Course Bio144:
Data analysis in biology

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

Week 1: Introduction and Outlook

23./24. February 2017

Organization

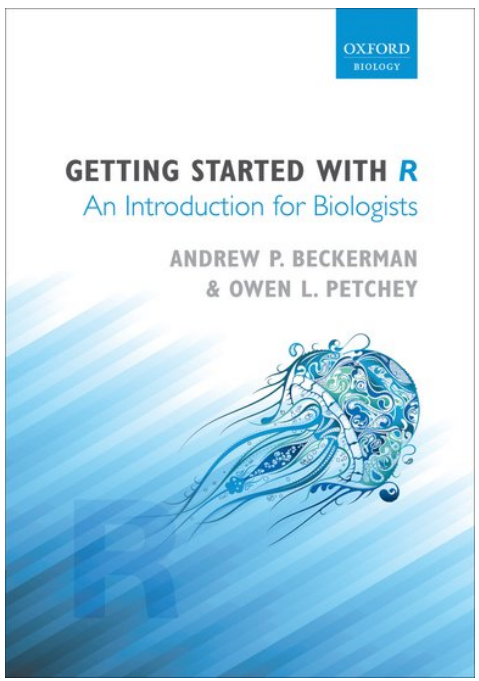
To do:

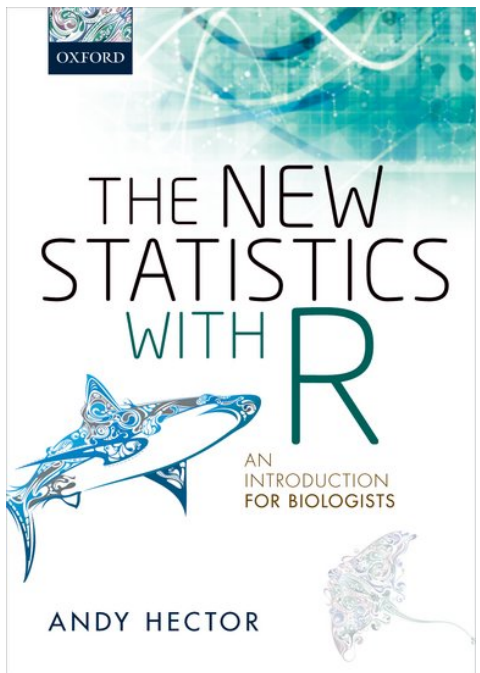
- Testate conditions
- Examination date
- Link to course webpage

Literature

Compulsory literature:

1. *Lineare Regression* by W. Stahel (pdf on course webpage)
2. *Getting Started With R* by A.P. Beckerman und O.L. Petchey, Oxford University Press; ISBN 978-0-19-960162-2
3. *The New Statistics With R* by A. Hector, Oxford University Press; ISBN 978-0-19-872906-8





Complementary literature:

- *Statistics – An Introduction Using R* by M.J. Crawley (similar to 3.) above)
- *The Analysis of Biological Data* by M.C. Whitlock and D. Schluter
- *Regression - Modelle, Methoden und Anwendungen* by Fahrmeier, Kneib, Lang

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!

Prerequisite for Bio144

- Mat183 “Stochastik für die Naturwissenschaften” (2nd semester)

Schedule (14 weeks, 12 lectures)

1. Introduction and outlook
2. Simple linear regression
3. Residual analysis, model validation
4. Multiple linear regression
5. ANOVA and ANCOVA
6. Matrix Algebra
7. Model choice
8. Self-study week
9. Interpretation of the results
10. Binary Data (logistic regression)
11. Count data (Poisson regression)
12. Measurement error, random effects
13. Self-study week
14. Selected topics, repetition and outlook

Short-dated changes possible!

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

What do you think?

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

What do you think?

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: Does a drug have a positive effect? 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?

Be careful! "Learning by doing" is 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 is itself an exciting part of research!

Data analysis is at the interface between mathematics and biology (and other research fields such as medicine, earth sciences, and so on).

What are the purposes of data analysis?

- To find and quantify associations through graphical representations and modelling.
- To draw conclusion from data.
- To quantify the uncertainty of these conclusions.

Own examples

Otter (lutra lutra) (Weinberger et al., 2016)

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.

Biological Conservation 199 (2016) 88–95

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/bioc

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

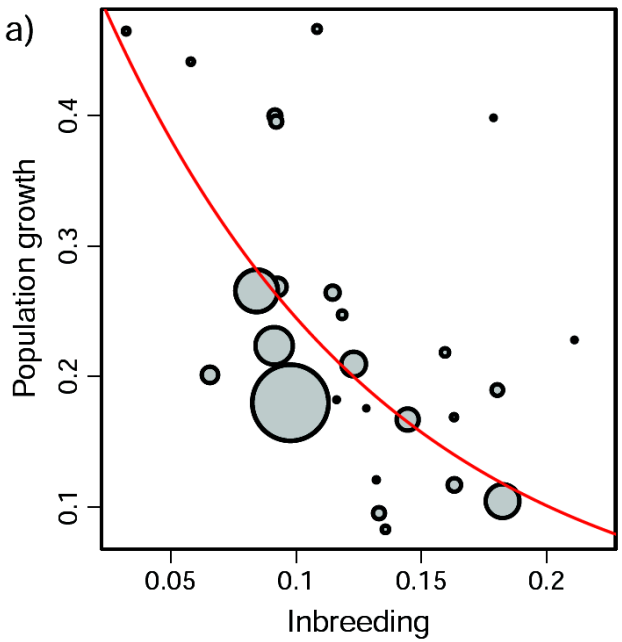
Irene C. Weinberger [a,*], Stefanie Muff [a,b], Addy de Jongh [c], Andreas Kranz [d], Fabio Bontadina [e,f]

[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, Wuhrstr. 12, 8003 Zurich, Switzerland
[f] Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology, 8903 Birmensdorf, Switzerland

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

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,...

▶ Link to Splashy study

Statistics in the news (April 2016)

NZZ am Sonntag 3. April 2016 **Wissen** 61

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**

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

Karriere machen können nicht nur Menschen, sondern auch statistische Grössen. Das gilt besonders für den sogenannten p-Wert, mit dem jeder Mittelschüler und jede Studentin in Kontakt kommt, vor allem aber jeder, der im weitesten Sinn etwas mit Statistik zu tun hat. Inzwischen ist der p-Wert indes auf die schiefe Bahn geraten. Denn was der Vater der modernen Statistik, der britische Genetiker Ronald Fisher, 1920 als eine Art informelles Sensorium für die Aussagekraft von Daten entwickelte, ist in der Praxis oftmals zu einem simplen Lackmustest verkommen.

Ergibt die statistische Analyse von Daten *einen p-Wert < 0,05 (5 Prozent)* oder noch *besser < 0,01 (1 Prozent), gelten diese als signifikant – den Daten wird dann automatisch Beweiskraft zugestanden. Das entscheidet etwa darüber, ob ein neues Medikament* als wirksam eingestuft wird oder ob ein Forscher seine Studie in einem angesehenen Fachblatt publizieren kann. «Der p-Wert war aber nie dazu gedacht, wissenschaftliches Denken ausser Kraft zu setzen», hat sich Ron Wasserstein, der Direktor der Amerikanischen Statistischen Vereinigung (ASA), jüngst öffentlich beklagt.

Wissenschafter verwenden den p-Wert immer häufiger, ohne zu verstehen, was er bedeutet – das fördert schlechte Forschung und untergräbt die Glaubwürdigkeit der Wissenschaft. Der Mediziner und Epidemiologe John Ioannidis von der Universität Stanford sprach in einem Kommentar von «Drogenabhängigen»: «Der falsche Gebrauch des p-Wertes ist dermassen einfach und erfolgt so automatisiert, dass manche süchtig werden danach – vor allem wenn sie mit Forschungsgeldern und Publikationen belohnt werden.» Angesichts der Missstände sah sich die ASA jetzt veranlasst, zum ersten Mal in ihrer fast 180-jährigen Geschichte Empfehlungen zu veröffentlichen, wie man mit einer statistischen Grösse vernünftig umgeht.

Wider die Null-Hypothese

«Der p-Wert sagt nicht das aus, was man gemeinhin von ihm erwartet», erklärt der Berner Epidemiologe Peter Jüni, der seit kurzem am Applied Health Research Centre der Universität Toronto tätig ist. Das bedeutet: Der p-Wert misst nicht die Wahrscheinlichkeit, ob eine bestimmte Hypothese zutrifft, und auch nicht, ob ein bestimmtes Resultat zufällig zustande gekommen ist, wie die ASA festhält. Vielmehr dienen p-Werte dazu, die sogenannte Null-Hypothese zu testen und möglichst zu verwerfen.

Die Hypothese in einer Patientenstudie könnte zum Beispiel lauten, dass ein Medikament A gegen Nierenprobleme besser wirkt als ein Medikament B. Die Null-Hypothese besagt dann genau das Gegenteil davon, nämlich dass das Medikament A nicht besser wirkt als das Medikament B. Beim Test berechnet der Forscher im Prinzip, wie gross die Wahrscheinlichkeit für das Auftreten eines tatsächlich festgestellten oder noch grösseren Unterschieds zwischen den beiden untersuchten Medikamenten ist – unter der Annahme, dass die Null-Hypothese stimmt. Diese Wahrscheinlichkeit ist der p-Wert, und je geringer er ist, desto weniger spricht für die Null-Hypothese. Ein p-Wert von 0,05 bedeutet, dass das festgestellte (oder ein noch extremeres) Resultat unter den Bedingungen der Null-Hypothese mit einer Wahrscheinlichkeit von lediglich 5 Prozent zustande kommen kann – und nicht, dass eine bestimmte Hypothese mit einer Sicherheit von 95 Prozent wahr ist.

«Studien führen heute zu dermassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von p-Werten kommt.»

Über die eigentlich untersuchte Hypothese kann der p-Wert nur indirekt etwas aussagen, weil er über zwei Ecken gedacht ist. Der ominöse Wert liefert also keine ultimativen Beweise für einen postulierten Unterschied oder Zusammenhang. «Der p-Wert ist eine bedingte und nicht eine absolute Wahrscheinlichkeit», erklärt Peter Jüni. «Doch genau das verstehen viele Forscher nicht, und es interessiert sie auch nicht.»

Hinzu kommt, dass die Signifikanzgrenzen von 5 Prozent bzw. 1 Prozent historisch entstanden und keineswegs klar definierte Werte sind. Ronald Fisher, der Erfinder des p-Werts, überliess es jedenfalls ausdrücklich der Interpretation des einzelnen Forschers, ab welcher Grösse ein p-Wert in einer Untersuchung Aussagekraft haben soll. «Trotzdem haben sich die willkürlich gewählten Signifikanzgrenzen ins Gehirn von Generationen von Forschern gebrannt», sagt Leonhard Held vom Institut für Epidemiologie, Biostatistik und Prävention der Universität Zürich.

Die britischen Statistiker Jonathan Sterne und George Davey Smith haben schon vor 15 Jahren im «British Medical Journal» dazu aufgerufen, die Resultate von medizinischen Studien nicht mehr als «signifikant» oder «nichtsignifikant» darzustellen, sondern im Kontext der gesamten Untersuchung und anhand anderer Ergebnisse zu interpretieren. Genützt hat das herzlich wenig, wie das Team von John Ioannidis in einer soeben erschienenen Studie festgestellt hat. Demnach sind in den vergangenen 25 Jahren in der biomedizinischen Forschung immer mehr Studien erschienen, die p-Werte angeben, die zudem immer klarer signifikant ausfallen. Gleichzeitig werden Zusatzinformationen zu den festgestellten Effekten immer seltener («JAMA», Bd. 315, S. 1141).

5%

Kleiner als dieser Wert muss der sogenannte p-Wert in einem statistischen Test sein, dann gelten die Daten aus einer Studie als aussagekräftig. Doch die Grenze ist willkürlich gewählt. (pim.)

Es gibt Alternativen

Das Problem ist dabei nicht nur, dass der p-Wert ein trügerisch einfaches statistisches Instrument ist. «Studien führen heute zu dermassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von p-Werten kommt», erklärt Leonhard Held. «Der eine oder andere fällt dann bestimmt signifikant aus, auch wenn kein Effekt vorhanden ist.» Leonhard Held und Peter Jüni verlassen sich deshalb bei der Planung und Auswertung von Studien schon lange nicht mehr nur auf die p-Werte.

Jüni empfiehlt, die Resultate von Studien mindestens mit Vertrauensintervallen zu versehen, die spezifische Aussagen über die Unsicherheit einer Schätzung machen. Und Held erforscht alternative Masse für statistische Evidenz – zum Beispiel Bayes-Faktoren. Mit ihrer Hilfe lässt sich die Wahrscheinlichkeit einer Hypothese anhand der Daten anpassen, statt dass diese wie beim p-Wert nach einem Schwarz-Weiss-Schema angenommen oder abgelehnt wird.

https://www.theguardian.com/science/2016/jul/17/politicians-dodgy-statistics-tricks-guide?&tc=eml

Data example 1: Prognostic factors for body fat

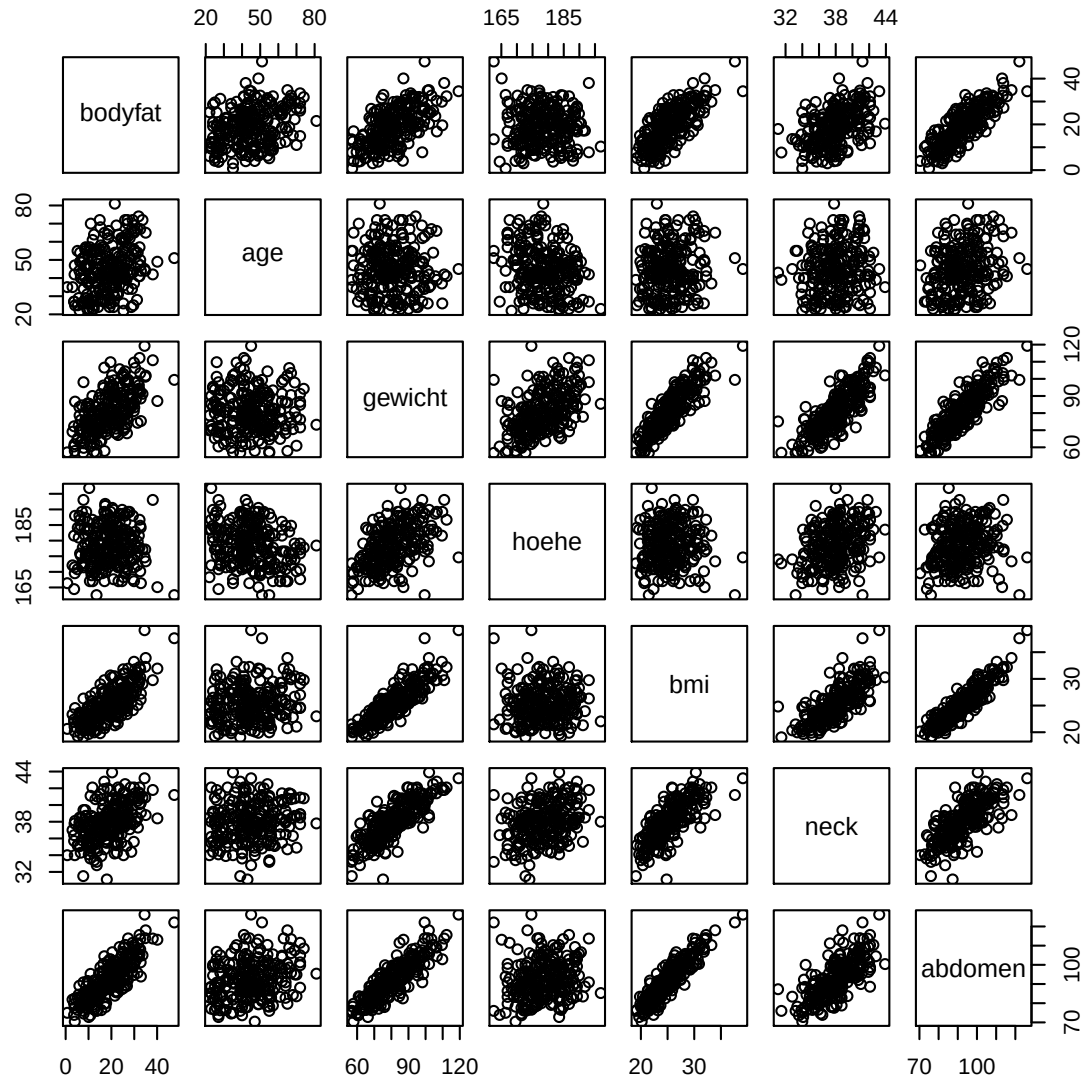
(From Theo Gasser & Burkhardt Seifert *Grundbegriffe der Biostatistik*)

Body fat is an important indicator for overweight, but difficult to measure.
Question: Which factors allow for precise estimation (prediction) of body fat?

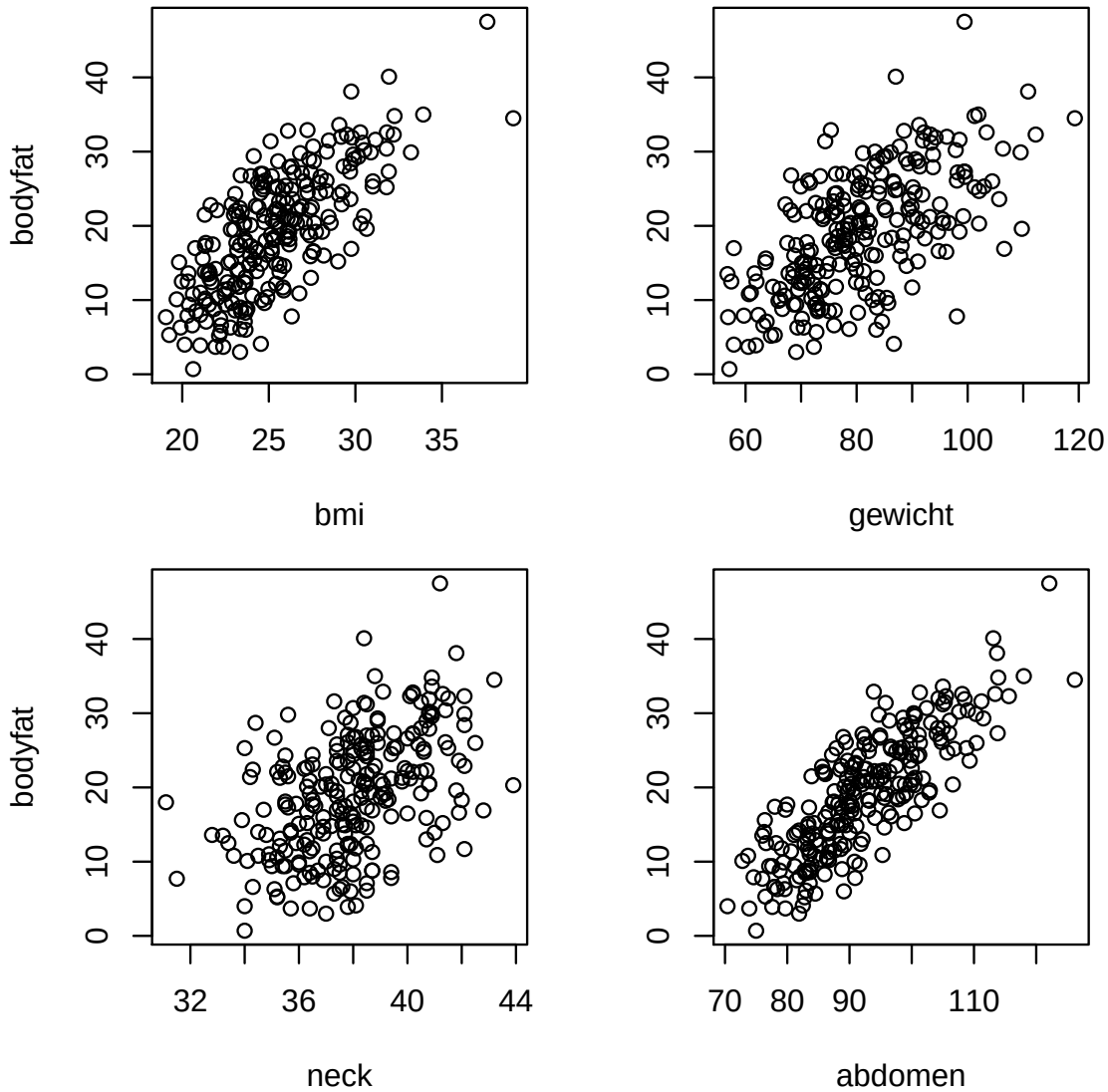
Study with 241 male participants. Measured variable were, among others, body fat (%), age, weight, body size, BMI, neck thickness and abdominal girth.

```
> str(d.bodyfat)

'data.frame':        243 obs. of  7 variables:
 $ bodyfat: num  12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...
 $ age    : int  23 22 22 26 24 24 26 25 25 23 ...
 $ gewicht: num  70 78.7 69.9 83.9 83.7 ...
 $ hoehe  : num  172 184 168 184 181 ...
 $ bmi    : num  23.6 23.4 24.7 24.9 25.5 ...
 $ neck   : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
 $ abdomen: num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
```

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



`pairs()` gives us the scatterplots of all against all variables.



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

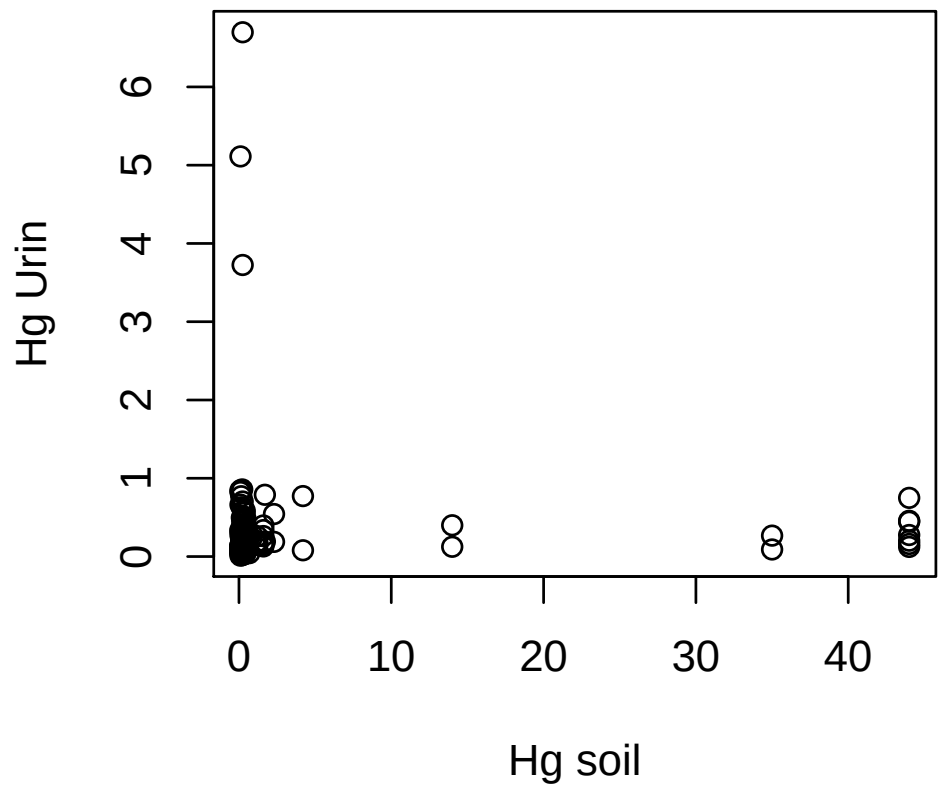
Data example 2: Mercury in Valais (Switzerland)

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

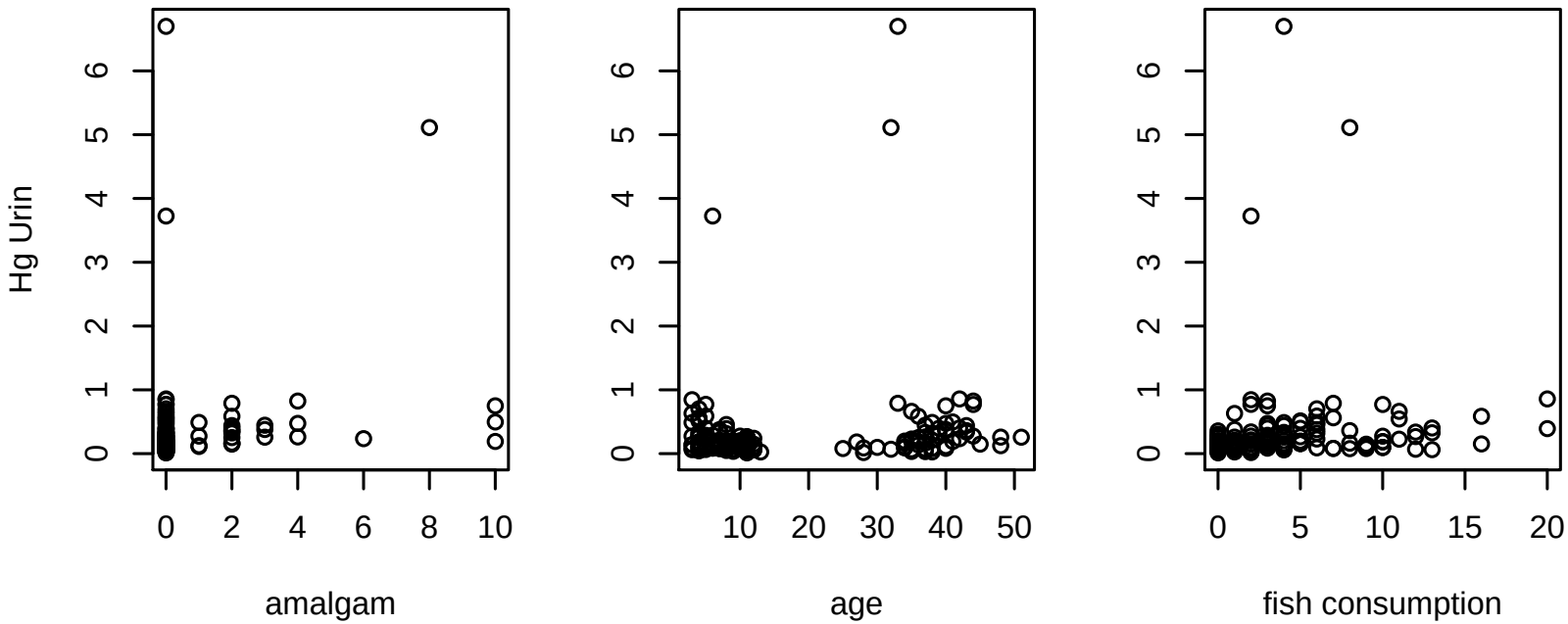
```
> str(d.hg)

'data.frame':        156 obs. of  10 variables:
 $ Hg_urin       : num  0.258 0.036 0.16 0.314 0.29 ...
 $ Hg_soil       : num  0.49 0.42 0.18 0.49 0.24 0.2 0.1 14 0.1 0.3 ...
 $ veg_garden    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ migration     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ smoking       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ amalgam       : int  3 0 2 0 0 0 0 1 0 0 ...
 $ age           : int  51 11 34 8 6 40 7 48 11 38 ...
 $ fish          : int  3 2 5 4 4 2 2 4 0 7 ...
 $ last_time_fish: int  0 0 0 0 0 0 0 0 0 0 ...
 $ mother        : Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 1 2 ...
```

A first visual inspection is not very informative. There is no association visible by eye:

```
> plot(Hg_urin ~ Hg_soil, data=d.hg, xlab="Hg soil", ylab = "Hg Urin")
```



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

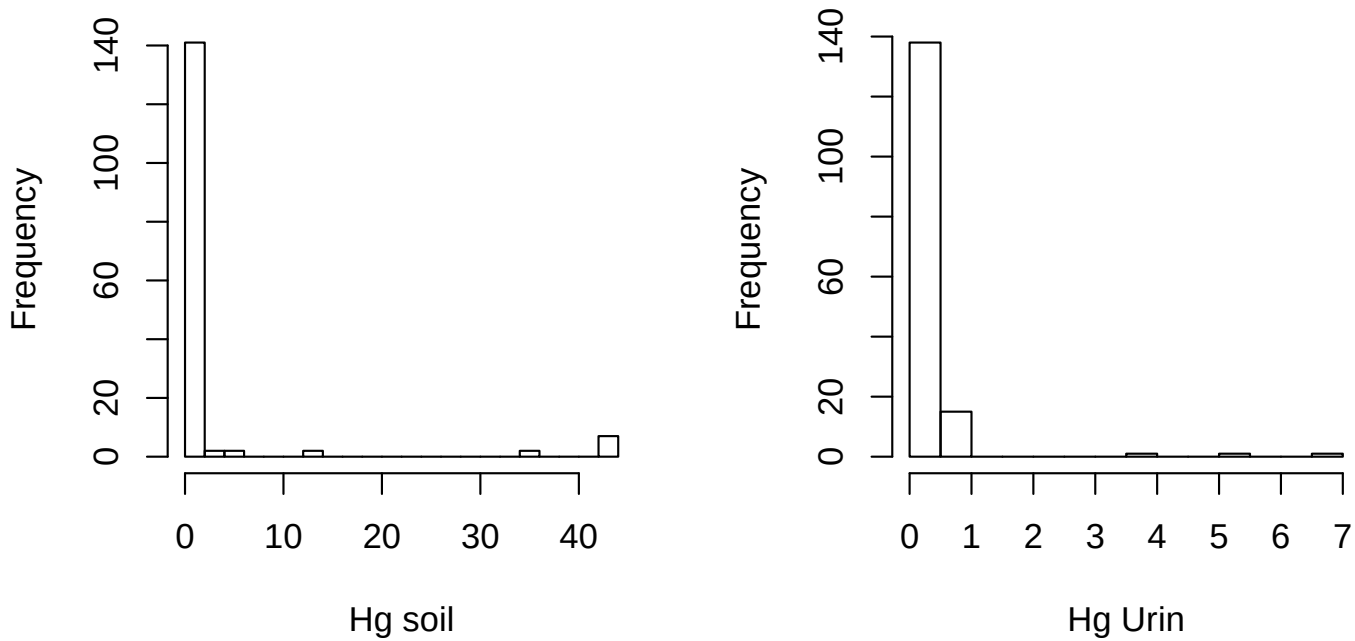


From these plots it is hard to tell which factors exactly influence the Hg pollution in humans.

It is always useful to look at the distribution of the variables in the model.

Let us plot the histogram of Hg concentrations:

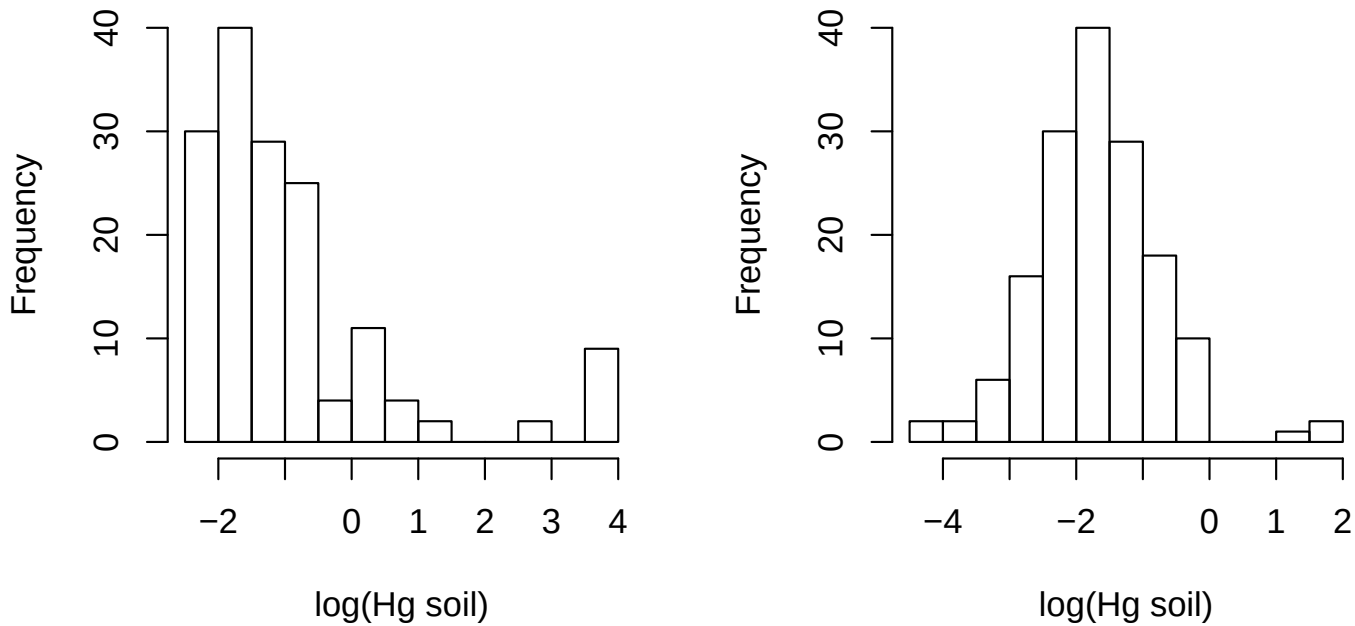
```
> par(mfrow=c(1,2))
> hist(d.hg$Hg_soil,xlab="Hg soil",nclass=20,main="")
> hist(d.hg$Hg_urin,xlab="Hg Urin",nclass=20,main="")
```



All Hg values seem to "stick" at 0.

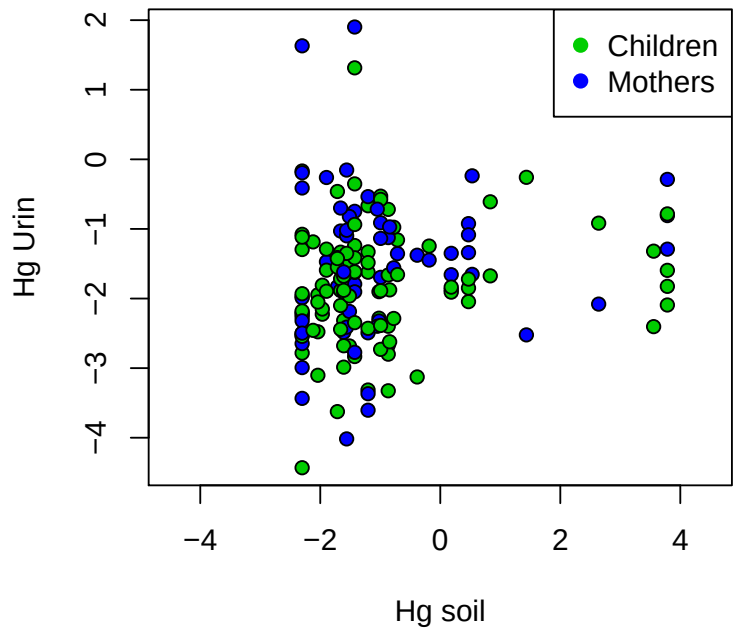
In such cases in can help to *log-transform* the respective variables.

```
> par(mfrow=c(1,2))
> hist(log(d.hg$Hg_soil),xlab="log(Hg soil)",nclass=20,main="")
> hist(log(d.hg$Hg_urin),xlab="log(Hg soil)",nclass=20,main="")
```



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

```
> plot(log(Hg_urin) ~ log(Hg_soil), data=d.hg, xlab="Hg soil",
+      ylab = "Hg Urin",pch=21,bg=as.numeric(mother)+2,xlim=c(-4.5,4.5))
> legend("topright",legend=c("Children","Mothers"),col=c(3,4),pch=21,pt.bg=c(3,4))
```



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

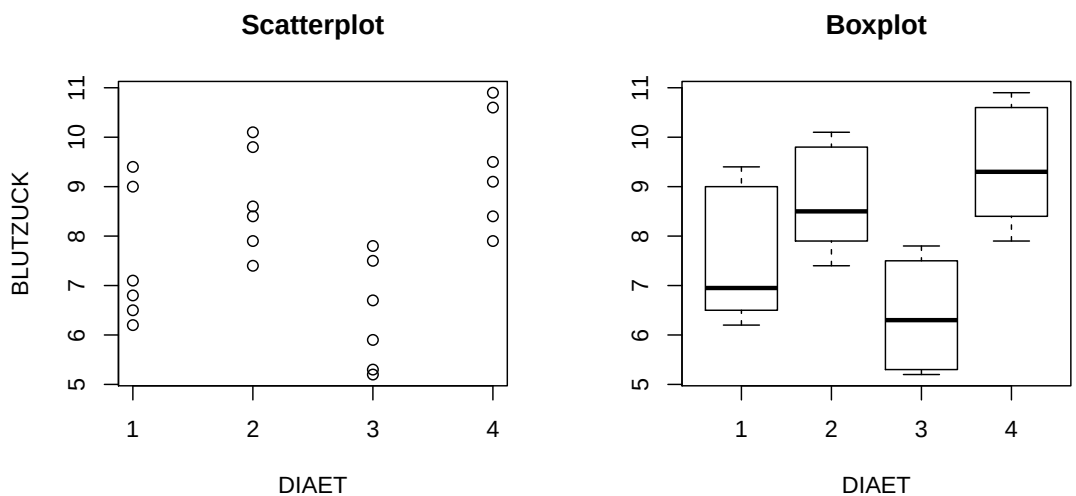
Data example 3: Diet and blood glucose level

(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?

```
> par(mfrow=c(1,2))
> plot(BLUTZUCK ~ DIAET,d.blz,xaxt="n",main="Scatterplot")
> axis(1,1:4)
> boxplot(BLUTZUCK ~ DIAET,d.blz,xaxt="n",xlab="DIAET",main="Boxplot")
> axis(1,1:4)
```



Does this question seem familiar to you?
Hint: what would you do for two groups?

For more than 2 groups we need the *ANOVA* (=ANalysis Of VAriance) approach.

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

The next question then is: which diets are *pairwise* different.

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$< 20 mm/hr$ and the concentrations of the plasma proteins Fibrinogen and Globulin?

Load data from the package that comes with Hothorn and Everitt (2014):

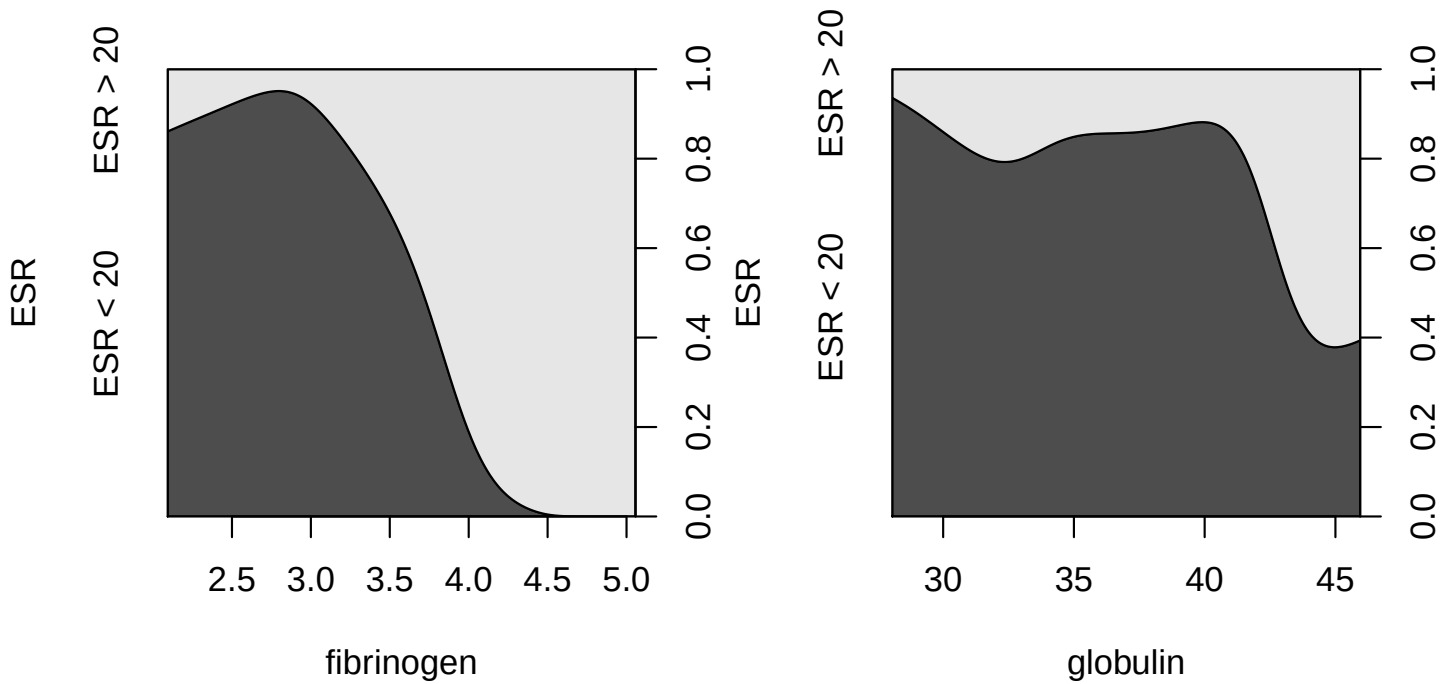
```
> library(HSAUR3)
> data("plasma",package="HSAUR3")
> plasma[c(1,5,9,10,15,29),]
```

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

The distinction ESR$<20mm/hr$ vs. ESR$\geq 20mm/hr$ leads to a binary variable.

The relation between the plasmaprotein levels and the binary indicator can be captured by a *conditional density plot*.

```
> par(mfrow=c(1,2))
> cdplot(ESR ~ fibrinogen,plasma)
> cdplot(ESR ~ 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 oft-cited remark that "all models are wrong, but some are useful" (Box, 1979).

Steps in a modelling process

1. Formulate a precise question
2. Plan your inquiry and the analysis of your data, collect the data (experiments or surveys).
3. Tidy and clean the data
4. Graphical representation of the data
5. Choose an appropriate *model*
6. Estimate model parameters and uncertainties
7. Check modelling assumptions
8. If needed, improve the model; back to step 7
9. Interpret your results and compare to step 1
10. Communicate results precisely and carefully (publication, articles..)

The scopes of statistical data analysis

a) Prediction, interpolation. Example body fat: use substitute measurements to predict body fat of a person.

b) Estimation of parameters.

c) Determine important variables. Example physical activity of children: The study aims to find factors that (positively or negatively) influence the movement behavior of children.

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 analyze all data examples introduced here (and of course many more), as well as to draw conclusions from them.

Graphical representation of data

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

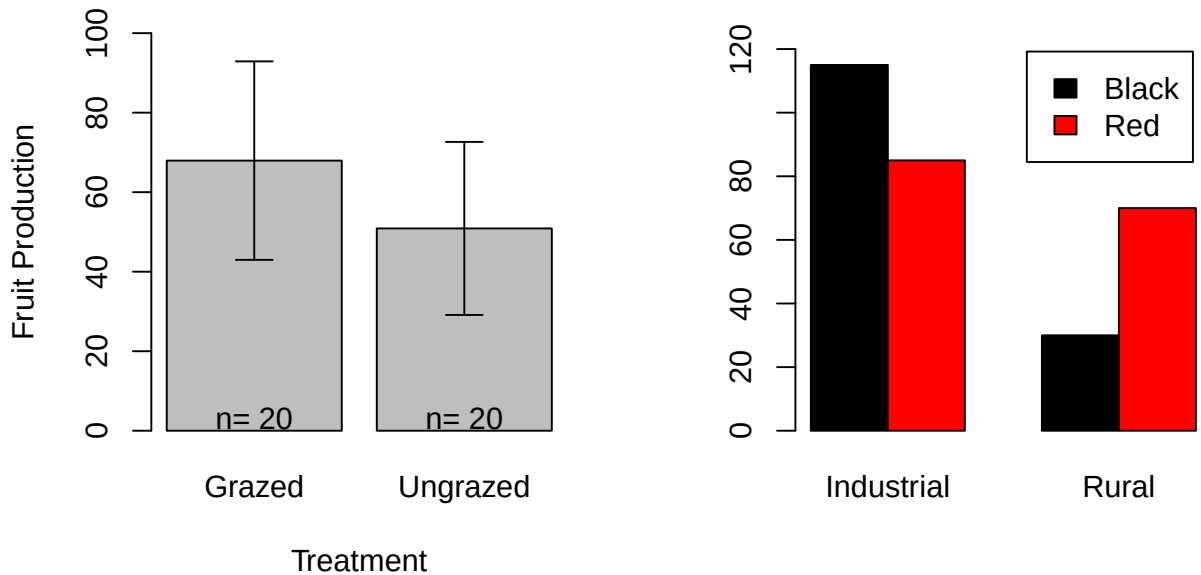
Representation	Useful for
Scatterplots	Pairwise dependency of continuous variables.
Histograms	Distribution of numerical variables.
Boxplots	Distribution of numerical variables, ev. conditionally on categories.
Conditional density plots	Dependency of a binary variable from a continuous variable.
Barplots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

Barplots

Examples from Beckerman & Petchey:

Left: Fruit production in grazed and ungrazed areas. Grazing reduces above-ground biomass. How is fruit production affected by this?

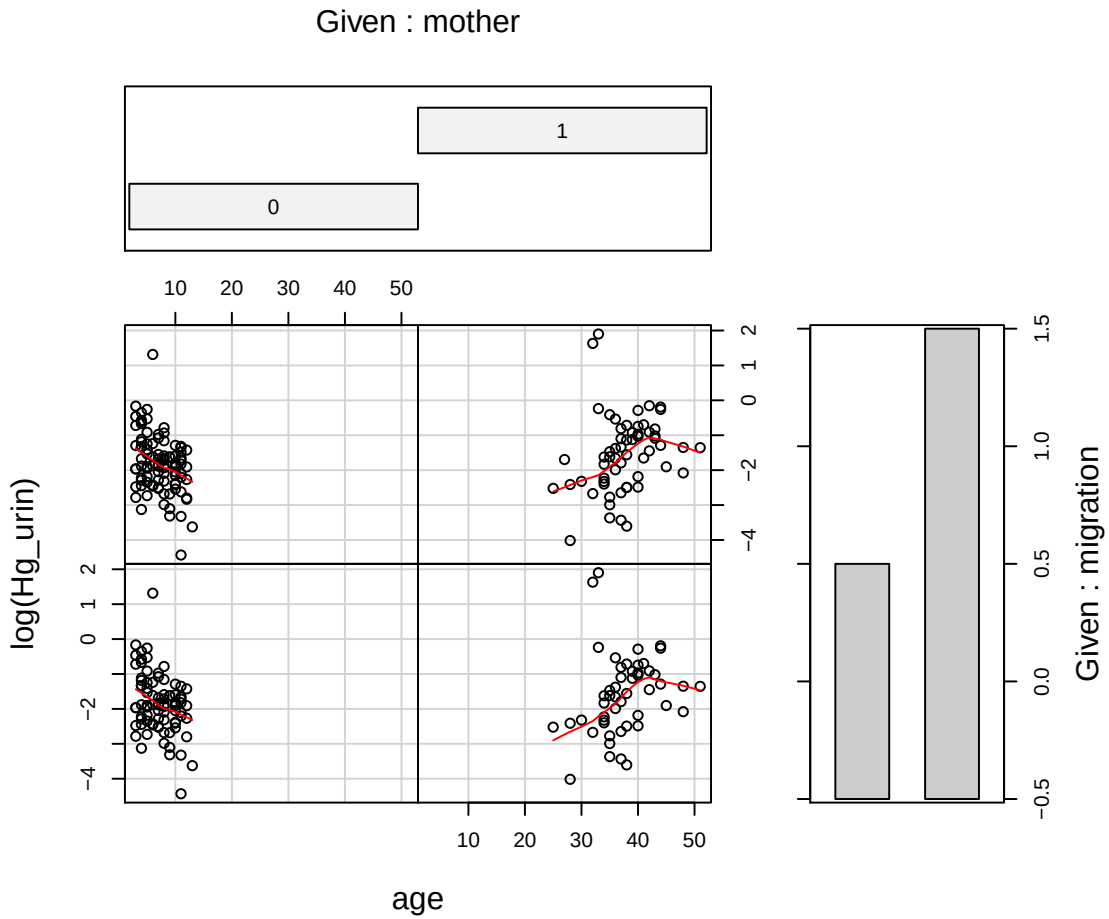
Right: Number of birds of certain colors in two environments.



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 (see census app).

Experimental vs observational data

To do, ev mention only later in weeks 7/8

Next week: Simple linear regression

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 Übungsbuchder angewandten Statistik*.

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

Weinberger, I. C., S. Muff, A. Kranz, and F. Bontadina (2016). Flexible habitat selection paves the way for a recovery of otter populations in the European Alps. *Biological Conservation 199*, 88–95.