  | Distribution of numerical 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)
```

Given : mother



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.



## References:

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson (Eds.), *In Robustness in Statistics*, pp. 201–236. New York: Academic Press.

Elpelt, B. and J. Hartung (1987). *Grundkurs Statistik, Lehr- und Übungsbuch der angewandten Statistik*.

Hothorn, T. and B. S. Everitt (2014). *A Handbook of Statistical Analyses Using R* (3 ed.). Boca Raton: Chapman & Hall/CRC Press.