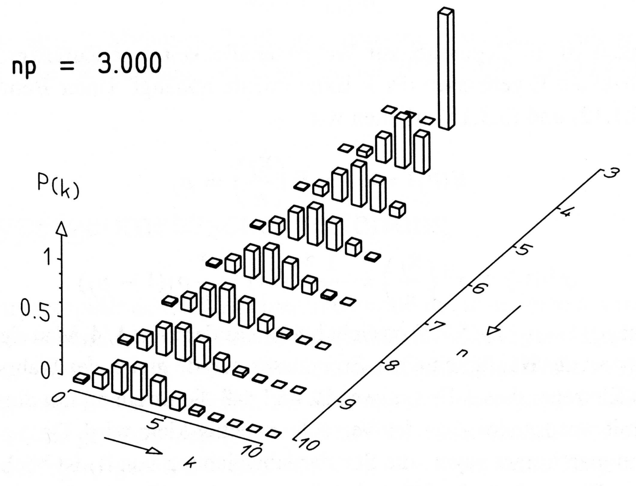
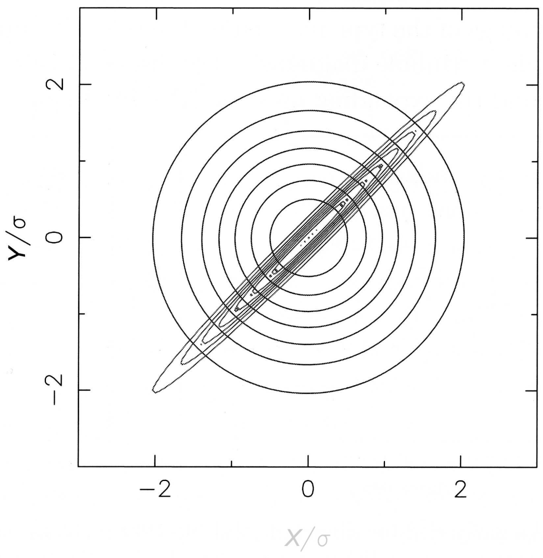
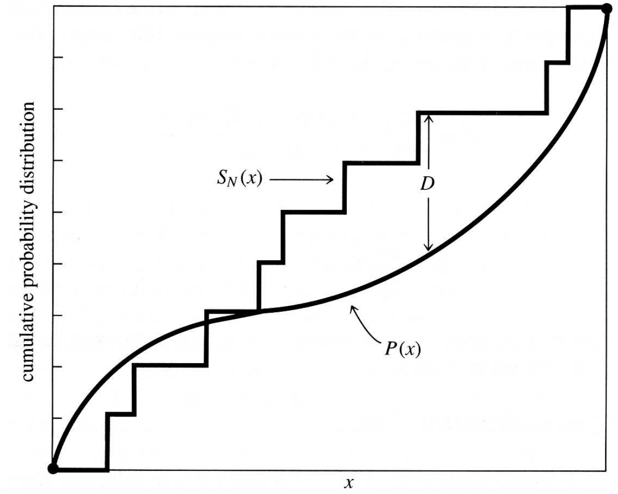






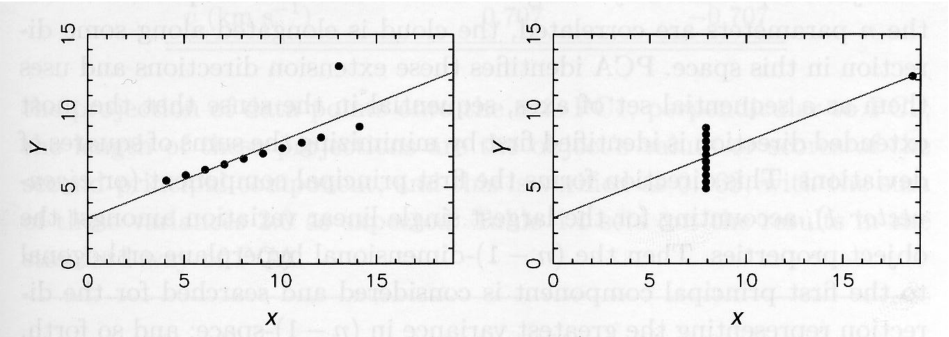
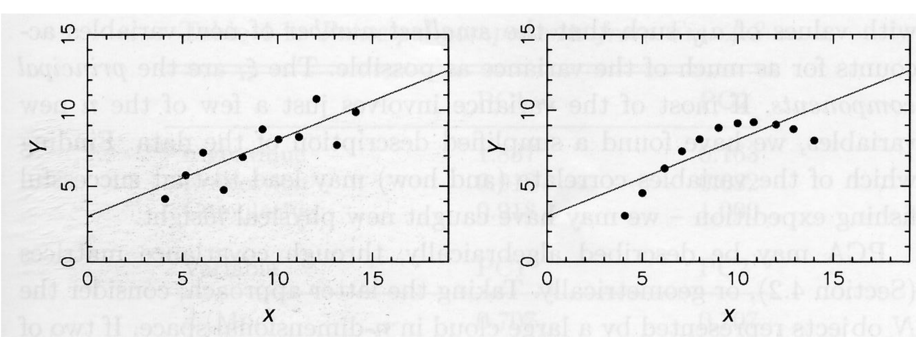
Statistical methods(UKSta)

Introduction

Dr. Yiannis Tsapras

Summer term 2024

(Based on original lectures by Prof. Dr. N. Christlieb and others)

Code: UKSta	Modulname: Statistical Methods
Art des Moduls	Wahlpflichtmodul
Modulbetreuer	
Sprache	Englisch
Leistungspunkte*	3
Lerninhalte des Moduls*	• Concept of probability, probability distributions, Bayesian reasoning • errors, error propagiation, estimation, uncertainty • orthodox hypothesis testing (e.g. t-test) and Bayesian model comparison • linear models and regression • binomial and poisson processes • likelihood-based modelling: prior, likelihood, posterior; maximum likelihood, least squares, chi-squared • Bayesian modelling using numerical (Monte Carlo) methods: sampling, integration • nonlinear and nonparametric methods: density estimation, kernel methods, regularization • statistics with the R pgrogramming language
Lernziele	learning the principles and methods of probability and statistics needed for analysing, modelling and interpreting data
Lehr- und Lernformen*	• Laboratory course, homework Literatur: Notes provided by lecture, plus book/internet recommendations Besonderheiten: course given in English; block course of 10 half days over two weeks (mornings)
Voraussetzungen für die Teilnahme, ggf. vorgeschriebenes oder empfohlenes Studiensemester*	Notwendige/nützliche Vorkenntnisse: basic (high school) statistics and first semester maths (for physicists). Recommended from the third semester
Verwendbarkeit des Moduls*	(siehe Präambel).
Voraussetzung für die Vergabe von Leistungspunkten, Arbeitsaufwand und Noten*	Prüfungsmodalitäten: Doing the exercises in class, submitting the homework, presenting the homework at least once
Häufigkeit des Angebots von Modulen*	Sommersemester
Dauer*	2 Wochen

What is statistics?

- Summary description of data
  - Mean; median; variance; quartiles of a distribution
  - Diagrams; Tables
  - Principal component analysis (PCA)
- Inference from data; decision making
  - Determination of the parameters of a model
  - Do the measurements agree with the model?
  - Do two sets of measurements/properties of two samples agree with each other?
- Understanding structure in data
  - Are two parameters correlated with each other?
  - Classification: can data be grouped according common properties?

The role of statistics

- "The logic behind science"
- Not only important for describing/analysing given datasets, but also for planning/executing experiments as well as designing surveys and compiling samples.

Statistical diversity

- Genetics, central role in bioinformatics
- Kinetic theory of gases
- Design of computer operating system (e.g., theory of queues)
- Noise in electrical devices
- Model atmospheric turbulence
- Insurance and finance
- Theory of complex systems
- … and much more

How to deal with the diversity?

- Here: emphasis on Monte Carlo approach
- Importance constantly increasing due to economic computing resources
- Books often deal with methods specific to a particular subject
- Overview is difficult to obtain

Course aims

Main aims:

- Learn basic concepts of statistics
- Learn how to use computational tools for (describing)/(analysing)/(inferences from) data
- Practical approach emphasized, only some theory
- Not a full fledged R-programming course

  (online https://www.coursera.org/course/rprog)

Side aims:

- Learn how to work with Jupyter notebooks, handling files in a Unix-like environment (exporting? sharing?)

Course topics: probability

$$p := \frac{\text{number of favourable events}}{\text{total number of events}}$$

$$p(\text{A}|\text{B}) = \frac{p(\text{A and B})}{p(\text{B})}$$

$$p(\text{A}|\text{B}) = \frac{p(\text{B}|\text{A})p(\text{A})}{p(\text{B})}$$

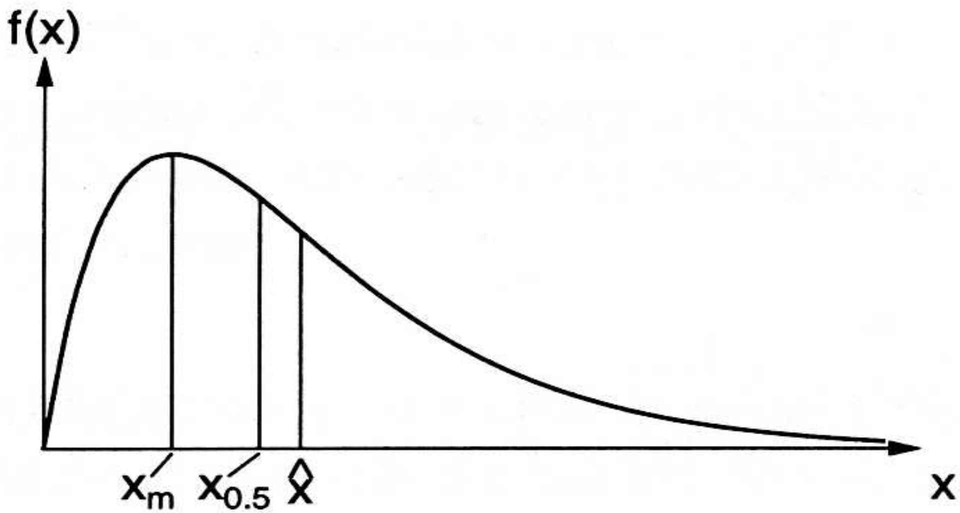
(1) For a random event A, $0 \leq p(\text{A}) \leq 1$.

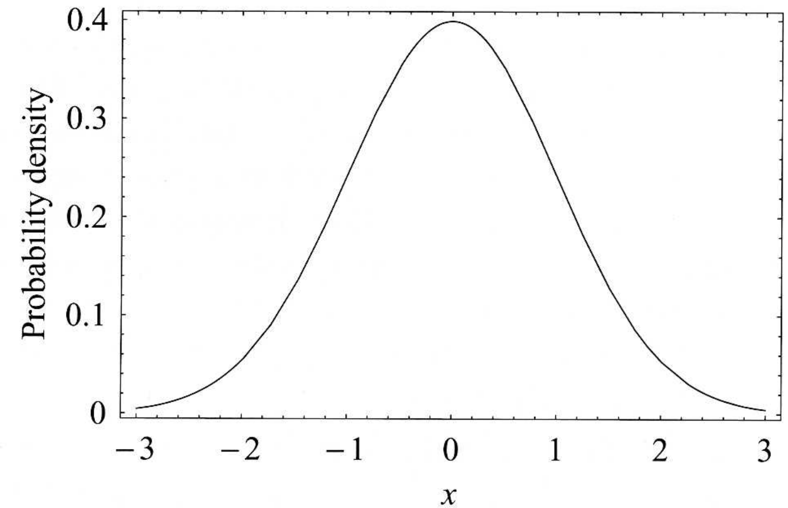
(2) For the sure event A, $p(\text{A}) = 1$.

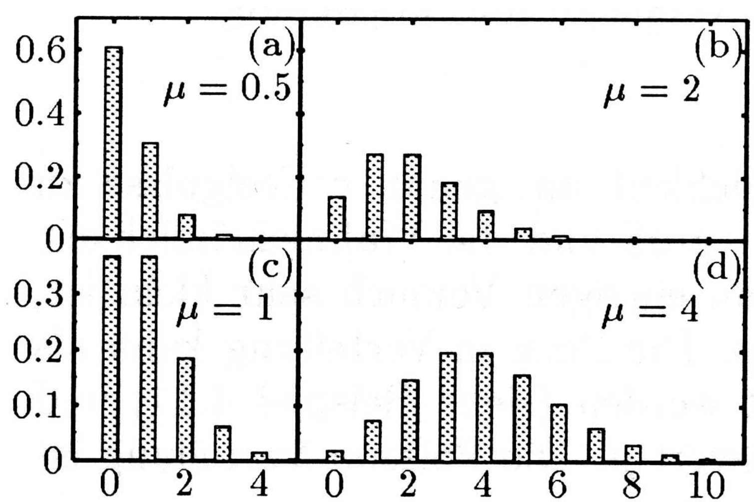
(3) If A and B are exclusive events, then

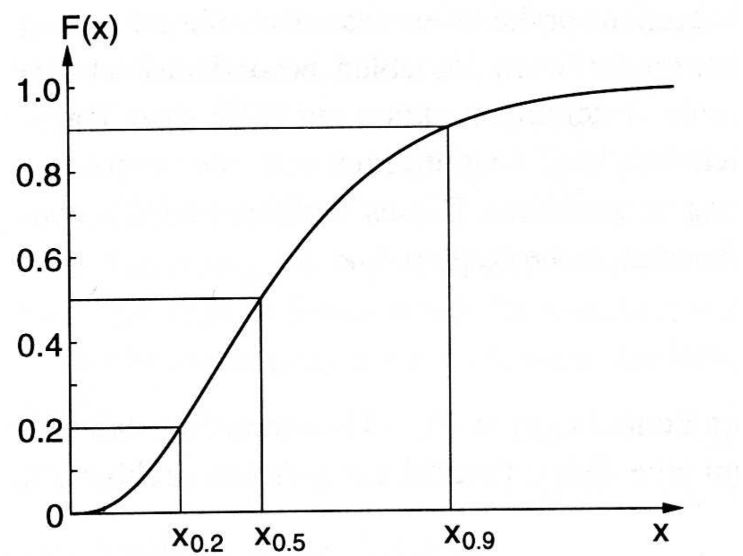
$$p(\text{A or B}) = p(\text{A}) + p(\text{B}).$$

Course topics: probability distributions









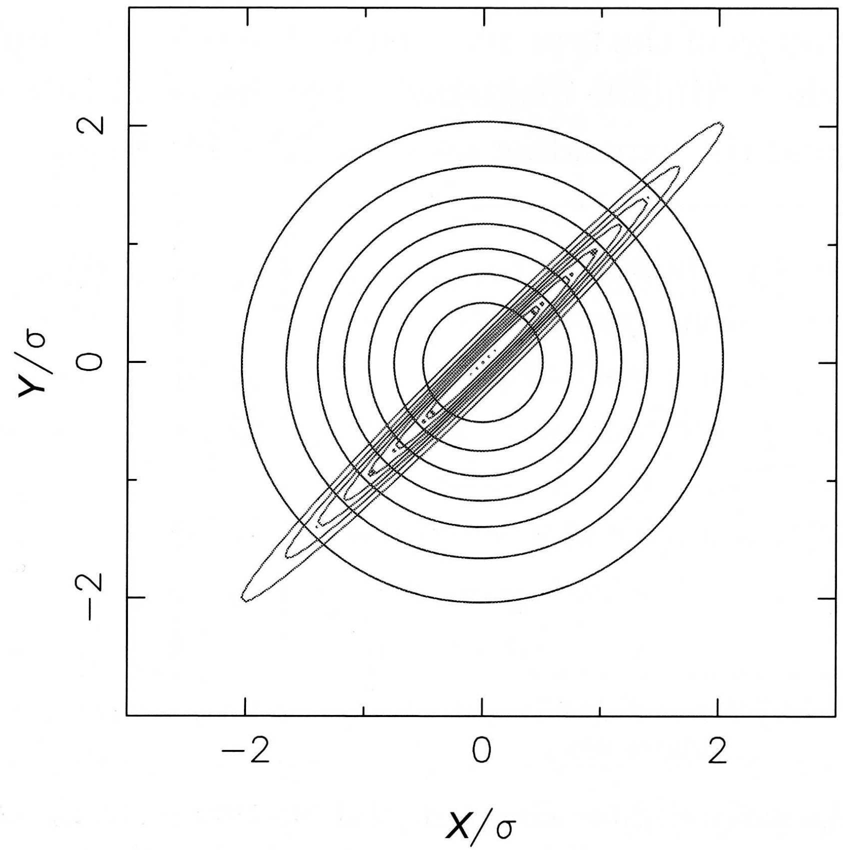
Course topics: covariance and correlation

Consider measurements x_i and y_i of the variables x and y. The covariance σ_{xy} is related to to the correlation coefficient $\rho(x, y)$,

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

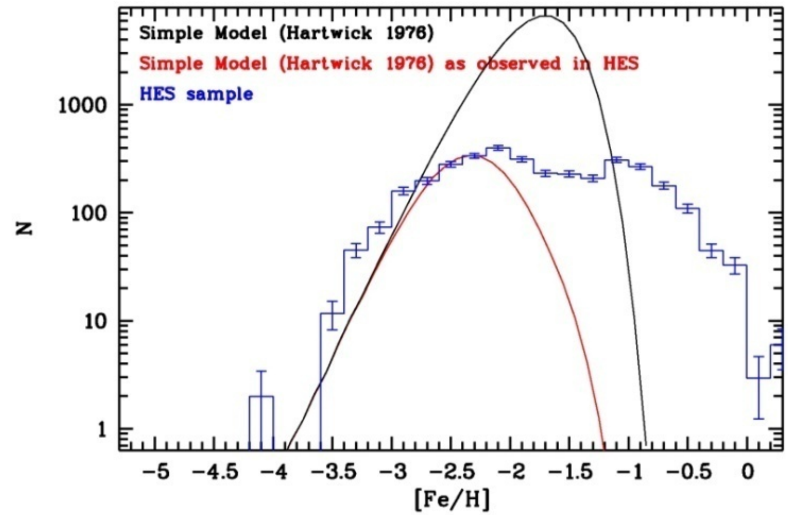
It can be estimated by

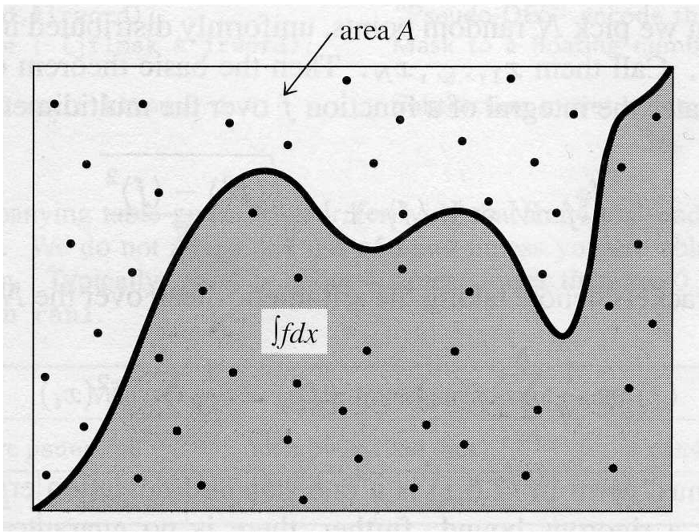
$$\hat{\rho}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$



Course topics: Monte Carlo methods

- Method of choice when statistical problems can not (easily) be solved analytically.
- Simulation of data sets; e.g. simulated measurements with uncertainties following a Gaussian distribution.
- Monte-Carlo integration.





Course topics: Parameter estimation

Let $x_1, x_2, \ldots, x_n$ be measurements which follow the probability distribution $f(x|a)$, where a is one or more free parameter(s). The likelihood function $L(a)$ is defined as

$$L(a) = f(x_1|a) \cdot f(x_2|a) \cdots f(x_n|a) = \prod_{i=1}^{n} f(x_i|a).$$

$L(a)$ is the probability for measuring the set of values $x_1, x_2, \ldots, x_n$, given the parameter(s) a and the probability distribution function $f(x|a)$.

According to the maximum likelihood principle, the best estimate $\hat{a}$ of a is the one which maximizes the likelihood function; i.e.,

$$L(a) \stackrel{!}{=} \text{maximum}.$$

Course topics: Error propagation

We consider a transformation

$$y_i(x_1, x_2, \ldots, x_n), \;\; i = 1 \ldots m.$$

The law of error propagation is

$$\mathbf{C}[\mathbf{y}] = \mathbf{B}\mathbf{C}[\mathbf{x}]\mathbf{B}^T,$$

where $\mathbf{C}[\mathbf{y}]$ and $\mathbf{C}[\mathbf{x}]$ are the covariance matrices for $\mathbf{y}$ and $\mathbf{x}$, respectively, and

$$\mathbf{B} = \begin{pmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \cdots & \partial y_1/\partial x_n \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & \cdots & \partial y_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial y_m/\partial x_1 & \partial y_m/\partial x_2 & \cdots & \partial y_m/\partial x_n \end{pmatrix}.$$

Course topics: Linear regression

$$L(a,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[y_i - (ax_i + b)]^2}{2\sigma_i^2}\right\}$$

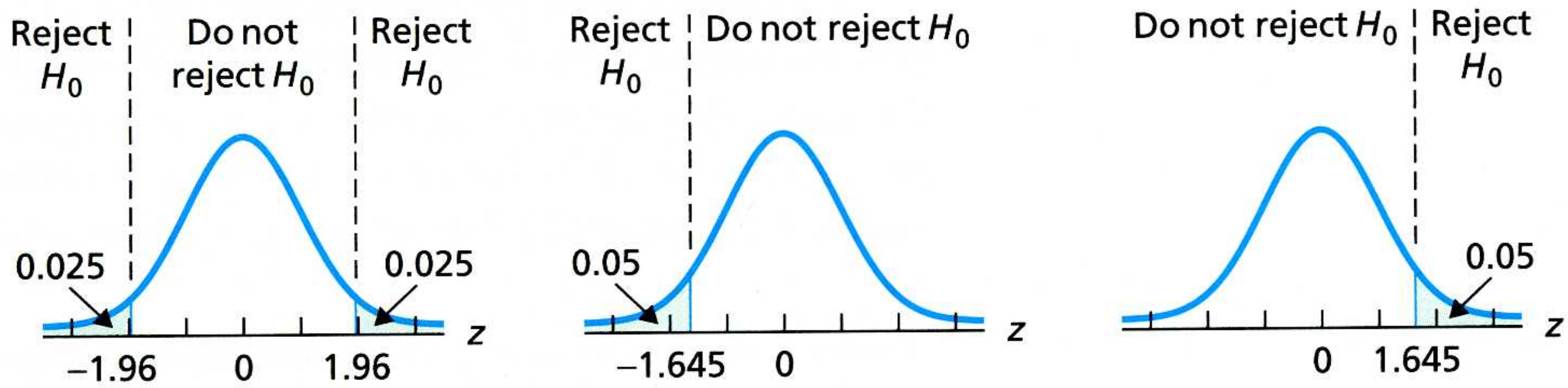
$$l(a,b) = \text{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (ax_i + b)]^2.$$

$$\begin{aligned} a &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ b &= \frac{1}{n}(\sum y_i - a\sum x_i) \end{aligned}$$

Course topics: hypothesis testing

Test	H_0	Assumptions	Parameters	Test Statistic
Student's *t* test	$\mu_x = \mu_y$	Data is Gaussian	$\mu_x, \mu_y, \sigma_x, \sigma_y$	t
F test	$\sigma_x = \sigma_y$	Data is Gaussian	σ_x, σ_y	F
χ^2 test	Same parent distribution	$(O_i - E_i)^2$ is Gaussian	—	χ^2
KS test	Same parent distribution	—	—	D
U test	Same parent distribution	—	—	U_A, U_B
Spearman	Data is uncorrelated	—	—	r_s
Runs test	Data is random	—	—	r

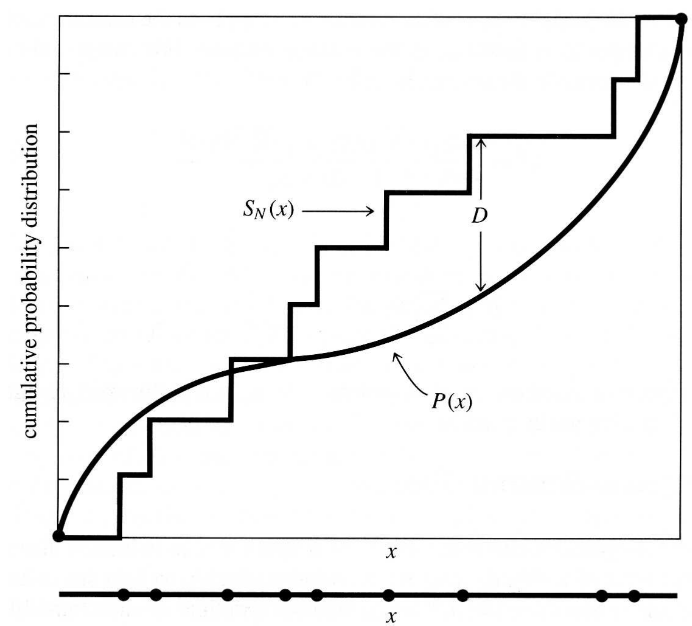
Course topics: hypothesis testing



(a) Two tailed (b) Left tailed (c) Right tailed

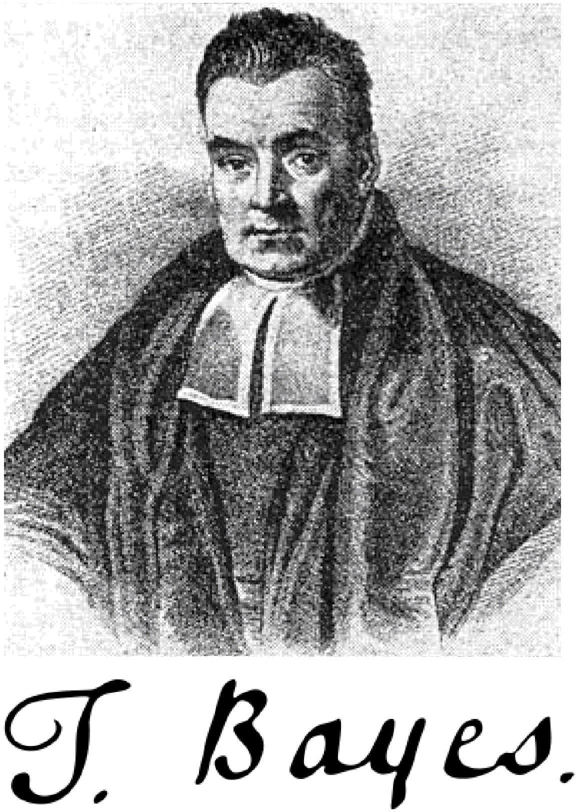
$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n(n-1)} + \frac{\sum(y_i - \bar{y})^2}{m(m-1)}}},$$

$$F = \frac{s_x^2}{s_y^2} = \frac{m-1}{n-1} \cdot \frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}.$$



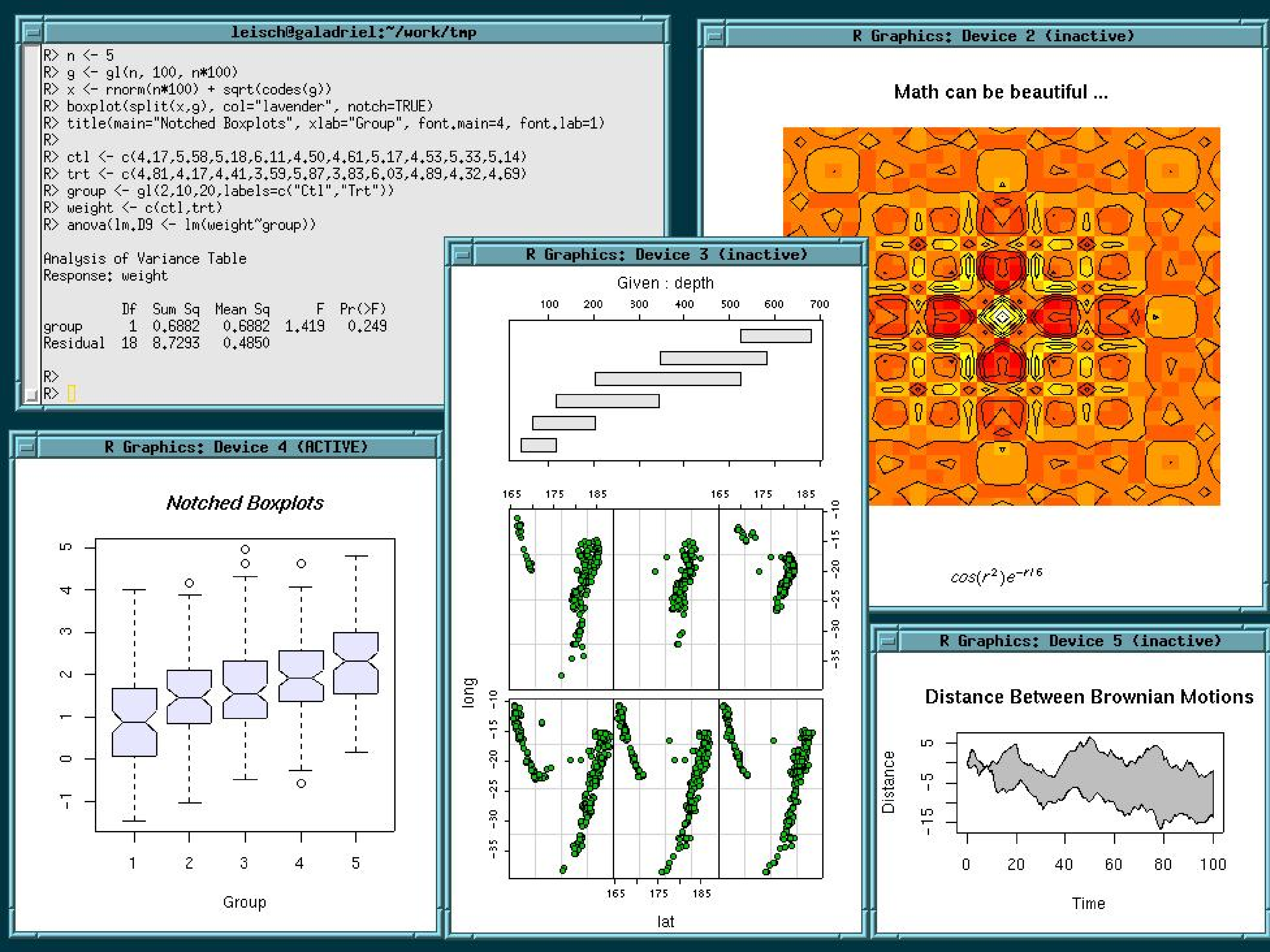
Bayesian methods

- Basic idea behind Bayesian approach
- The role of a prior
- Bayesian parameter estimation
- Bayesian model selection

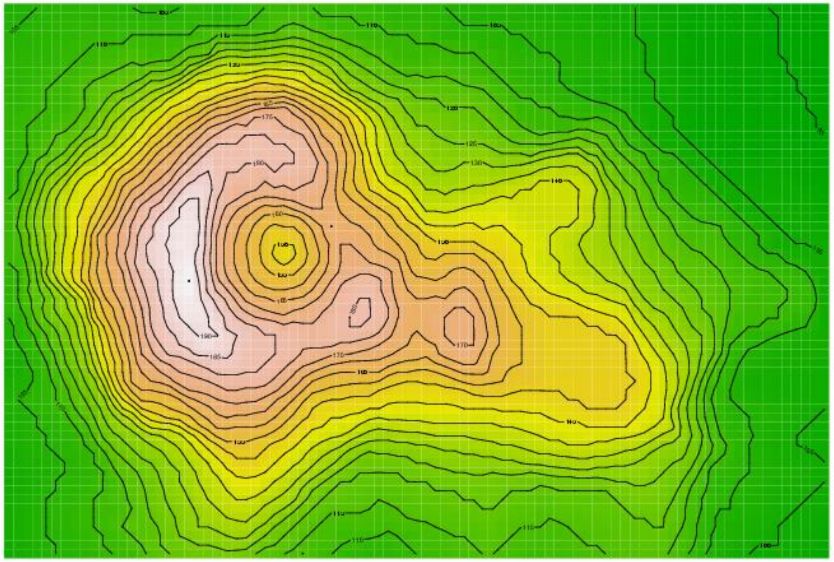


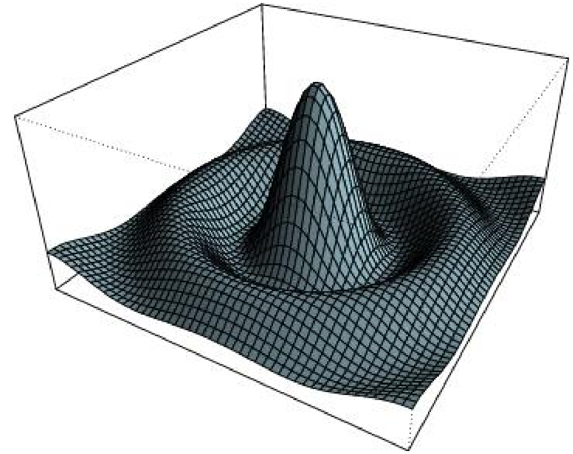
R

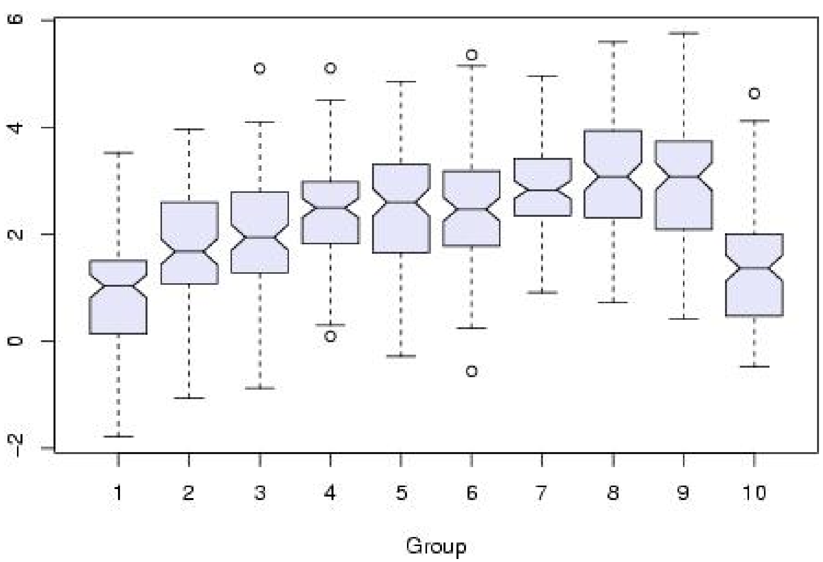
- Programming language and environment for statistics and data analysis
- Platforms: Linux, MacOS X, Windows
- Published under GNU General Public License (GPL); i.e., freely available (see `www.r-project.org`)
- Command-line; interpreter
- Object oriented (will not play a big role)
- Own programs can easily be integrated
- Extensive statistics library – but here a lot DIY
- Very powerful graphics package(s)

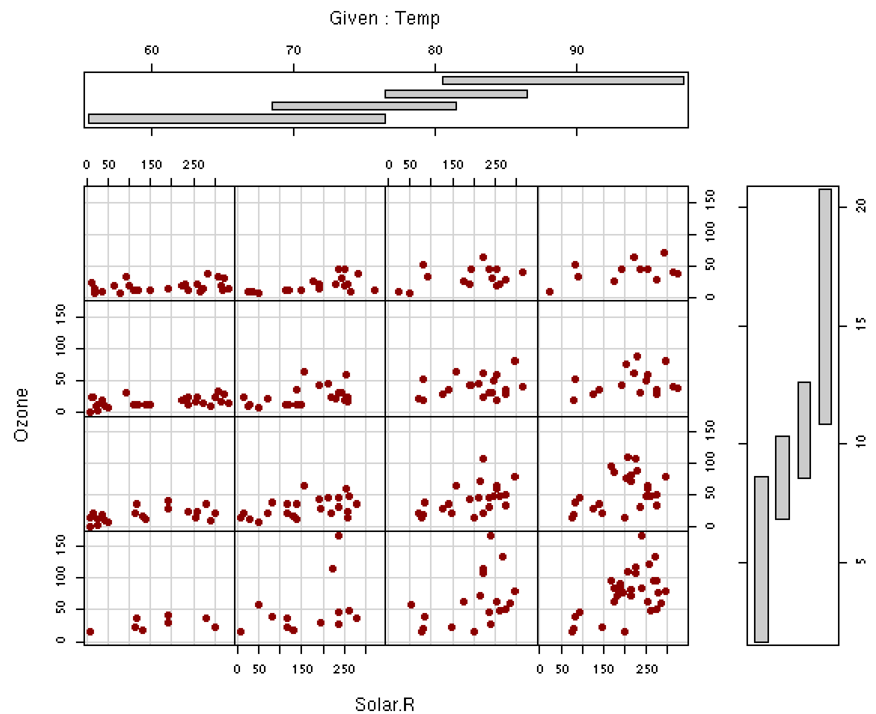

leisch@galadriel:~/work/tmp
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))
Analysis of Variance Table
Response: weight
Df Sum Sq Mean Sq F Pr(>F)
group 1 0.6882 0.6882 1.419 0.249
Residual 18 8.7293 0.4850
R>
R>
R Graphics: Device 2 (inactive)
Math can be beautiful ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
100 200 300 400 500 600 700
165 175 185
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
-1 0 1 2 3 4 5
1 2 3 4 5
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
-15 -5 5
0 20 40 60 80 100
Time

Graphics produced in R









Preliminary course plan

Day		Topic(s)
Mon	**5 Aug**	Org.; running R, jupyter, markdown; R tutorial
Tue	**6 Aug**	Probability; Probability distributions, more R
Wed	**7 Aug**	Random numbers; Monte Carlo methods 1
Thu	**8 Aug**	Monte Carlo methods 2
Fri	**9 Aug**	Bootstrap, Maximum Entropy
Mon	**12 Aug**	Maximum Likelihood Estimation
Tue	**13 Aug**	Bayesian parameter estimation
Wed	**14 Aug**	Hypothesis testing 1
Thu	**15 Aug**	Hypothesis testing 2
Fri	**16 Aug**	EM procedure? Gaussian Processes?

Course is under development! (Dauerbaustelle)

- Time management? Overlap between days?
- Feedback appreciated

Course format

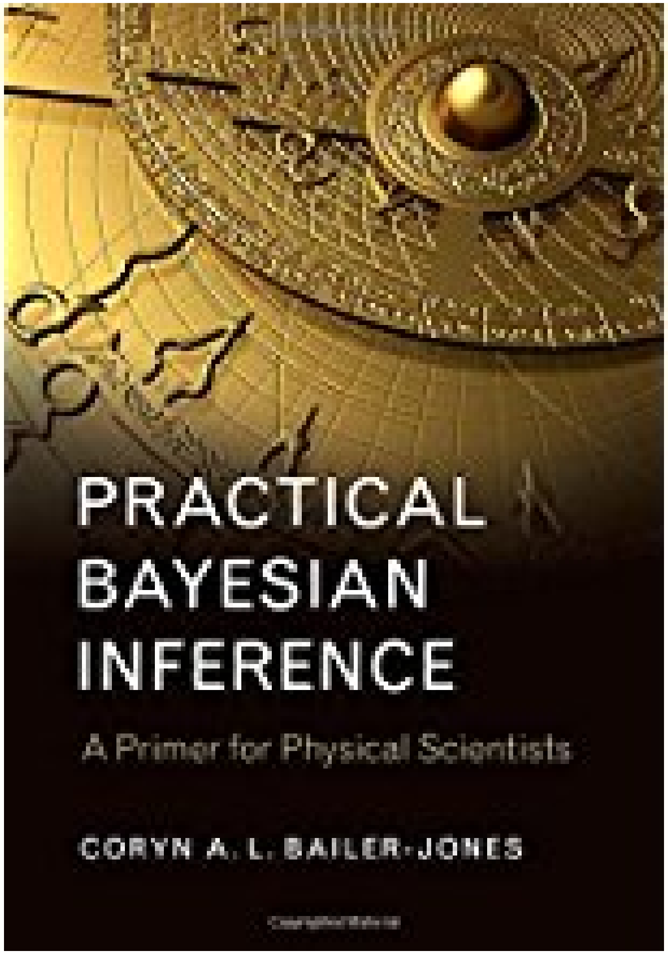
- Time: **Mo/Tu/Th/Fr 9:00-13:00, break 10:45-11:15**
- Presence is mandatory; exceptions have to be discussed with me in advance.
- **14:00-17:00** Work on assignments; up to 3 people
- The results of homework assignments have to be submitted in writing by **9:15** the next day as **single PDF** (in addition perhaps R-notebook) via Ü-system
- To pass the course and earn the 3 ECTS credit points, *60%* of **every** homework assignment have to be solved in a satisfactory manner.
- In cases, work-over of unsatisfactory solutions

Resources

- Lecture slides on the web
  Note: lecture slides are not a script!
- Other handouts
- Online help pages and tutorials
- Course page with lecture material on the web
- Books

Statistics books

Coryn Bailer-Jones
Practical Bayesian Inference:
A Primer for Physical Scientists
1st edition, 2017
29 €



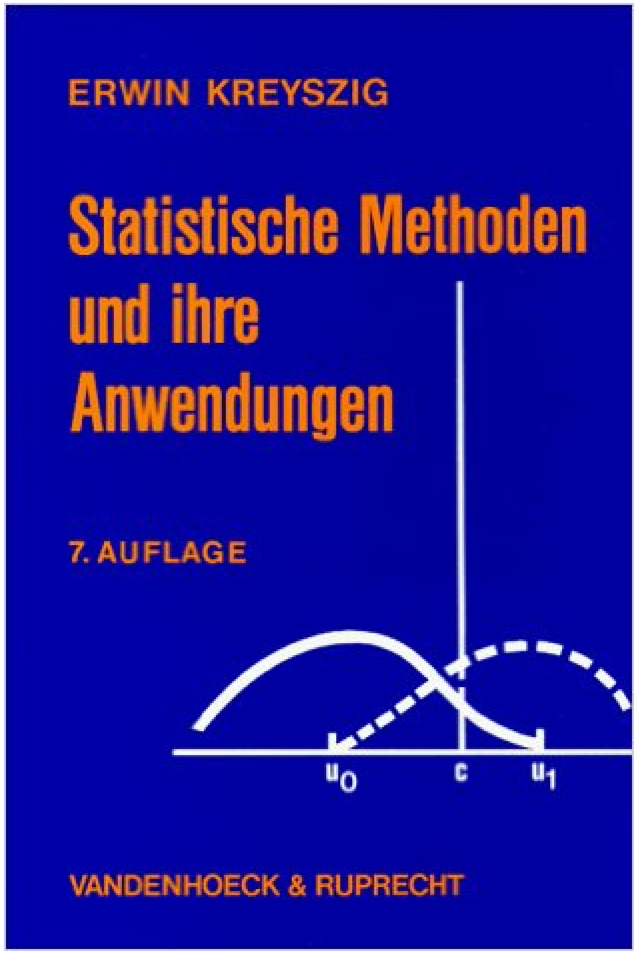
Very useful for the course.
Some examples taken from the Book. Available online at UB Heidelberg.

R-scripts of the book can be found at the web site of Coryn Bailer-Jones:
http://www2.mpia-hd.mpg.de/homes/calj/

Statistics books

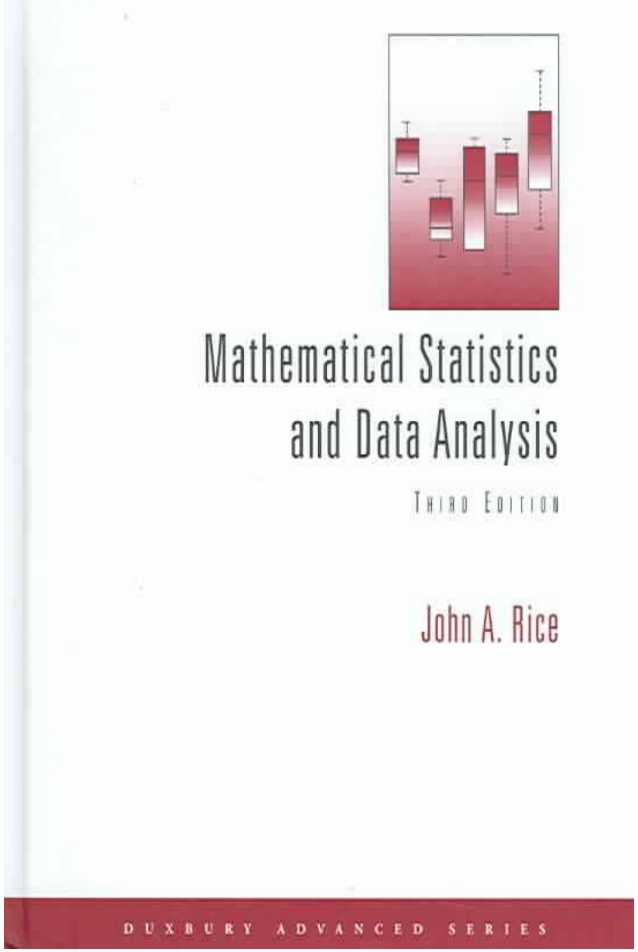
Erwin Kreyszig,
Statistische Methoden und ihre Anwendungen
7th edition, 1979 (!)
40 €

(in German only)



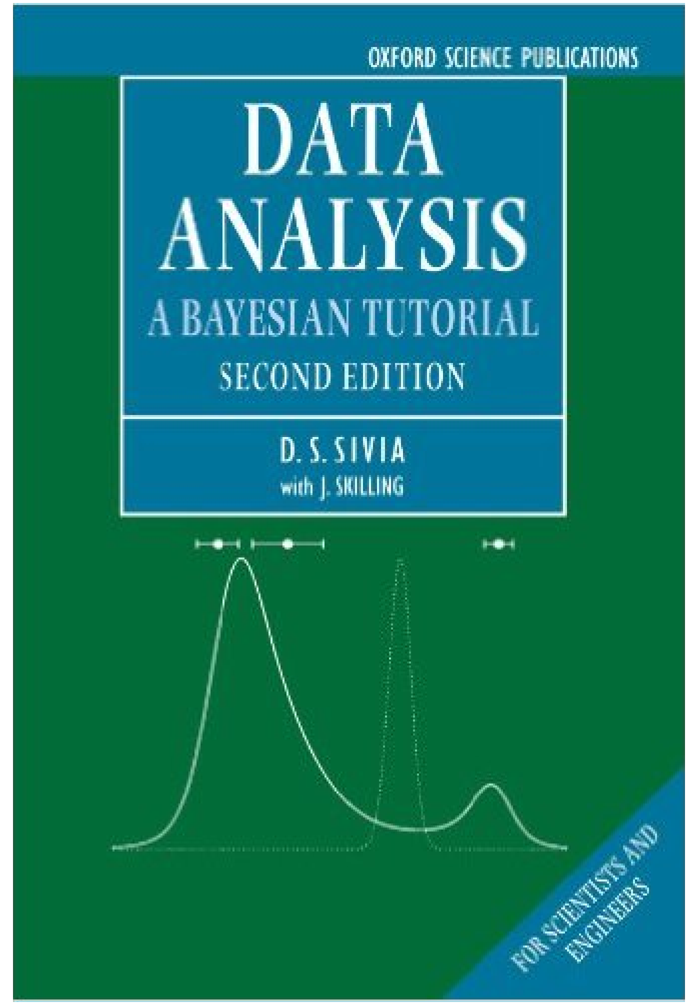
Statistics books

John A. Rice,
Mathematical Statistics and Data Analysis
3th edition, 2007
26 €



Statistics books

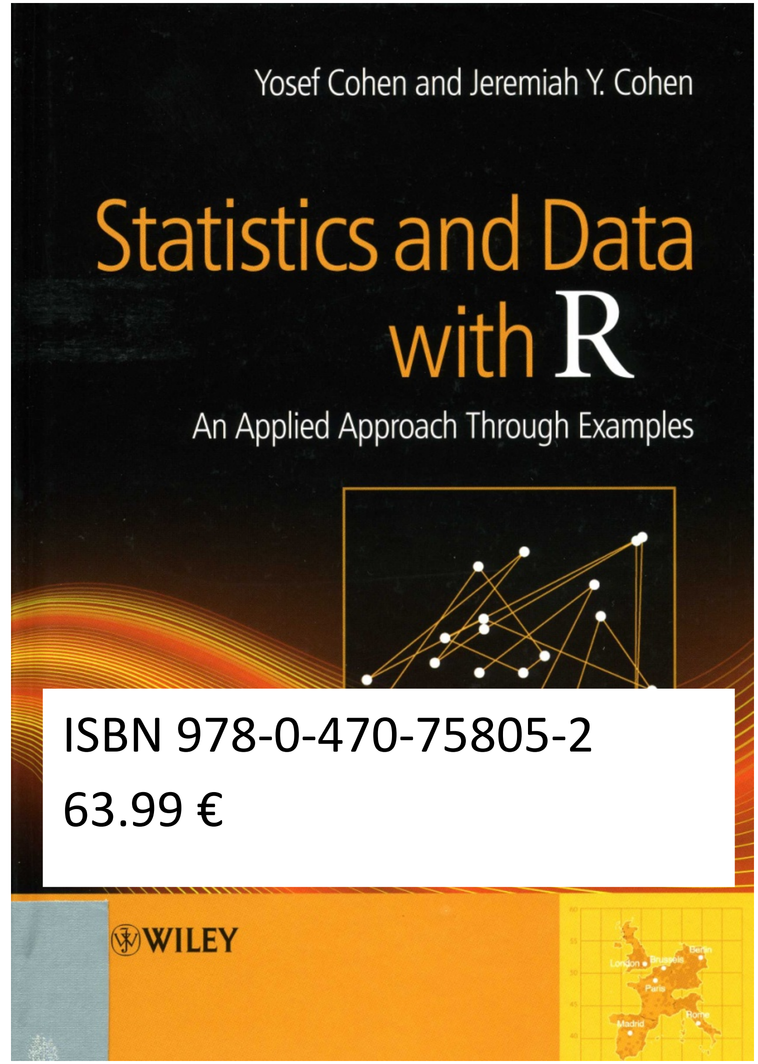
Sivia & Skilling
Data Analysis: A Bayesian Tutorial
1st edition, 2006
30 €



R books

ISBN 978-3-540-79997-9

29.95 €

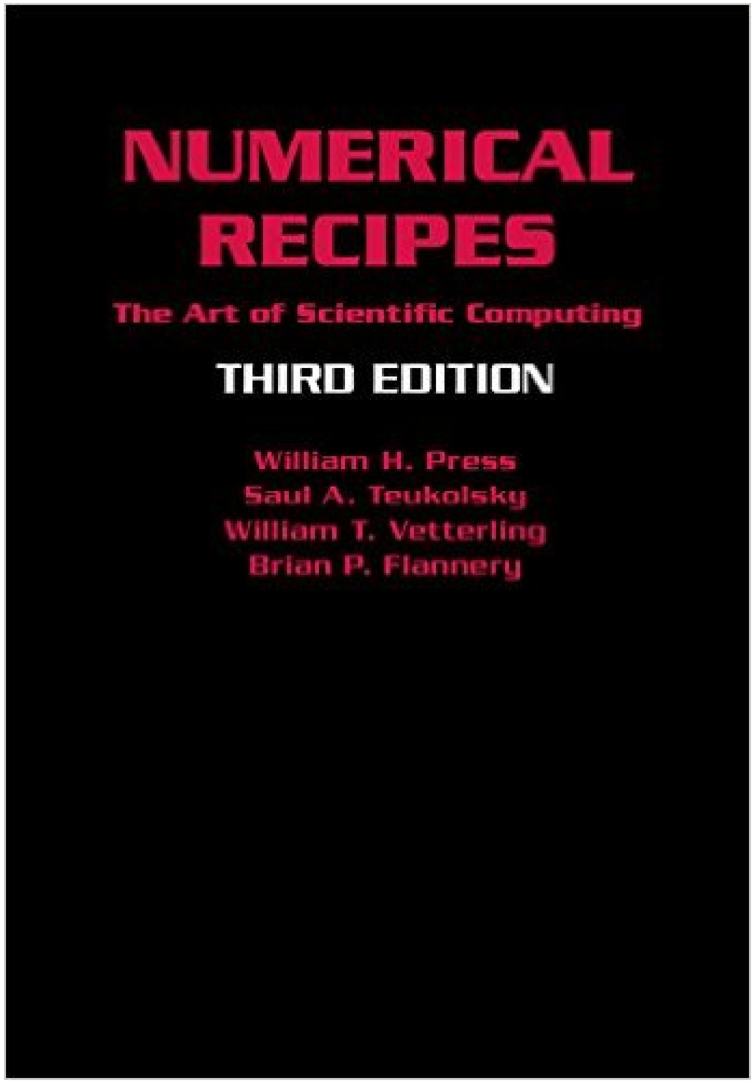


ISBN 978-0-470-75805-2

63.99 €

Statistics books

Press/Teukolsky/Vetterling/Flannery
Numerical Recipes
Cambridge Univ. PreSS 2307
70 €



Further resources

- Coryn Bailer-Jones' lecture notes on Computational Statistics, outdated! (on course web page)
- Article by David Hogg et al. (2010): Data Analysis Recipes (on course web page)
- Reference Cards for R and Emacs (on course web page)
- R project online:

  www.r-project.org
- R project related quick reference: www.statmethods.net
- Wikipedia, in particular English pages!