

# Kurs Bio144: Datenanalyse in der Biologie

Stefanie Muff (Vorlesung) & Owen L. Petchey (Praktikum)

1. Woche: Einführung und Ausblick  
23./24. Februar 2017

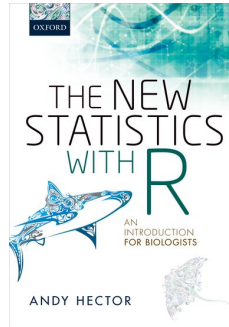
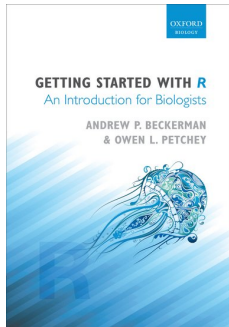
# Organisatorisches

Gebe hier genaue Testatbedingungen, Prüfungsdaten (falls bekannt);  
OpenEdx Kursseite (Link). Studis müssen sich da anmelden.

# Lehrmittel und Literatur

Obligatorische Lehrmittel:

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



## Ergänzende Literatur:

- *Statistics – An Introduction Using R* von M.J. Crawley (ähnlich wie 3.) oben)
- *The Analysis of Biological Data* von M.C. Whitlock und D. Schluter
- *Regression - Modelle, Methoden und Anwendungen* von Fahrmeier, Kneib, Lang

# Ziele dieses Kurses

- Solides Fundement an statistischen Methoden erarbeiten, um biologische oder biomedizinische Fragen mit Daten quantitativ und objektiv zu beantworten.
- Fähigkeit vermitteln, Resultate in Forschungsartikeln zu verstehen, zu interpretieren und evtl. kritisch zu hinterfragen.
- Die *Sprache* des Statistikers verstehen lernen.
- Wir möchten Euch eine herausfordernde, spannende und freudvolle Lernerfahrung geben. **Etwas, was man wirklich brauchen kann und darum Spass macht.**

Meine Überzeugung: Fundierte Kenntnisse in Statistik machen Sie unabhängig!

# Voraussetzungen für Bio144

- Mat183 Stochastik für die Naturwissenschaften

# Kurs-Fahrplan (12 Wochen)

- |  |  |
|--|--|
| 1. Einführung und Ausblick               | 7. Modellwahl  |
| 2. Einfache lineare Regression           | 8. Interpretation von Resultaten                       |
| 3. Residuenanalyse,<br>Modellvalidierung | 9. Binäre Daten (logistische<br>Regression)            |
| 4. Multiple lineare Regression           | 10. Zähldaten (Poisson Regression)                     |
| 5. ANOVA und ANCOVA                      | 11. Messfehler, zufällige Effekte                      |
| 6. Matrix Algebra                        | 12. Ausgewählte Themen,<br>Wiederholungen und Ausblick |

**Kurzfristige Anpassungen vorbehalten!**

# Warum ist Statistik für die Biologie und Medizin so wichtig?

Was denken Sie?



# Warum ist Statistik für die Biologie und Medizin so wichtig?

Was denken Sie?

Erkenntnis, dass ohne Kenntnisse in statistischer Datenanalyse eigene Daten in Bachelor-, Master- oder Doktorarbeiten nicht ausgewertet werden können.

Beispiele:

- Medizin: Hat ein bestimmtes Medikament eine Wirkung? Welche Faktoren führen zu Krebs?
- Oekologie: Was für einen Lebensraum braucht ein Tier zum Leben? Was bevorzugt es?
- Evolutionsbiologie: Haben Tiere mit hohem Inzuchtgrad schlechtere Chancen zu überleben oder sich fortzupflanzen?

Achtung! "Learning by doing" ist in der Statistik praktisch unmöglich. Es braucht viel Erfahrung, es gibt sehr viele Fallstricke.

Wer ein gutes Grundlagenwissen in Statistik hat, kann unabhängiger arbeiten. Wer sich nicht auskennt, ist immer auf die Hilfe anderer Leute angewiesen...

Datenanalyse ist selber ein spannender Teil der Forschung!

Datenanalyse ist die Schnittstelle zwischen Mathematik und Biologie (oder anderen Forschungsfeldern, z.B. Medizin, Geographie etc.).

# Was leistet die Datenanalyse?

- Auffinden und Quantifizierung von Zusammenhängen durch graphische Darstellung und Modellierung.
- Aus Daten gültige Schlussfolgerungen ziehen.
- Die Unsicherheit der Schlussfolgerung quantifizieren.

# Eigene Beispiele

**Fischotter** (Weinberger et al., 2016)

*Forschungsfrage:* Welche Lebensräume werden von den Fischottern bevorzugt?

*Methode:* Studie in Österreich, 9 Otter mit Radiotelemetriesendern versehen und während 2-3 Jahren überwacht.

Biological Conservation 199 (2016) 88–95



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Biological Conservation

journal homepage: [www.elsevier.com/locate/bioco](http://www.elsevier.com/locate/bioco)



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

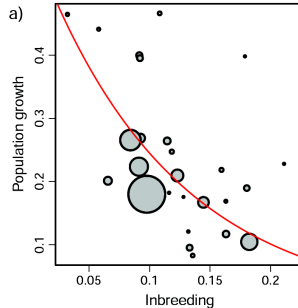
<sup>e</sup> SWILD – Urban Ecology & Wildlife Research, Wuhstr. 12, 8003 Zurich, Switzerland

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

## Inzucht bei Steinböcken

*Forschungsfrage:* Hat Inzucht in Steinbockpopulationen eine negative Auswirkung auf das Langzeit-Populationswachstum? Inzuchtdepression!

*Methoden:* Genetische Information aus Blutproben gibt Aufschluss über Inzucht der Steinböcke. Langzeit-Monitoring von Populationsgrößen und Abschussquoten.



### 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

*Forschungsfrage:* Gibt es einen Zusammenhang zwischen Quecksilber(Hg)-Bodenwerten von Wohnhäusern und der Hg-Belastung im Körper (Urin, Haar) der Bewohner?

*Methode:* Bodenmessungen auf den Grundstücken, sowie Messungen und Befragungen von Kindern und deren Müttern.

Hoch brisante, politisch aufgeladene Fragestellung!

► Schweiz Aktuell, 20. Juni 2016

## Bewegungsverhalten bei Kindern

*Forschungsfrage:* Welche Einflussfaktoren beeinflussen das Bewegungsverhalten von 3-5 jährigen Kindern?

*Methode:* Die Kinder werden während mehreren Tagen mit Bewegungsmessern ausgestattet. Die Eltern müssen mehrmals einen detaillierten Fragebogen ausfüllen.

Erfasste Variablen sind z.B. Medienkonsum, Verhalten der Eltern, Gewicht, Alter,...

► [Link zur Splashy Studie](#)

NZZ Donnerstag 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**

**K**ritiker machen keinen Unterschied zwischen Menschen, sondern auch statistischen Größen. Das gilt besonders für den sogenannten  $p$ -Wert, ein den jeder Mediziner und viele Statistiker in Kontakt kommt, vor allem aber jeder, der im weitesten Sinne mit Statistik zu tun hat. Insbesondere der  $p$ -Wert indes auf die sichere Bahn geraten. Denn was der Vater der modernen Statistik, der britische Genetiker Ronald Fisher, 1925 als eine Art informelles Kriterium für die Aussagekraft von Daten entwickelte, ist in der Praxis oftmals zu einem simplen Lackmeyer verkommen.

Dreht die statistische Analyse von Daten einen  $p$ -Wert (0,05) oder auch besser «0,01 (1 Prozent), gehen diese als signifikant» aus. Dann wird dann automatisch als signifikant» angesehen. Das erhebt der etwas darüber, ob ein neues Medikament als wirksam eingestuft wird oder als Testsubstanz. In einem angeblichen Fachbuche publizieren kann. «Der  $p$ -Wert war aber nie dazu gedacht, wissenschaftliches Denken ausser Kraft zu setzen, bis sich dann Wasserstein, der Direktor der amerikanischen Statistischen Vereinigung (ASA), jüngst öffentlich beklagt.

Wissenschaftler verwenden den  $p$ -Wert immer häufiger, ohne zu verstehen, was er bedeutet – das fordert wichtige Forschung und untergräbt die Glaubwürdigkeit der Wissenschaft. Der Mediziner und Epidemiologe John Ioannidis von der Universität Stanford gewährt in einem Kommentar von «drogenabhängigen» über falsche Gebrauch des  $p$ -Wertes ist demnach einfach und erfolgt so entstehend, dass manche sich

wenden danach: «vor allem wenn sie mit Forschungsgeldern und Publikationen beehrt werden.» Angesichts der Missstände sah auch die ASA jetzt veranlasst, zum ersten Mal in ihrer fast 180-jährigen Geschichte Empfehlungen zu veröffentlichten, wie man mit einer statistischen Grösse umzugehen umgeht.

### Wider die Null-Hypothese

«Der  $p$ -Wert sagt nicht das aus, was man gewöhnlich von ihm erwartet», erklärt der Berner Epidemiologe Peter Künzli, der seit Jahren am Appleton Health Research Center 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 fordert. Vielmehr misst er die Wahrscheinlichkeit, dass ein bestimmtes Resultat zu finden ist, wenn die Null-Hypothese zutrifft, dass es keine Unterschiede zwischen den Gruppen gibt.

Die Hypothese in ihrer Patiententests könnte zum Beispiel lauten, dass ein Medikament A gegen Herz-Kreislauferkrankungen wirkt, während das Medikament B. Die Null-Hypothese besagt dann genau das Gegenteil: dass beide Medikamente gleich wirksam sind. Ronald Fisher, der Erfinder des  $p$ -Wertes, beschrieb die Forscher im Prinzip, wie gross die Wahrscheinlichkeit für die Aufnahme

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

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

**5%**  
kleiner als der sogenannte  $p$ -Wert in einem statistischen Test sein, dann gelten die Daten aus einer Studie als aussagekräftig. Doch die Grenze ist willkürlich gewählt. (jein.)

von einem tatsächlich festgestellten oder noch grossen Unterschied zwischen den beiden untersuchten Medikamenten ist – unter der Annahme, dass die Null-Hypothese zutrifft. 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 Resultat aus noch kleinerem Resultat unter den Bedingungen der Null-Hypothese mit einer Wahrscheinlichkeit von lediglich 5 Prozent zustande kommen könnte – und umgekehrt, dass eine bestimmte Hypothese mit einer Sicherheit von 95 Prozent wahr ist. Über die eigentlich untersuchte Hypothese kann der  $p$ -Wert nur indirekt etwas aussagen, weil er eben zwei Seiten goldene ist. Der zentrale Wert liefert also keine alternativen Beweise für einen positiven Unterschied oder Zusammenhang. «Der  $p$ -Wert ist eine bedingte und nicht eine absolute Wahrscheinlichkeit», erklärt Peter Künzli. «Deshalb kann das Verhalten viele Forscher nicht, und es interessiert sie auch nicht.» Künzli kommt, dass die Signifikanzgrenzen von 5 Prozent bzw. 1 Prozent historisch entstanden sind, aber niemals klar definierte Werte sind. Ronald Fisher, der Erfinder des  $p$ -Wertes, beschrieb in jeder Fall ausnahmslos die Interpretation des erprobten Falsches, ab welcher Grösse ein  $p$ -Wert in einer Untersuchung Aussagekraft haben soll. «Ertraben haben sich die willkürlich gewählten Signifikanzgrenzen im Gehirn von Generationen von Forschern gebildet», sagt Leonhard Held vom Institut für Epidemiologie, Biostatistik und Prävention der Universität Zürich. Die britischen Statistiker Jonathan Sterne und George Dave Smith haben schon vor 15 Jahren im «British Medical Journal» dazu

aufgerufen, die Resultate von medizinischen Studien in 10 Teile als signifikant oder nichtsignifikant zu unterteilen, sondern im Kontext der gesamten Untersuchung und anhand anderer Ergebnisse zu interpretieren. Gezeigt hat das herkömmliche, wie das Team von John Ioannidis in einer neuen einschlägigen Studie festgestellt hat. Demnach sind in den vergangenen 15 Jahren in der biomedizinischen Forschung immer mehr Studien erschienen, die  $p$ -Werte angaben, die nicht immer klar signifikant ausfielen. Gleichzeitig wuchs der Zusatzenformierung zu den festgelegten Effekten immer stärker («JAMA», Bd. 315, S. 334).

### Es gibt Alternativen

Das Problem ist dabei nicht nur, dass der  $p$ -Wert ein eigentlich einfaches statistisches Instrument ist. «Strichen führen kann zu demassen vielen Daten, dass man allen Unfug testen kann und so zu Hunderten von  $p$ -Werten kommt», erklärt Leonhard Held. «Ob eine oder andere Fall dann bestimmt signifikant aus, auch wenn kein Effekt vorhanden ist.» Leonhard Held und Peter Künzli verweisen sich deshalb bei der Planung und Auswertung von Studien schon lange nicht mehr nur auf  $p$ -Werte.

Jetzt empfiehlt, die Resultate von Studien mindestens mit Vertrauensintervallen zu versehen, die spezifische Aussagen über die Unsicherheit einer Schätzung machen. Und Held empfiehlt alternative Mass für statistische Evidenz – zum Beispiel Bayes-Faktoren. Mit diesen Hilfe lässt sich die Wahrscheinlichkeit einer Hypothese anhand der Daten arguieren, statt dass wie beim  $p$ -Wert nach einem Schwellen-Wert-Schema angenommen oder abgelehnt wird.



# Beispiel 1: Prognostische Faktoren für Körperfett

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

Körperfett ist ein wichtiger Indikator für Übergewicht, aber schwer zu messen.

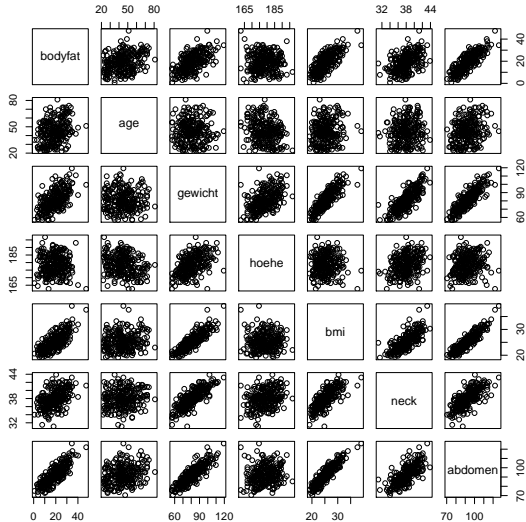
**Frage:** welche Faktoren erlauben eine gute Schätzung des Körperfetts?

Studie mit 241 Männern, von welchen der Körperfett-Anteil (in %) und andere Variablen wie Alter, Gewicht, Körpergrösse, BMI, Nackenfett und Bauchumfang gemessen wurden.

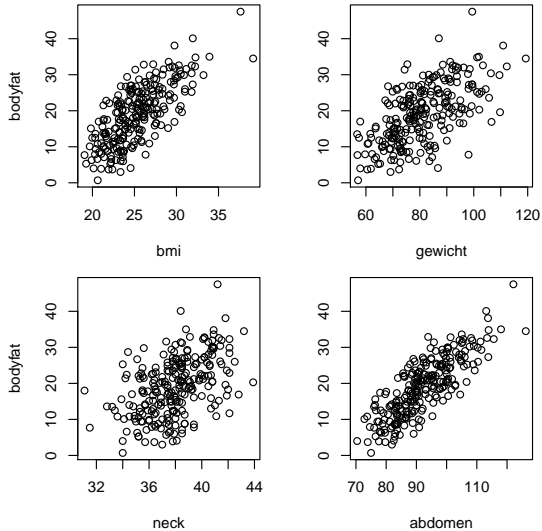
```
> str(d.bodyfat)
```

```
'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()` liefert die Streudiagramme (scatterplots) von allen Variablen gegen alle.



Gesucht ist ein *Modell*, welches das Körperfett aus einfach zu messenden Größen möglichst genau vorhersagen kann.

## Beispiel 2: Quecksilber im Wallis

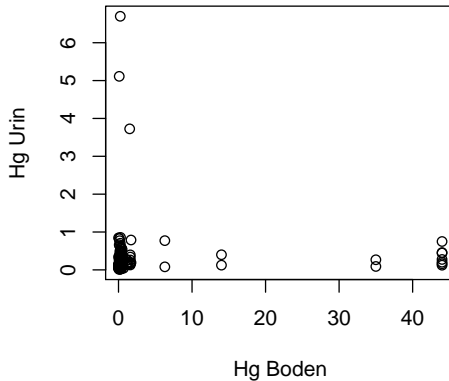
**Frage:** Zusammenhang zwischen Hg-Werten im Boden und Werten im Urin? Wir verwenden hier ein leicht modifiziertes Datenset.

```
> str(d.hg)
```

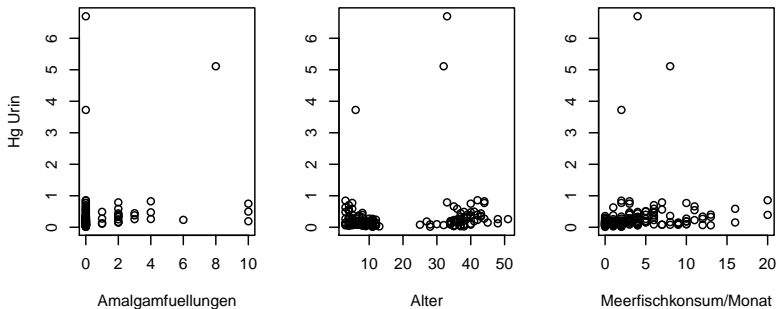
```
'data.frame':      156 obs. of  10 variables:
 $ Hg_urin      : num  0.258 0.036 0.16 0.314 0.29 ...
 $ Hg_soil      : num  0.49 0.42 0.18 0.49 0.24 0.2 0.1 14 0.1 0.3 ...
 $ veg_garden   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ migration    : int  0 0 0 0 0 1 1 0 0 0 ...
 $ smoking      : int  0 0 0 0 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 ...
```

Erste visuelle Inspektion ist nicht sehr informativ. Es ist kein Zusammenhang von Auge ersichtlich:

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



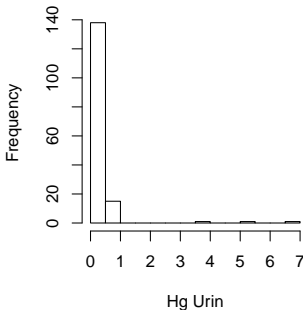
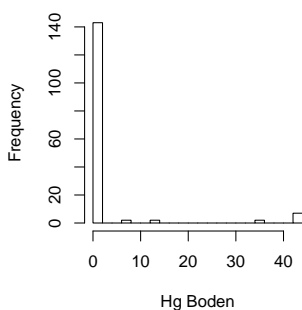
Haben andere Faktoren einen Einfluss auf Hg im Urin?



Aus diesen Grafiken ist es sehr schwer zu sagen, welche Faktoren die Quecksilberbelastung im Menschen genau beeinflussen.

Es ist immer nützlich, die Verteilungen der Variablen im Modell anzuschauen. Zeichnen wir mal das Histogramm der Quecksilberwerte:

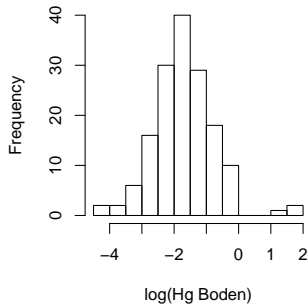
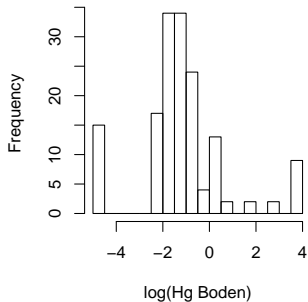
```
> par(mfrow=c(1,2))  
> hist(d.hg$Hg_soil,xlab="Hg Boden",nclass=20,main="")  
> hist(d.hg$Hg_urin,xlab="Hg Urin",nclass=20,main="")
```



Es zeigt sich: fast alle Hg-Werte “kleben” bei 0.

In solchen Fällen kann es helfen, die Variable zu *logarithmieren*.

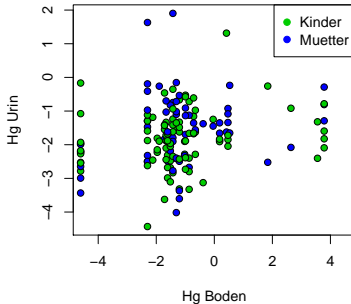
```
> par(mfrow=c(1,2))  
> hist(log(d.hg$Hg_soil),xlab="log(Hg Boden)",nclass=20,main="")  
> hist(log(d.hg$Hg_urin),xlab="log(Hg Boden)",nclass=20,main="")
```





Mit logarithmierten Werten sieht auch das Streudiagramm etwas sinnvoller aus:

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



Merke: Auf die Idee, die Variablen zu logarithmieren, sind wir nur dank visueller Inspektion gekommen.

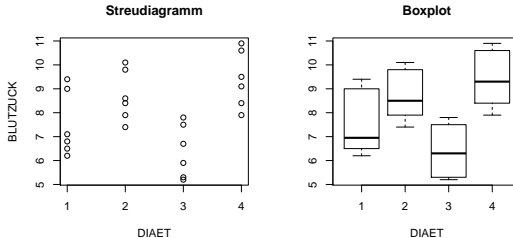
## Beispiel 3: Ernährung und Blutzucker

(Elpelt and Hartung, 1987, p. 190)

24 Personen werden in 4 Gruppen unterteilt. Jede Gruppe erhält eine andere Diät (DIAET). Es werden zu Beginn und am Ende (nach 2 Wochen) die Blutzuckerwerte gemessen. Die Differenz wird gespeichert (BLUTZUCK).

**Frage:** Unterscheiden sich die Gruppen in der Veränderung der Blutzuckerwerte?

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



Kommt Ihnen diese Fragestellung irgendwie bekannt vor?

Stichwort: 2 Gruppen.

Für mehrere Gruppen braucht man die *Varianzanalyse* oder *ANOVA* (=ANalysis Of VAriance).

Es wird sich herausstellen (Woche 5), dass sich die Diäten tatsächlich voneinander unterscheiden.

Die nächste Frage ist dann, welche Diäten sich *paarweise* voneinander unterscheiden.

## Beispiel 4: Blut-Screening

(Aus Hothorn and Everitt, 2014, Chapter 7.1)

Untersucht wird, ob eine hohe ESR (erythrocyte sedimentation rate) ein Indikator für gewisse Krankheiten (Rheuma, chronische Entzündungen etc) ist.

**Konkret:** Gibt es einen Zusammenhang zwischen einem ESR Level  $ESR < 20 \text{ mm/hr}$  und den Plasmaproteinen Fibrinogen und Globulin?

Lade die Daten aus dem Package, welches für Hothorn and Everitt (2014) geschrieben wurde:

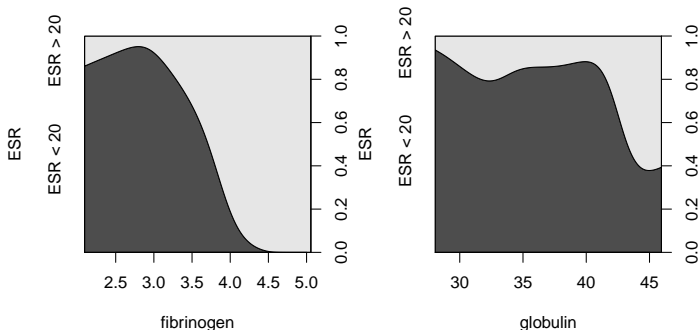
```
> library(HSAUR3)
> data("plasma", package="HSAUR3")
> plasma[c(1,5,9,10,15,29),]
```

	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

Die Unterteilung  $ESR < 20\text{mm/hr}$  vs.  $ESR \geq 20\text{mm/hr}$  führt zu einer **binären** Variable.

Der Zusammenhang der einzelnen Plasmaprotein-Levels kann mit einer grafischen Darstellung, dem *conditional density plot*, gut erfasst werden:

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



# Was ist ein Modell?

Ein Modell ist eine Annäherung an die Realität. Das Ziel der Statistik und Datenanalyse ist es immer, dank Vereinfachungen der wahren Welt gewisse Zusammenhänge zu erkennen.

David Hand schrieb 2014:

*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).*

# Vorgehen bei einem Modellierungsprozess

- 1 Präzise Fragestellung formulieren
- 2 Datenerhebung und -analyse planen, Daten sammeln (Experimente, Erhebungen)
- 3 Daten aufbereiten und bereinigen
- 4 Daten graphisch darstellen
- 5 Ein geeignetes *Modell* auswählen
- 6 Modellparameter und deren Unsicherheit schätzen
- 7 Modellannahmen überprüfen
- 8 Falls notwendig, Modell verbessern; zurück zu Schritt 7
- 9 Resultate interpretieren und mit Schritt 1 vergleichen
- 10 Resultate präzise und verständlich kommunizieren (Publikation, Zeitungsbericht...)

# Fragestellungen der Datenanalyse

- a) **Vorhersage, Interpolation.** Beispiel Körperfett: verwende Ersatzmessungen, um Körperfett einer Person vorherzusagen.
- b) **Schätzung von Parametern** (Beispiel: ...)
- c) **Bestimmung von Einflussgrößen.** Beispiel Aktivitätsstudie bei Kindern:  
Es werden Faktoren gesucht, welche das Bewegungsverhalten der Kinder (positiv oder negativ) beeinflussen.
- d) Optimierung
- e) Eichung

Hier befassen wir uns vor allem mit Fragestellungen a)-c).



## Ziele des Kurses (Teil 2)

Am Ende des Kurses sind wir in der Lage, alle hier eingeführten Beispiele zu Analysieren und Schlussfolgerungen daraus zu ziehen.

# Graphische Darstellung von Daten

Die folgenden graphischen Möglichkeiten sollten Sie kennen. In den obigen Beispielen haben wir einige wichtige Darstellungsarten bereits kennengelernt.

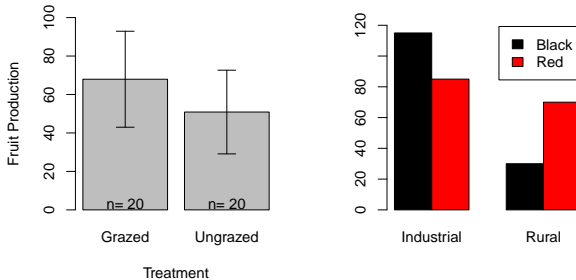
Darstellung	Nützlich bei
Streudiagramme (Scatterplots)	Paarweiser Abhängigkeiten kontinuierlicher Variablen.
Histogramme	Verteilungen kontinuierlicher Variablen.
Boxplots	Verteilung kontinuierlicher Variablen, ev. in Abhängigkeit von Kategorien.
Conditional density plots	Abhängigkeit einer binären Variable von kontinuierlichen Variablen.
Barplots	Ähnlicher Anwendungsbereich wie Boxplots.
Coplots	Darstellung von Abhängigkeiten von mehreren Variablen.

# Barplots

Beispiele aus Beckerman & Petchey:

**Links:** Untersuchung der Fruchtproduktion in abgegrasten und nicht abgegrasten Gebieten. Grasen reduziert die Biomasse über dem Boden. Wie wird Fruchtproduktion dadurch beeinflusst?

**Rechts:** Anzahl Vögel einer bestimmten Farbe in zwei Umgebungen.

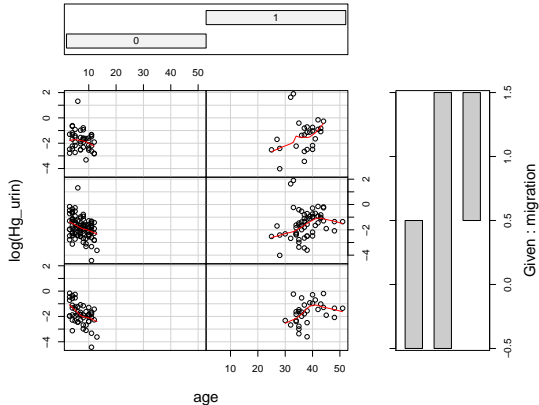


# Coplots

Ideal zur Darstellung von Abhängigkeiten, wenn mehrere Variablen involviert sind. Eignet sich sehr gut bei kategoriellen Variablen. Beispiel: Quecksilber im Wallis.

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

Given : mother



Es gibt viele “fancy” Arten, Daten graphisch darzustellen (**nice-to-know**):

- 3D-plots
- Räumliche Darstellungen (mit Geodaten)
- Interaktive Grafiken und Animationen

Dazu gibt es etlich R Pakete. Interaktive Darstellungen können beispielsweise mit Shiny Apps generiert werden (see census app).

# Nächste Woche: Einfache lineare 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 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.
- 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.