

Lecture 1: Introduction and Outlook

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

University of Zurich

31 January, 2021

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**.

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

Only with knowledge of data and statistical data analysis will it be possible 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?

Data and statistics are essential

A good foundation in statistics **makes you more independent** from consultants or the goodwill of colleagues. Without such a knowledge, you will have to heavily rely on and trust 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).

Examples of insights from data

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.

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, Spanjaardskanaal 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

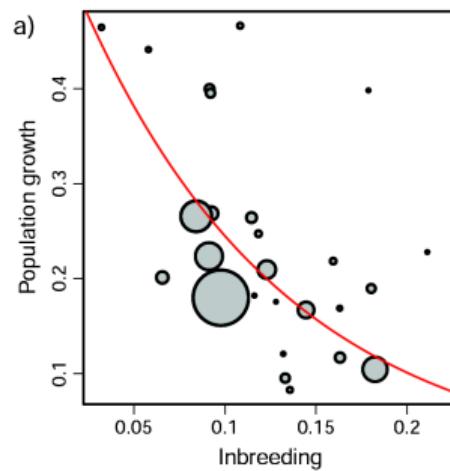
^f Swiss Federal Research Institute WSL, Biodiversity and Conservation Biology, 8903 Birmensdorf, Switzerland

<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)

NZZ am Sonntag | 3. April 2016

Wissen

61

Überschätzte Statistiken

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



Kritische Stimmen klonen sich fast so schnell wie die Daten, die mit statistischen Methoden bearbeitet werden. Angestellt der Massenmedien sah sich die ASA jetzt veranlasst, zum ersten Mal in ihrer Geschichte einen Bericht über die Auswirkungen von veröffentlichten Wissenschaftsergebnissen zu veröffentlichen und wie man mit einer statistischen Ortsweise vorsichtig vorgeht.

Was ist anders? In der Praxis ist der *p*-Wert indes auf die schiefre Bahn geraten. Dieser Wert ist ein Maß für die Wahrscheinlichkeit, dass eine Hypothese falsch ist. Der britische Geisteskrieger Ronald Fisher, (1920) als eine Art Informations-Sensoren für Auszüge aus dem Datenmaterial, ist in der Praxis oftmals zu einem simplen Lackmuster verkennet.

Bei einer statistischen Analyse von Daten entweder $p < 0.05$ (5 Prozent) oder noch besser < 0.01 (1 Prozent), gelten diese als signifikant. Das bedeutet, dass die statistisch Beweislast zugewandert. Das entscheidet etwa darüber, ob ein neues Medikament wirkt oder nicht. Ein *p*-Wert von 0,05 könnte zum Beispiel lauten, dass ein Medikament A gegen eine Newzealand-Krebs-Hypothese besagt dann genau das Gegenpol zu Medikament B. Ein *p*-Wert von 0,01 hingegen besagt dann, dass das Medikament A nicht wirkt. Bei einer Hypothese, dass die Signifikanzgrenze 5 Prozent bzw. 1 Prozent statistisch ist, kann man also sagen, dass Medikament A die Hypothese bestätigt, während Medikament B sie widerlegt.

Wissenschaftler verwenden den *p*-Wert, um Aussagen über die Wirkung eines Produktes zu treffen. Das klingt schlechte Presse und unangenehme Glaubwürdigkeit der Wissenschaft. Der Mediziner Michael Hsu ist hier ein Beispiel. Der Universität Stanford sprach in einem Konzertat von «dengen-abhängigen» oder «falsche Belege des *p*-Wertes». Erst wenn dieses erreicht ist, darf es ausgewertet, dass manche wichtig

werden darum – von allen wegen der oft Fehlinterpretationen von Publikationen belohnt werden. Das Angestellte der Massenmedien sah sich die ASA jetzt veranlasst, zum ersten Mal in ihrer Geschichte einen Bericht über die Auswirkungen von veröffentlichten Wissenschaftsergebnissen zu veröffentlichen und wie man mit einer statistischen Ortsweise vorsichtig vorgeht.

Der *p*-Wert misst nicht die Wahrscheinlichkeit, dass die Hypothese falsch ist, sondern die Wahrscheinlichkeit, dass sie wahr ist, und je geringer er ist, desto weniger spricht für sie. Ein *p*-Wert von 0,05 ist z.B. kein Beweis, dass das Ergebnis der Studie ein hoch interessantes Resultat unter den Fakten ist. Ein *p*-Wert von 0,01 ist ebenso kein Beweis, dass die Hypothese falsch ist. Der *p*-Wert misst nicht die Wahrscheinlichkeit, dass die Hypothese wahr ist, sondern die Wahrscheinlichkeit, dass sie zufällig zufällig wahr ist.

Der *p*-Wert ist eigentlich ein Hinweis, um zu unterscheiden, ob ein beobachtetes Resultat zufällig zufällig gekennzeichnet ist. Der kontinuierliche Wert (0,05) also setzt voraus, dass es sich um eine kontinuierliche Unterscheidung oder Zusammenhang. Der *p*-Wert ist aber kein Maß für die Wahrscheinlichkeit, dass die Hypothese falsch ist, erklärt Peter Flax. «Doch genau das wünschen viele Forscher und Journalisten», sagt er.

Haus kommt, dass die Signifikanzgrenze 5 Prozent bzw. 1 Prozent statistisch ist, kann man also sagen, dass Medikament A die Hypothese bestätigt, während Medikament B sie widerlegt.

Die Signifikanzgrenze 5 Prozent bzw. 1 Prozent statistisch ist, kann man also sagen, dass Medikament A die Hypothese bestätigt, während Medikament B sie widerlegt.

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

Die britischen Statistiker Jonathen Sterne und David Smith haben schon vor 15 Jahren im «British Medical Journal» den

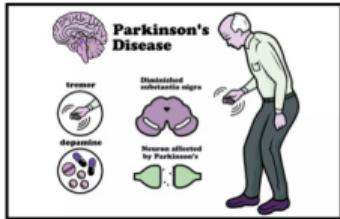
zu unterscheiden, dass die Hypothese von medizinischen Studien nicht mehr als signifikante oder erstaunliche Wirkung darstellen, sondern als wahrscheinlich. Und die Hypothese ist das, was man in Wirklichkeit zu untersuchen plant. Gerade das ist ein Problem, das die NZZ am Sonntag in einer anderen erschienenen Studie festgestellt hat. Denn nachdem die britischen Statistiker einen kleinen britischen Forschungstrupp gemacht hatten, der Studien untersuchte, die *p*-Werte angegeben haben, fanden sie, dass nur 10 Prozent davon tatsächlich signifikant waren. Gleichzeitig werden Datenzusammenfassungen zu den entsprechenden Wirkungen immer häufiger, wie z.B. in *Science*, S. 3, THI.

Richtig Alternativen?

Das Problem ist dabei nicht nur, dass der *p*-Wert ein trügerisch einfaches statistisches Instrument ist, «einfaches Pfeilschussinstrument», wie es der britische Statistiker Andrew Gelman in seinem Artikel über die Probleme mit dem *p*-Wert schreibt. Er fordert, dass man andere Maße für statistische Signifikanz, zum Beispiel *p*-Werte, benutzen sollte. Gelingt das nicht, dann müssen andere Maße für die Hypothese anhand der Daten angesehen, statt dass diese wie beim *p*-Wert nach einer Hypothese abgeschaut wird.

Hier exemplifiziert, die Beispiele von Studien sind leicht verständlich, wenn man Vertrautheit mit dem *p*-Wert hat.

Question you will work on



Producing nonsense with statistics...

... is too easy ...

The risks of alcohol (by David Spiegelhalter, 23. August 2018)

“Calling bullshit” course (University of Washington)

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

How do we get insights from data... .

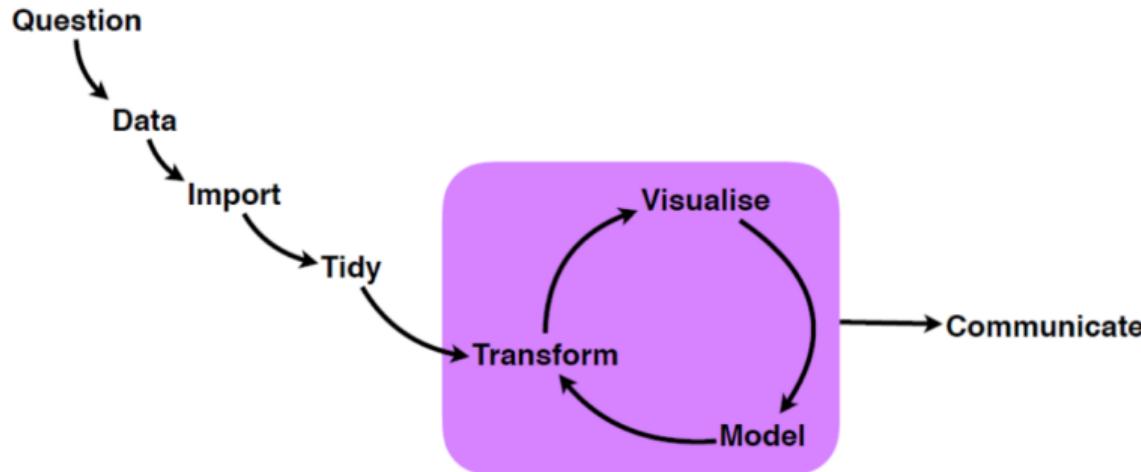
... rather than nonsense.

- ▶ Excellent data management practices.
- ▶ Informative graphical visualisations to explore data.
- ▶ Informative numerical summarise to explore data.
- ▶ Appropriate transformations of data.
- ▶ Appropriate statistics tests / models.

Awareness of our “realm”:

- ▶ Description of patterns, including associations (we will do this).
- ▶ Predicting (we will do this).
- ▶ Inferring causation (we will do this, by analysing experiments randomised manipulations).

Steps in a getting insights from data (“work flow”)



Visualising 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.
Box and whisker plots	Distribution of continuous variables for different categories.

All can be augmented, for example by “conditioning” (e.g. colouring points according to the values of a variable).

A career in visualising data??!!

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/>

What is a (statistical) 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."

Goals of the course (part 2)

By the end of the course you will be able

- ▶ to **explore** and **analyze** data with appropriate methods, including statistical models,
- ▶ 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.

Literature

Recommended 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.

