Hands-on Data Analysis with R
University of Neuchatel, 10 May 2016

Welcome & Road Map

Bernadetta Tarigan, Dr. sc. ETHZ

Agenda

08:45 – 12:45	Welcome and Road map Linear & Piecewise Regression (Hands on ***critic*** data) GLM & Feature Selection (Hands on ***bug data***)
12:45 – 14:15	Lunch
14:15 – 18:00	Continue with *bug data* Two-Way ANOVA (Hands on ***variabletype data***) Wrap up and Evaluation

When coffee breaks?
In between!

Who is talking to you?

	TEACHING/TUTORING	RESEARCH	COACHING	CONSULTING & DATA ANALYSIS
2010-2016 ETHZ-CH Self-employed (ZH)		www.mis.ethz.ch www.statistical-coaching.ch (until 2015)	• Probability & Statistics • Regression & GLM • Optimization • Quantitative Methods • Stochastic Processes (discrete & continue) • Test theory • Extreme values	• Clinical Trial • Management IS: Wikipedia, Sail, Comparis • Earthquake • Climate • Finance • Re-Insurance
2009-2010 VUA-NL		Realistic Neural Connectivity Networks model		
2008-2009 ITB-Indonesia	• Mathematical stats • Statistics for engineers & scientists			
2003-2008 Leiden–NL ETHZ–CH	• Computational statistics with R • Mathematical stats • Master-thesis	• Classification • Multivariate discriminant analysis • Pattern recognition • Machine Learning		
2000-2003 CWI–NL		• Probability of ruin model • Bootstrap		
1995–2000 ITB-Indonesia	• Basic Statistics • Basic Probability • Calculus	• Winning research grant to NL • Master in Math/Stat with scholarship		

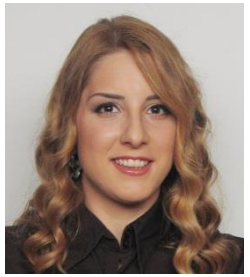
Who is organizing this workshop?

Prof. Oscar Nierstrasz,
University of Bern

Mascha Kurpicz-Briki,
University of Bern

Prof. Pascal Felber,
University of Neuchâtel

Nevena Milojkovic,
University of Bern



Please introduce yourself too

- Your name
- A few words about your research project(s)

First name	Last name	First name	Last name
Antoine	Bellwald	Mirco	Kocher
Ionel Tudor	Calistru	Mascha	Kurpicz
Michele	Catasta	Alain	Mermoud
Claudio	Corrodi	Nevena	Milojkovic
Ana	De Abreu	Oscar	Nierstrasz
Ljiljana	Dolamic	Haidar	Osman
Mohammad	Ghafari	Iuliia	Proskurnia
Amit	Gupta	Jonnahtan	Saltarin
Yaroslav	Hayduk	Kasun	Samarasinghe
Aigul	Kaskina	Yuriy	Tymchuk

Do not hesitate…

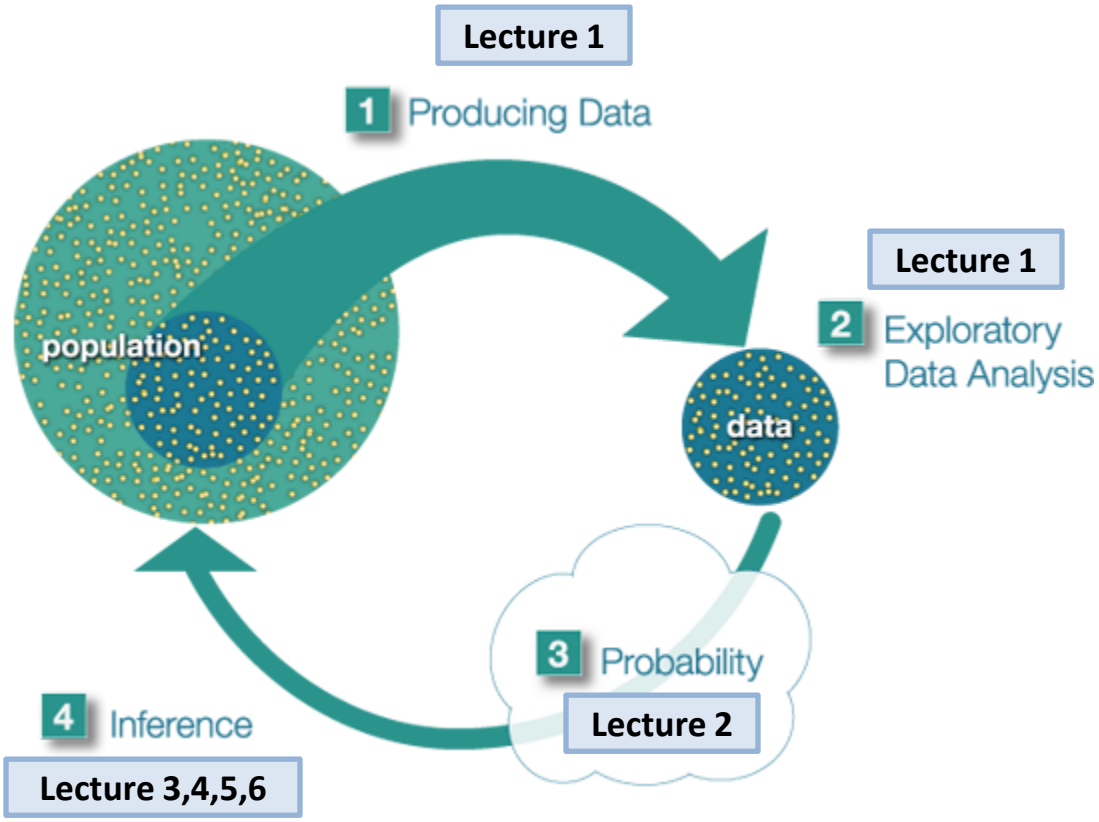
- to interrupt me
- to participate actively
- to say

  Bernadetta,

  - *You talk too fast…*
  - *Can you please repeat this and that…*
  - *I don't understand why…*
  - *How did we get that results/numbers/values…*
  - *I think you are wrong… (who knows ☺)*

2015, remember?

Statistical Inference: drawing conclusion about population from sample with some calculated **un**certainty



What we have learned in 2015?

1. Data Analysis & Producing

2. Introduction to Probability

3. Foundation for Statistical Inference
   - Parameter vs Statistics
   - Sampling Distributions
   - Point Estimation
   - Set Estimation
   - Hypothesis Testing

4. Two-Way Table and One-Way ANOVA

5. Multiple Linear Regression

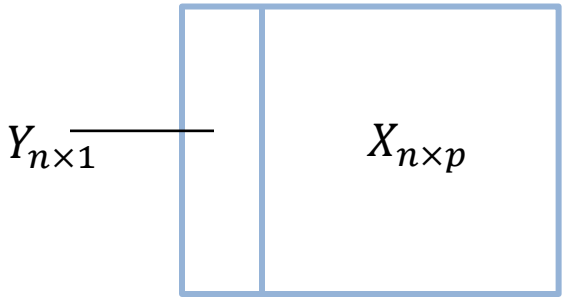
6. Binary Logistic Regression

And many R commands…

Data Analysis = Function Estimation

Data = numbers with context

- consists of variables and cases/observations
- data ≠ information



Y: dependent, response, outcome, output

X: independent, explanatory, predictor, input

We assume

$$Y = f(X) + \varepsilon$$

ε is random part with $E[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$

f is a function connecting Y to X

But f is unknown, want to estimate it with $\hat{f}$

Function Estimation

- What is the goal/purpose of estimating f?
  - To describe how Y depends on
  - To explain causal effect of $X_1, X_2, \ldots, X_p$ on Y (testing causal theory)
  - Predicting Y from $X_1, X_2, \ldots, X_p$ (association)
- How to choose an estimate $\hat{f}$?
- What criterion to assess the performance of an estimate $\hat{f}$?

Data sets we have today

	Data set 1 ***critic*** (from Yuriy)	**Data set 2** ***bug*** (from Haidar)	**Data set 3** ***variabletype*** (from Nevena)
Goal	Explanation Hypotheses test	Prediction	Explanation Hypotheses test
Methods	Simple & Piecewise Linear Regression	GLM & Feature Selection	Analysis of Variance (ANOVA) - Two way with repeated measures

Linear Regression Models

- Simple, easy to understand
- For explanation and description purpose, often provide an adequate and interpretability description on if and how the inputs affect the output
- For prediction purpose they can sometimes outperform fancier nonlinear models, especially in situations with:
  - Small numbers of training data
  - Low signal-to-noise ratio
  - Sparse data ($n \ll p$)
- Moreover, linear methods can be applied to transformations of the inputs and this expands their scope, e.g., basis function methods

General steps of data analysis