

# Course Bio144: Data Analysis in Biology

## Lecture 1: Introduction and Outlook

Owen Petchey

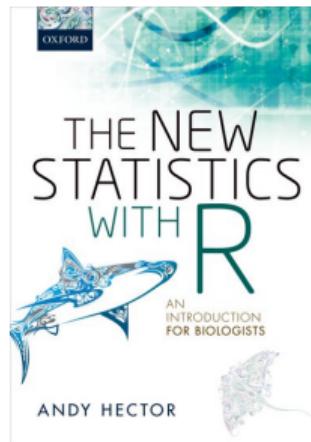
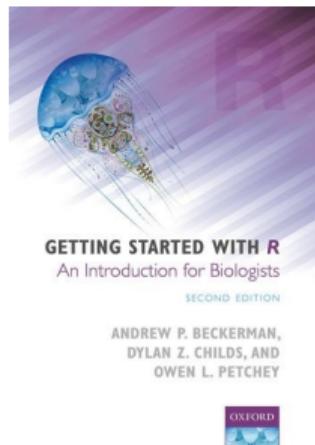
University of Zurich

06 January, 2021

# 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 <sup>a,\*</sup>, Stefanie Muff <sup>a,b</sup>, Addy de Jongh <sup>c</sup>, Andreas Kranz <sup>d</sup>, Fabio Bontadina <sup>e,f</sup>

<sup>a</sup> Institute of Ecology and Evolutionary Biology, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland

<sup>b</sup> Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

<sup>c</sup> Dutch Otterstation Foundation, Spanjaardslaan 136, 8917 AX Leeuwarden, Netherlands

<sup>d</sup> alka-kranz Ingenieurbüro für Wildökologie und Naturschutz, Am Waldgrund 25, 8044 Graz, Austria

<sup>e</sup> SWILD – Urban Ecology & Wildlife Research, Wührstr. 12, 8003 Zurich, Switzerland

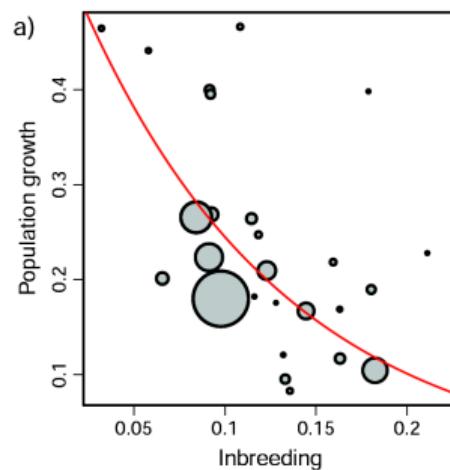
<sup>f</sup> Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology, 8903 Birmensdorf, Switzerland

<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**

**Überschätzte Statistiken**

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

**K**inder machen können nicht nur Menschen, sondern auch statistische Größen. Das gilt besonders für die *p*-Werte, die in der Biostatistik eine Rolle spielen. Ein *p*-Wert ist mit dem jüngsten Medikamenten-Hype verbunden, der in Kontakt kommt, vor allem aber jeder, der im wissenschaftlichen Bereich arbeitet. Und das ist statthaft.

**Wieder auf die schiefen Bahn geraten.** Denweise ist Vater der modernen Statistik, der es in den 1920er Jahren erstmals als eine Art formelles Servietten für die Ausmargrafie von Daten entwickelte. Ist in der Praxis durch ein etwas lächerhaftes Verhalten verdeckt.

Eigentlich die statistische Analyse von Daten bestimmt, ob eine Aussage signifikant ist. Wenn ein *p*-Wert kleiner als 0,05 Prozent ist, gehen diese als signifikant. Den *p*-Wert wird dann automatisch in die Presseberichte übernommen und darüber dargestellt, ob ein neues Medikament als wirksam eingestuft wird oder ob ein entsprechendes Ergebnis zufällig ist. *p*-Werte sind also wichtig, wissenschaftlich aber nie so gedacht, wissenschaftliche Presseberichte aber eben. Ob *p*-Werte nun tatsächlich gescheitert sind, wie die ASA festhält. Vielleicht dämmert es Ihnen, dass *p*-Werte, die so wichtig zu verweisen.

**Die Hypothese in einer Patientenstudie könnte nun bestätigt werden.** Eine solche Wissenschaftsbedrohung ist eine *Hypothese*. Sie ist eine Aussage, die man überprüfen kann – und sollte, dass sie bestätigt ist. Sicherheit ist kein Ziel.

Über die eigentlich unbestreitbare Hypothese kann der *p*-Wert zu irgendeiner Aussage führen. Aber es kann auch bestätigt sein, dass die Hypothese falsch ist. Der *p*-Wert liefert also keine absoluten Beweise für einen postulierten Sachverhalt. Er kann nur bestätigen. Der *p*-Wert ist eine Bedrohung und nicht eine Abschutz-Wissenschaftsbedrohung, erklärt Peter Hirsch, Professor für Epidemiologie, Biostatistik und Prevention der Universität Zürich. Er ist einer der Autoren des Berichts der American Statistical Association (ASA), der gerade die Wissenschaftlichkeit des *p*-Werts kritisiert.

**Statistiken führen heute zu demnassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von *p*-Werten kommt.**

Der Mediziner und Epidemiologe John Ioannidis von der University of Stanford sprach in einem Interview mit dem US-Magazin *Scientific American*: „Der falsche Gebrauch des *p*-Werts ist dermaßen einfach und einfallsreich, dass manche ausgenutzt, dass manche ausgenutzt,

verdeutlicht – vor allen wenn sie mit Forschungsgeldern und Publikationsbelohnungen belohnt werden. Das ist der Grund, warum die ASA jetzt veranlasst, zum ersten Mal in ihrer fast 100-jährigen Geschichte Empfehlungen zu formulieren, um die Verwendung statistischer Methoden verzerrungsfrei zu gestalten.“

**Daten werden meist von Leuten analysiert, die nicht dafür ausgebildet sind.**

**5%**

**Kleiner als dieser Wert muss der Unterschied zwischen den beiden statistischen Testwerten sein, damit die Daten aus einer Studie als ausreichend für die Grenze ist willkürlich gewählt.**

„Um einen statistisch signifikanten Unterschied zu finden, muss der Unterschied zwischen den beiden Testwerten größer als 5% sein. Ganz gut hat das bereits wenig, wie das Team von John Ioannidis in seinem Bericht geschrieben: „Wir haben gesehen, dass die *p*-Werte nach und in den vergangenen 25 Jahren in den klinischen Studien zunehmend niedrig geworden sind, was die Aussicht auf eine tatsächliche Verbesserung der Therapie aufgeweckt hat. Die *p*-Werte zeigen, die studien immer klar signifikant ausgewiesen haben, die Aussichten auf eine Verbesserung der Therapie aufgeweckt haben.““

**Es gibt Alternativen.**

Das Problem ist dabei nicht, dass der *p*-Wert falsch ist. Es ist, dass er ein schlechtes Instrument ist, eingeschränkt eingesetzt zu dienen. Dazu müssen Daten, muss man allein. Und es kann nicht bestätigt werden, dass der *p*-Wert korrekt ist, erklärt Leoridan Held. „Der *p*-Wert oder andere Falle durchaus eigentlich korrekt, aber er ist nicht optimal für vorhandene Daten.“ In Leonard Held und Peter Hirsch verteidigen sich die Autoren der Plausibel-Anwendung der *p*-Werte lange nicht mehr nur auf die *p*-Werte, sondern auf die *p*-Werte in Bezug auf die Plausibel-Anwendung. Sie fordern, nur Verzerrungsuntersuchungen zu verwenden, die spezifische Aussagen über die Unsicherheit einer Schätzung und nicht über die Güte einer Hypothese für statistische Evidenz – zum Beispiel Bayes-Parameter. Mit Peter Hirsch lautet sich die Wahrnehmung: „Wir sind der Meinung, dass es besser ist, die *p*-Werte abzulehnen, statt dass diese wie beim *p*-Wert nach einem Schwarz-Weiß-Schema angekommen 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?

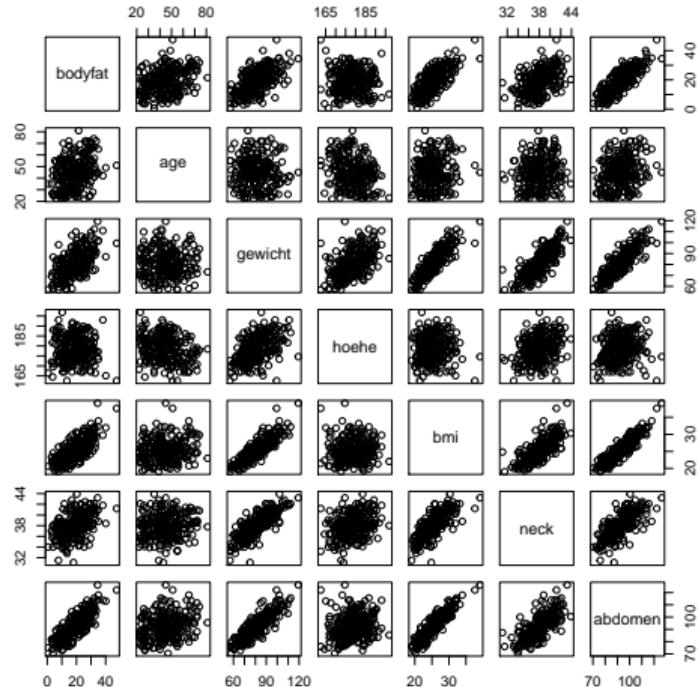
Study with 241 male participants. Measured variable were, among others, body fat (%), age, weight, body size, BMI, neck thickness and abdominal girth.

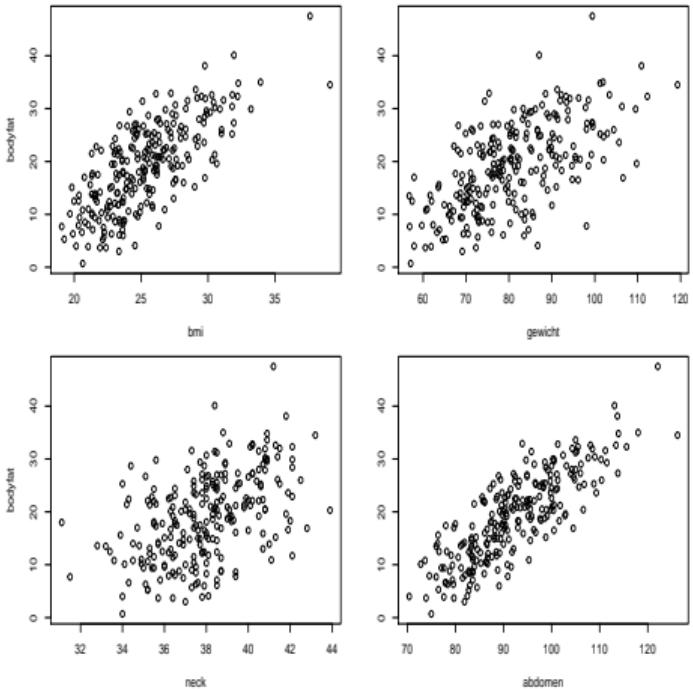
```
glimpse(d.bodyfat)
```

```
## 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.

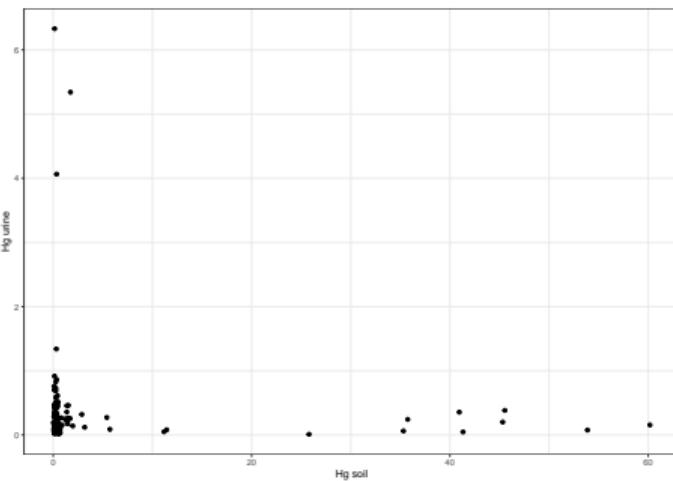
## Data example 2: Mercury (Hg) in Valais (Switzerland)

**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.

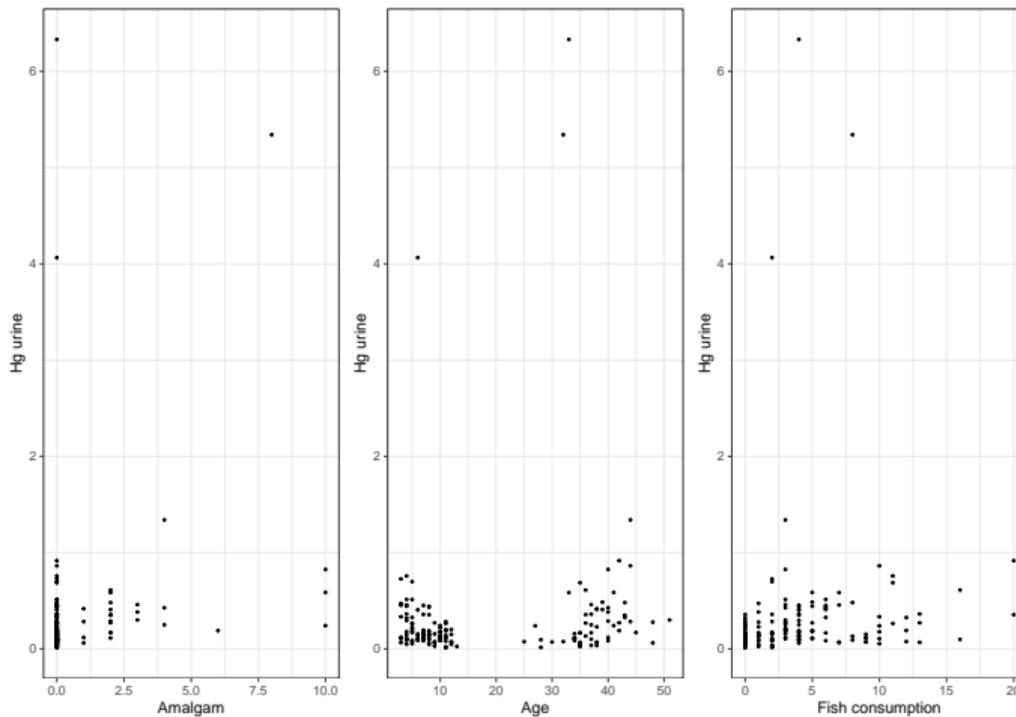
```
glimpse(d.hg)

## # Rows: 156
## # Columns: 10
## # $ Hg_urin      <dbl> 0.30047287, 0.02242110, 0.11347359, 0.44100951, 0.40...
## # $ Hg_soil       <dbl> 0.1174701, 0.3396658, 0.1236580, 0.1477723, 0.221915...
## # $ 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, 0, 2, 0, 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, 0, 0...
## # $ last_time_fish <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## # $ mother         <fct> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0...
```

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

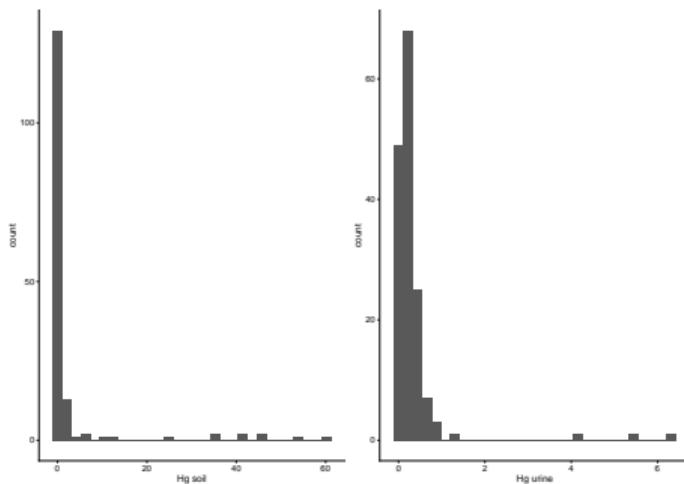


Which other factors might be responsible for high Hg concentrations in urine?



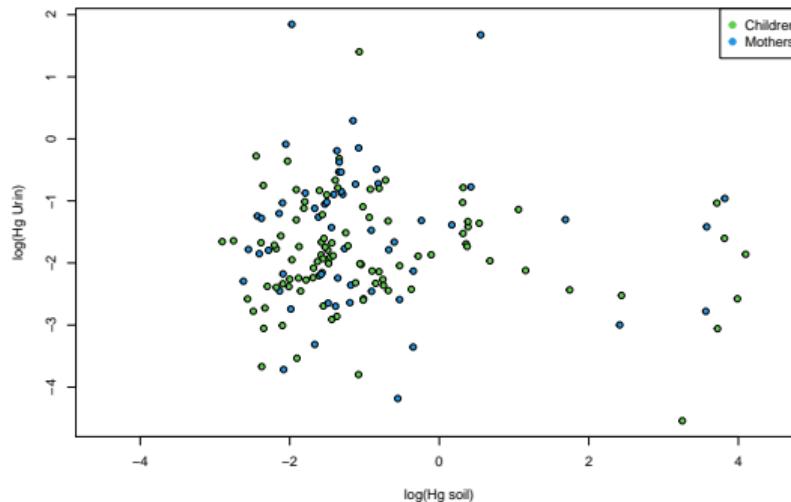
From these plots it is hard to tell which factors exactly influence the Hg pollution in humans.

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.

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



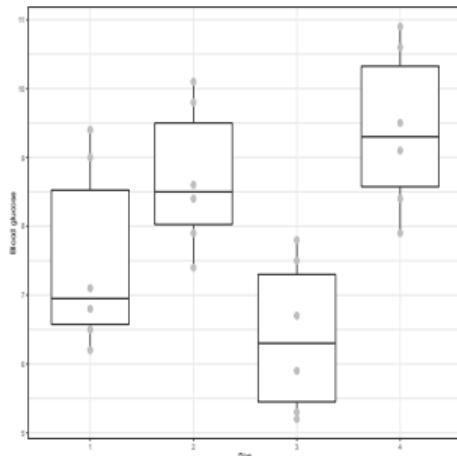
Remember: The idea to log-transform the variables was mainly obvious thanks to **visual inspection!**

## 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  $ESR < 20\text{mm/hr}$ ?

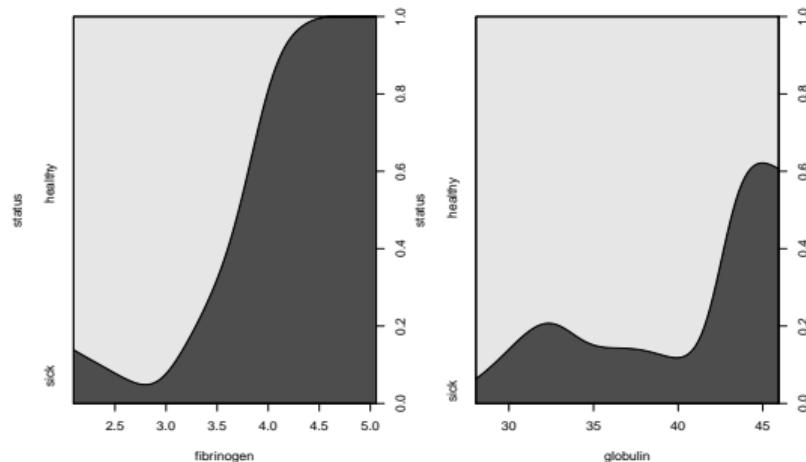
The plasma data come with the package HSAUR3

```
##      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?

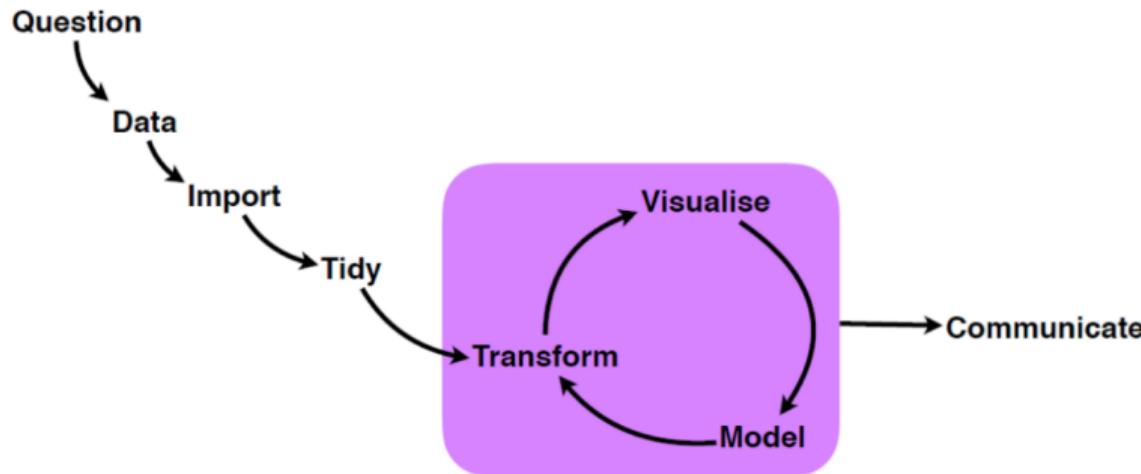
A model is an approximation of the reality. **Understanding how the real world works** is usually only possible thanks to simplifying assumptions.

→ 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.

## 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.