

Data analysis I+II

Sigbert Klinke

Chair of Statistics
Humboldt-Universität zu Berlin



General

November 3, 2022

Software packages • Data structure • Data visualization • Cheating with graphics • Famous graphs • John Snow (1855) • Florence Nigthingale (1858) • Charles Joseph Minard (1869) • Excellent graphics • Grammar of graphics • Interactive graphics • Graphics overview • Readable tables

Software packages

- Types of statistical software
 - ▶ Spreadsheets
- Which is the most used statistical software? Excel!**
- Advantages and disadvantages
 - ▶ Spreadsheets & menu-oriented statistical software: fast to learn, but not extandable
 - ▶ Statistical programming languages: difficult to learn, but can be extended
- Most major commercial statistical software packages have (free) open source counterparts
 - ▶ Excel - Apache Open/LibreOffice Calc
 - ▶ SPSS - GNU PSPP
 - ▶ S-Plus - GNU R, MATLAB - GNU Octave

- Most statistical softwares consist of
 - ▶ main program with a graphical user interface (menus)
 - ▶ programming language(s), extendable by libraries, packages, etc.
- Excel
 - ▶ Excel functions
 - ▶ Visual Basic for Applications (VBA)
- Apache OpenOffice/LibreOffice Calc
 - ▶ mimics most Excel features
- Gnumeric
 - ▶ can embed Python code
- SPSS
 - ▶ menu interface that generates a program
 - ▶ Graphics production language (GPL)
 - ▶ can embed R and Python code

- GNU PSPP
 - ▶ concentrates on building an open source programming language
- GNU R
 - ▶ only libraries, e.g. [Rcmdr](#), offer a user interface that generates a program
 - ▶ basically only programming language
- Stata
 - ▶ menu interface that generates a program
 - ▶ earlier only programming language
- GNU Gretl
 - ▶ menu interface that generates a program
 - ▶ earlier only programming language
- MATLAB & GNU Octave
 - ▶ no menu interface
 - ▶ earlier only programming language

Data structure

Variable

- collection of one measurement at several statistical units
- usually the column of a data matrix
- various data types: numeric, text, logical, date, location, ...

Observation

- collection of several measurements at one statistical unit
- usually the row of a data matrix

Dataset

- collection of observations and variables
- if possible integrated into a data matrix

Complex data

- time series
- panel data
- hierarchical data
- unstructured data
- big data
- ...

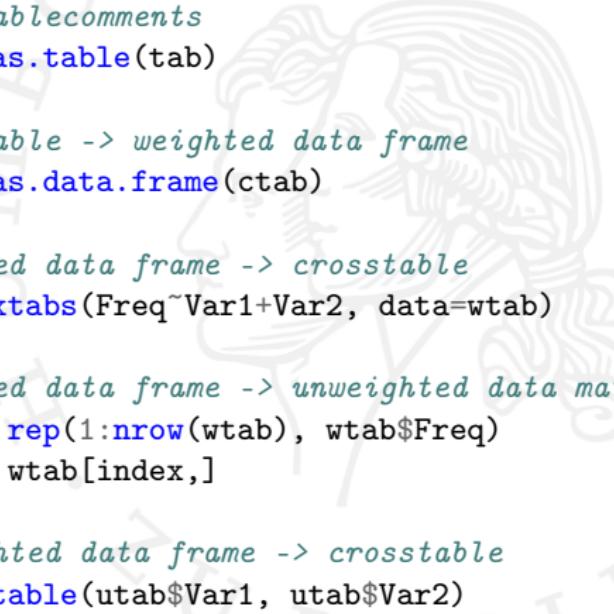
⇒ Data management e.g.

- data types and conversion
- reading and writing data
- subsetting, merging
- sorting
- aggregating
- reshaping

Example 1.1

Dataset (multivariate categorical data) : 46 persons had to say for two drugs whether they are favorable or not

	Person	1	2	...	45	46
unweighted data matrix	Drug 1	Y	Y	...	Y	N
weighted data matrix	Drug 2	N	Y	...	Y	N
	Drug 1		N	N	Y	Y
	Drug 2		N	Y	N	Y
	Frequency	12	6	6	22	
contingency table	Drug 2/Drug 1		N	Y		
	N		12	6		
	Y		6	22		

R Listing 1.1: example_data_format.R

```
1 tab <- matrix(c(12, 6, 6, 22), ncol=2)
2 tab
3 # crosstable comments
4 ctab <- as.table(tab)
5 ctab
6 # crosstable -> weighted data frame
7 wtab <- as.data.frame(ctab)
8 wtab
9 # weighted data frame -> crosstable
10 xtab <- xtabs(Freq~Var1+Var2, data=wtab)
11 xtab
12 # weighted data frame -> unweighted data matrix
13 index <- rep(1:nrow(wtab), wtab$Freq)
14 utab <- wtab[index,]
15 utab
16 # unweighted data frame -> crosstable
17 ctab <- table(utab$Var1, utab$Var2)
18 ctab
```

```
R as.table(x)
R as.data.frame(x)
R rep(x, times=1)
R xtabs(formula, data, na.action, exclude=c(NA, NaN))
R table(x, y, exclude=if(useNA=="no") c(NA, NaN), useNA=c("no",
  "ifany", "always"))
```

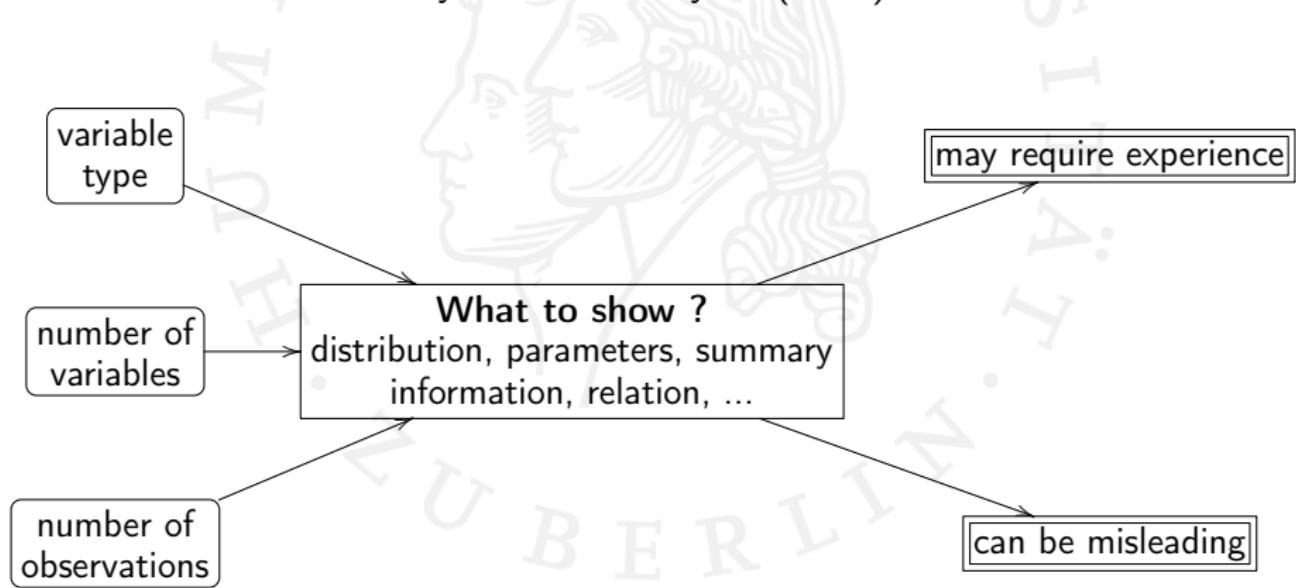
Data visualization

One Look is Worth A Thousand Words

by japanese philosopher

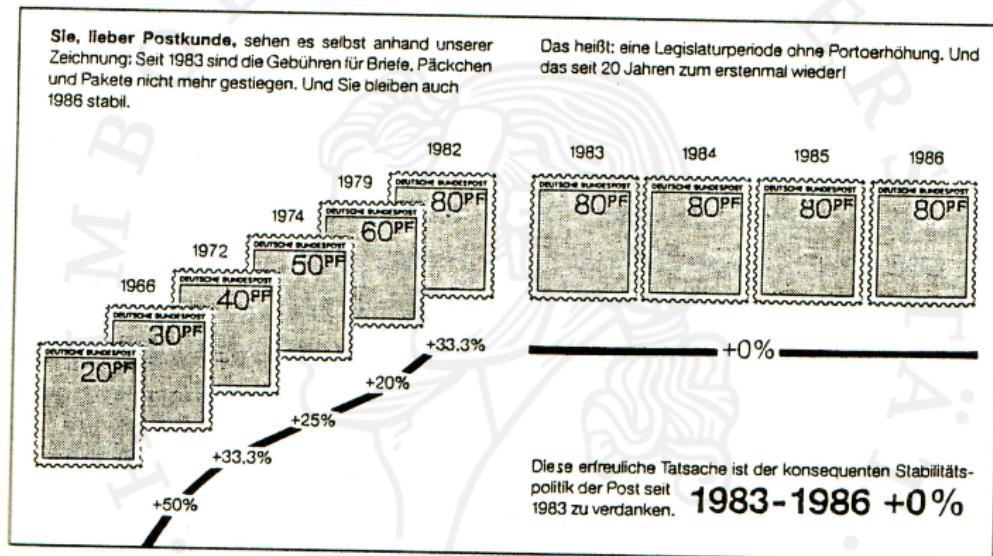
One Picture is Worth Ten Thousand Words

by Fred R. Barnyard (1927)



Cheating with graphics

Seit 1983 stabile Gebühren



Die Preisstabilität wird durch das Dehnen
der waagrechten Achse erzeugt

Krämer, W. (2009). *So lügt man mit Statistik*. Piper Taschenbuch.

Famous graphs

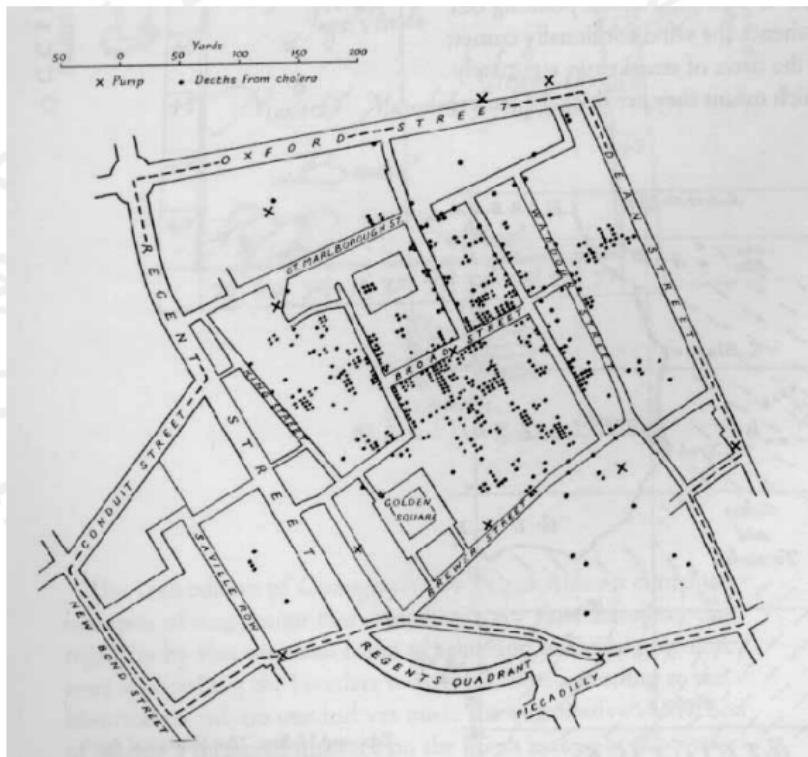
- John Snow (1855)
 - ▶ proves that cholera does not come from “bad air”
 - ▶ each point indicates a cholera case, each cross a public water pump
- Florence Nightingale (1858)
 - ▶ nurse in the Crimean war
 - ▶ red: death from wounds, black: death from other causes, blue: death from preventable diseases
- Charles Joseph Minard (1869)
 - ▶ visualizes Napoleon's invasion of Russia
 - ▶ the army's location and direction, splitting and rejoining
 - ▶ the declining size of the army, low temperatures during the retreat

Snow, J. (1855). *On the mode of communication of cholera*. London: John Churchill.

Nightingale, F. et al. (1858). *Mortality of the British Army, at home, at home and abroad, and during the Russian war, as compared with the mortality of the civil population in England*. London: Harrison and Sons.

Minard, C.J. (1869). *Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813*. Graphics Press.

John Snow (1855)

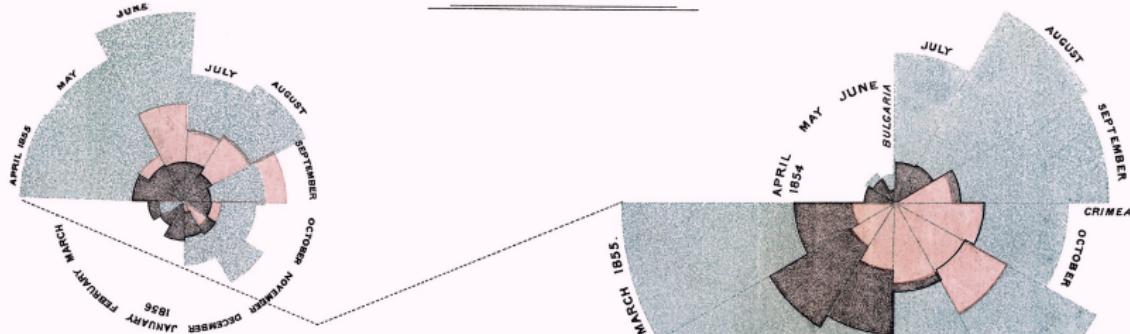


Florence Nightingale (1858)

2.
APRIL 1855 to MARCH 1856.

1.
APRIL 1854 to MARCH 1855.

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

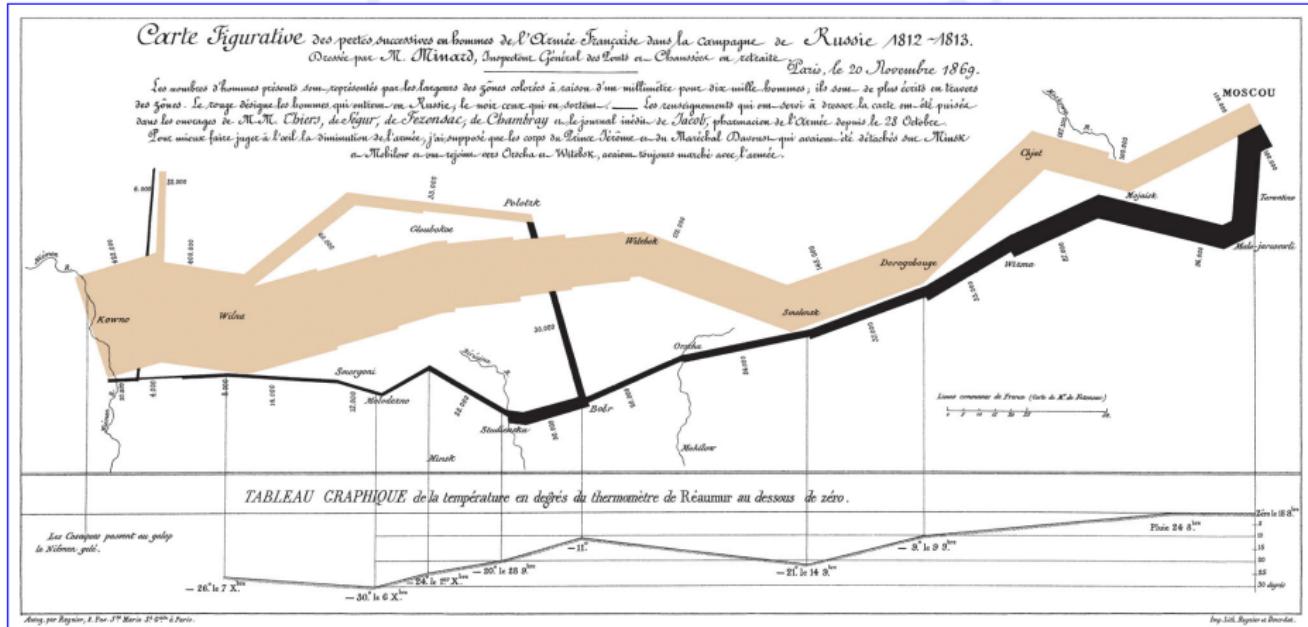
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Charles Joseph Minard (1869)



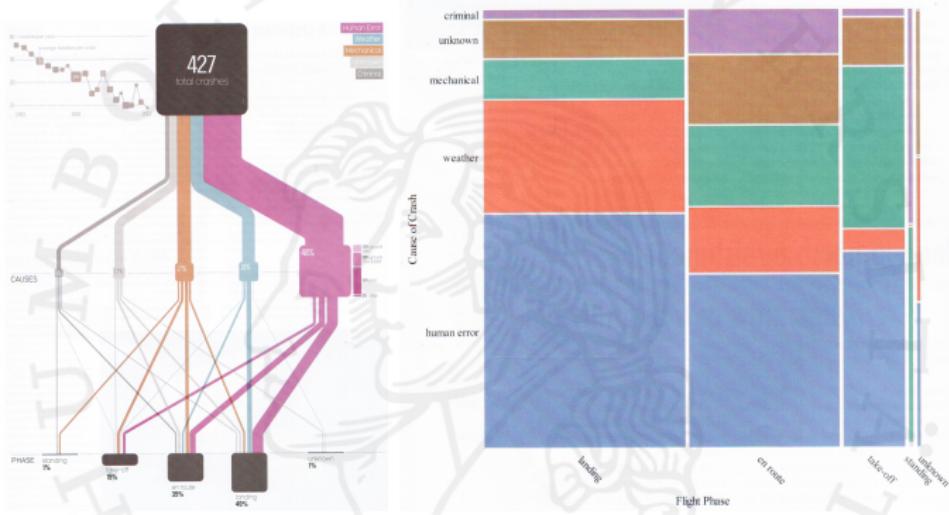
Excellent graphics

- show the data
 - ▶ think about substance rather than methodology, design, etc.
 - ▶ avoid distorting
- reveal data at several levels (broad overview → fine structure)
 - ▶ present many numbers in a small space
 - ▶ make large data sets coherent
 - ▶ encourage the eye to compare different pieces of the data
- serve a reasonable purpose: description, exploration, tabulation, decoration
- are closely integrated with statistical and verbal description of a data set

Tufte, Edward R. (2001). *The visual display of quantitative information*. 2nd ed. Cheshire, Conn: Graphics Press. 197 pp. ISBN: 978-0-9613921-4-7.

- data ink ratio = $\frac{\text{data ink}}{\text{total ink}}$ → maximize within reason
 - ▶ erase non data ink, for example chart junk as axes, boxes, grids, shading, color etc.
 - ▶ erase redundant ink
- data density = $\frac{\text{number of numbers}}{\text{area}}$ → maximize within reason
 - ▶ shrink the graphic
 - ▶ small multiples (like pictures of a movie)
- further advice
 - ▶ choose proper format and design
 - ▶ use words, numbers and drawing together
 - ▶ display an accessible complexity of detail
 - ▶ tell a story about the data

Air crash statistics (cause vs. phase)



McCandless, David (2014). *Knowledge is beautiful*. London: Collins. 255 pp. ISBN: 978-0-00-742792-5.

Wicklin, Rick (June 2015). "In praise of simple graphics". In: *Significance* 12.3, pp. 4–5. ISSN: 17409705. DOI: 10.1111/j.1740-9713.2015.00822.x. URL: <http://doi.wiley.com/10.1111/j.1740-9713.2015.00822.x> (visited on 08/15/2015).

Grammar of graphics

- Graphics generation formalized by Wilkinson et al.
 - ▶ Data: a set of data operations that creates variables from datasets,
 - ▶ Trans: convert data into a format suitable for the intended visualization,
 - ▶ Scale: convert to physical “drawing” units, e.g. require an anchor point
 - ▶ Coord: choose a coordinate system, e.g. cartesian, polar, ...
 - ▶ Element: a graph and its aesthetic attributes,
 - ▶ Guide: one or more guides, e.g. axes and legends
- implemented in R (Wickham) and SPSS (Wilkinson)
- Opposite approach: Excel with various (non-combinable) graphics

Wickham, Hadley (2009). *Ggplot2: elegant graphics for data analysis*. Use R! New York: Springer. 212 pp. ISBN: 978-0-387-98140-6.

Wilkinson, Leland and Wills, Graham (2011). *The grammar of graphics*. 2. ed., softcover reprint of the hardcover 2. ed., 2005. Statistics and computing. New York, NY: Springer. 690 pp. ISBN: 978-1-4419-2033-1.

Example 1.2 (Histogram)

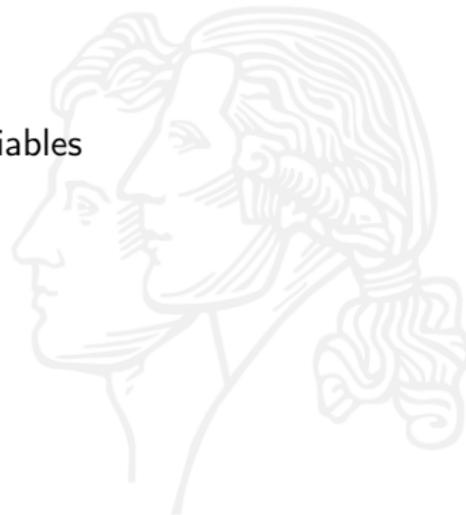
- Data: select one variable from the data set
- Trans: bin data based on class borders and compute frequencies and densities
- Scale: linear transformation, ensure that zero is visible on y-axis
- Coord: choose a cartesian coordinate system for plotting
- Element: choose bars for drawing
- Guide: add an axis

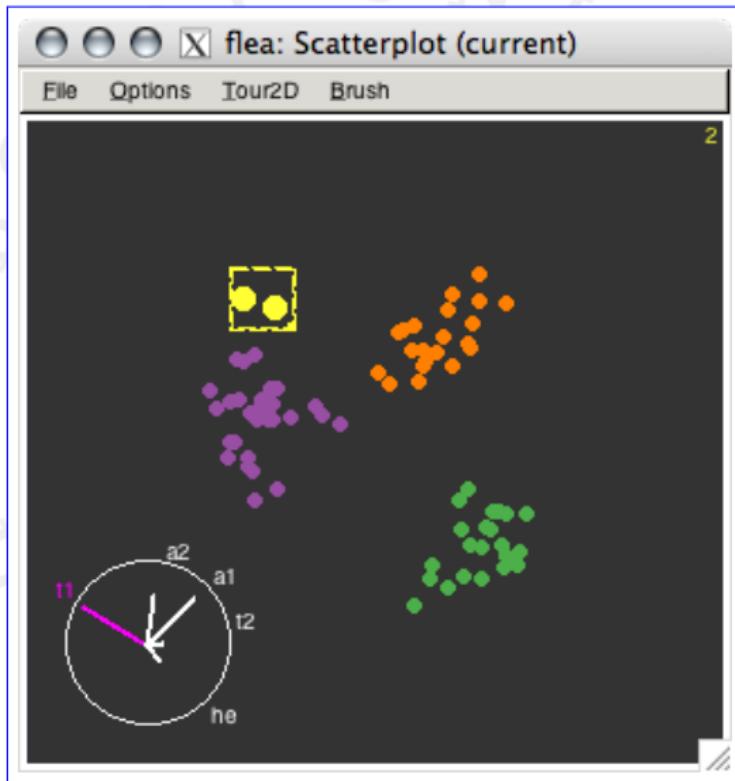
Example 1.3 (Scatterplot)

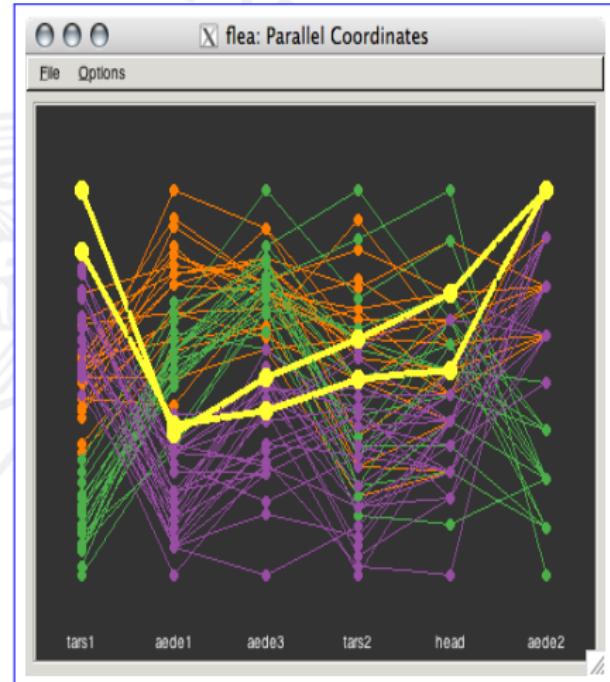
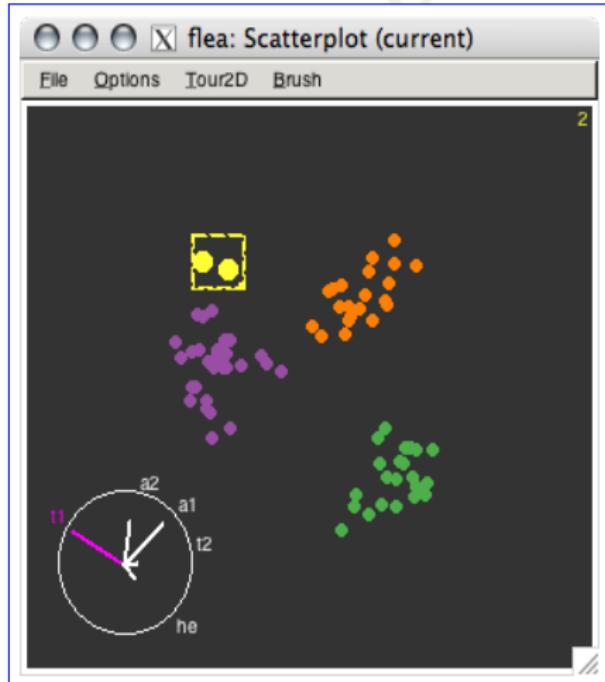
- Data: select two variables from the data set
- Trans: set x to the first variable and y to the second variable
- Scale: linear transformation, e.g. ensure $(0, 0)$ is visible
- Coord: choose a cartesian coordinate system for plotting
- Element: choose points for drawing
- Guide: add axes

Interactive graphics

- Techniques:
 - ▶ rotating
 - ▶ touring
 - ▶ reordering of variables
 - ▶ linking
 - ▶ querying
 - ▶ selecting
 - ▶ highlighting
 - ▶ brushing
 - ▶ panning
 - ▶ zooming
 - ▶ fisheye view
- Problem: some techniques can not be shown on paper







R Listing 1.2: example_ggobi.R

```
1 library("MASS") # for Boston Housing data
2 if (interactive()) {
3   library("rggobi")
4   g <- ggobi(Boston)
5   # now do some brushing
6   col <- glyph.colour(g["Boston"])
7   table(col)
8 }
```

R rggobi::ggobi(x, display)
R rggobi::glyph_...(obj)

Graphics overview

The screenshot shows the MM*Stat software interface. On the left, there are three dropdown menus: 'Analysis variable(s)', 'Grouping variable(s)', and 'Unused variable(s)'. The 'Unused variable(s)' menu contains variables: x1, x2, x3, x4, g1, g2, g3, g4, g5, g6, g7. At the top, there is a navigation bar with tabs: Plot, R code, R Help, Variables, and Log. Below the navigation bar, the title 'smvgraph 0.2.0' is displayed. The main content area is a table with two columns: '#Analysis' and '#Grouping'. The '#Analysis' column lists various plotting and analysis functions. The '#Grouping' column lists functions that require grouping variables. A note at the bottom states: 'Note: The availability of a plot depends on the required libraries (see "Log") and the number of unique values in a variable!' Below the table, a section titled 'Plots and algorithms used from' is shown, listing packages and their authors.

#Analysis	#Grouping
1 Bar chart, Box plot, Cumulative Distribution, Histogram, Index plot, Kernel density, Needle chart, Q-Q norm plot, Strip chart, Time series, Violin plot	0 Box plot, Cumulative Distribution, Dot plot, Histogram, Index plot, Kernel density, Pyramid plot, Q-Q norm plot, Strip chart, Violin plot
2 Bagplot, Bar chart, Data distances, DBSCAN clustering, EM clustering, Factor analysis, Hierarchical clustering, K-Means clustering, Kernel density, Mosaic plot, Q-Q plot, Scatter plot, Scatter plot matrix, Sunflower plot, Table plot	1 Box plot, Cumulative Distribution, Dot plot, Histogram, Index plot, Kernel density, Q-Q norm plot, Strip chart, Violin plot
3 Andrews curves, Bagplot, Bar chart, Chernoff faces, Data distances, DBSCAN clustering, EM clustering, Factor analysis, Hierarchical clustering, K-Means clustering, Mosaic plot, Parallel coordinate plot, Radar chart, Scagnostics, Scatter plot 3D, Scatter plot matrix, Table plot	2 Factor analysis, Q-Q plot, Scatter plot, Scatter plot matrix
4+ Andrews curves, Bagplot, Bar chart, Chernoff faces, Data distances, DBSCAN clustering, EM clustering, Factor analysis, Hierarchical clustering, K-Means clustering, Mosaic plot, Parallel coordinate plot, Radar chart, Scagnostics, Scatter plot matrix, Table plot	3 Andrews curves, Factor analysis, Parallel coordinate plot, Radar chart, Scatter plot 3D, Scatter plot matrix
	4 Andrews curves, Factor analysis, Parallel coordinate plot, Radar chart, Scatter plot matrix

Note: The availability of a plot depends on the required libraries (see "Log") and the number of unique values in a variable!

Plots and algorithms used from

Package	Author(s)
apck	Hans-Peter Wolf [aut, cre]
car	John Fox [aut, cre], Sanford Weisberg [aut], Brad Price [aut], Daniel Adler [ctb], Douglas Bates [ctb], Gabriel Baud-Bovy [ctb], Ben Bolker [ctb], Steve Ellison [ctb], David Firth [ctb], Michael Friendly [ctb], Gregor Gorjanc [ctb], Spencer Graves [ctb], Richard Heiberger [ctb], Pavel Krtikovsky [ctb], Rafael Labossiere [ctb], Martin Maechler [ctb], Georges

Listing 1.3: `smvgraph.R`

```

1 # devtools::install_github("sigbertklinke/smugraph")
2 library("smvgraph")
3 splot() # or splot(iris)

```

Learn ggplot2 by web-r.org

Make plot with `ggplot2` package of R. Please select data and variables for draw plot. You can download plot(s) as a [pdf](#) or [png](#) format. You can make multiplot. You can download upto 10 plots as a [powerpoint file](#) with or without R code.

Select Language

- English 한국어(Korean)

Web-R.org ggplot2 MultiPlot Powerpoint List Sessioninfo About Interactive plot

Please [select](#) one of sample data or [upload](#) your own. You can upload data as a [csv](#), [xlsx](#) (Microsoft Excel), [dbf](#) (dbase 3+), [sav](#) (SPSS), [dta](#) (STATA), [sas7bdat](#) (SAS) format. If you have any error, please upload as a [csv](#) format.

Upload data(*.xlsx,*.csv)

No file selected

Select data

Salaries
 mtcars
 acs
 radial
 iris
 heightweight

Please press [Reset Variables/Options](#) button before select a sample data.

show help for data

Browse Examples

Do preprocessing

You can preprocess the data by entering the R command(s) here. Please uncheck the checkbox before enter/change the R command(s) and recheck the checkbox.

Enter the name of data

Select data preprocessing

Example gallery

Source: [Web-R.org](#), a Shiny app to build ggplot2 plots

Readable tables

- do you really need the table?
- caption
 - ▶ should be detailed enough
 - ▶ might be placed on different page
- text/table linking
 - ▶ number your tables separately
 - ▶ explain your tables in the text
- layout
 - ▶ guides the eye: no vertical lines, no double lines, use background colors instead of a net
 - ▶ the number of columns and rows should be as small as possible
- table head
 - ▶ should contain the measurement units

- columns and rows
 - ▶ should be ordered naturally or by size
 - ▶ close with them with sums, means etc. if possible
- content
 - ▶ make two tables if the information is unrelated
 - ▶ emphasize important information
- text
 - ▶ should be left justified
- numbers
 - ▶ should be aligned by decimal points
 - ▶ should be right justified
 - ▶ should be limited to significant digits
 - ▶ should have a leading zero, e.g. *0.1* instead of *.1*

Ostermann, R., Wilhelm, AFX., and Wolf-Ostermann, K. (2004). "Komplizierter als man denkt. Präsentation statistischer Daten in der Pflege, Teil 1 - Tabellen". In: *Pflegezeitschrift* 57.1, pp. 18–21.

	Brown	Blue	Hazel	Green
Black	40.135135	39.222973	16.966216	11.675676
Brown	106.283784	103.868243	44.929054	30.918919
Red	26.385135	25.785473	11.153716	7.675676
Blond	47.195946	46.123311	19.951014	13.729730

With `xtable::xtable`, but no caption by default

Hair / Eye	Brown	Blue	Hazel	Green	Sum
Black	40.1	39.2	16.9	11.6	108.0
Brown	106.2	103.8	44.9	30.9	286.0
Red	26.3	25.7	11.1	7.6	71.0
Blond	47.1	46.1	19.9	13.7	127.0
Sum	220.0	215.0	93.0	64.0	592.0

Figure 1: Expected frequencies under independence for the *Hair and Eye Color of Statistics Students* data set.

Data analysis I+II

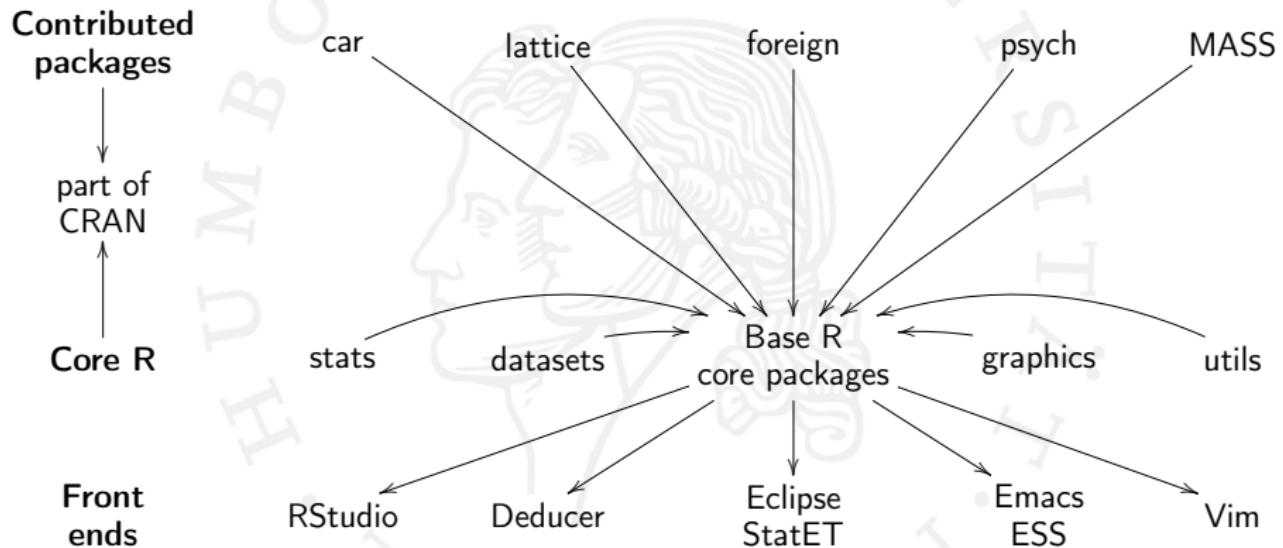
R

November 3, 2022

- Components • Packages • Help • Vector • Matrix/Array • List • Data frame • Object conversion • Tibble • Reading and writing data sets • Data management • Formula interface • Creating graphs • Parameters for plot
• Saving graphs

Components

- R is a programming environment
 - ▶ has a well-developed but simple programming language
 - ▶ allows rapid development of new tools according to user demand
 - ▶ tools are distributed as packages, which any user can download to customize the R environment
 - ▶ new R version appears approximately every six months
- Base R and most R packages are available for download
 - ▶ from cran.r-project.org (CRAN)
 - ▶ base R comes with a number of basic data management, analysis, and graphical tools
 - ▶ R's power and flexibility arises from the packages (more than 8000!)



• Rstudio, an Integrated Development Environment (IDE)

Source pane

```

File Edit Code View Plots Session Project Build Tools Help
tempfeuchte.R
  Source on Save Run Source
1 setwd("~/Arbeitsfläche/windowsXP/datalogger")
2
3 files <- c("130221_wohnzimmer.txt", "130223_bad.txt", "130224_schlafzimmer.txt", "130303_fuer.txt", "130310_wohnzimmer.txt",
4 sel <- select.list(files, multiple=F, title="Datalauswahl", graphics=FALSE)
5
6 d = 18
7 temp <- c()
8 rlef <- c()
9 for (l in seq(files)) {
10   if (l == 1) {
11     xl <- read.csv(files[l], skip=0, header=F, sep="\t", dec=",")
12     xl <- xl[,!(row.names(xl) == 1)]
13     temp <- c(temp, xl$V4)
14     rlef <- c(rlef, xl$V5)
15   }
16
17 temp <- range(temp) + c(-1,1)
18 rlef <- range(rlef) + c(-5,5)
19
20 par(mfcol=c(2, length(sel)), mar=c(1,4,1,0))
21 hrs <- 3600*6
22 day <- 3600*24
23 yaxs <- "l"
24 for (l in seq(files)) {
25   if (l != 1) rm(sel)
26   xl <- read.csv(files[l], skip=0, header=F, sep="\t", dec=",")
27   xl <- xl[,!(row.names(xl) == 1)]
28
29 names(xl) <- c("Zeit", "Temp", "RelFeuchte", "Taupunkt")
30 xl <- strptime(paste(xl$Zeit, xl$Zeit), "%d-%m-%Y %H:%M:%S")
31 xl <- strptime(paste(xl$Taupunkt), "%H:%M:%S")
32 vdt <- floor(max(xl$V1)*vdt)
33 hrs <- vdt*hrs
34
35 sel <- sel[which(sel != sel[1])]
```

Environment pane

Plot and help pane

Console pane

```

Console ->~/Arbeitsfläche/windowsXP/datalogger/ >R
1 tfl1 <- filter(xlSelFeuchte, rep(1/60, 60))
2 ylim <- range(tfl1, na.rm=TRUE)c(-5,5)
3 plot(xlSel1, x1SelFeuchte, type="l", lwd=2, ylim=ylim, axes=FALSE)
4 axis(1, at=1:60, col="grey50")
5 if (yaxs) axis(2, col="red")
6 box()
7 lines(xlSel1, tfl1[,5, col="red"])
8 lines(xlSel1, tfl1[,1, col="red"])
9 abline(v=vt, col="grey50")
10 abline(v=vt, col="grey50")
11 if (tfl1 <- row(xl$V1)/62)
12 text(xl$V1[tfl1], tfl1[,2], files[1])
13
14 par(mar=c(1,0,5,1,0))
15 yaxs <- F
16 x <- rbIn(x1,x1)
17
[1] "2013-03-10 CET" "2013-03-11 CET" "2013-03-12 CET" "2013-03-13 CET" "2013-03-14 CET" "2013-03-15 CET" "2013-03-16 CET"
[8] "2013-03-17 CET" "2013-03-18 CET" "2013-03-19 CET" "2013-03-20 CET" "2013-03-21 CET" "2013-03-22 CET"
```

Packages

- Install package(s), to be done only once per installation
 - update all packages, e.g. after updating R itself
 - ⌚ `update.packages()`
 - Load/attach package
 - ⌚ `library(package)`
 - ⌚ `require(package)`
 - Information about attached packages
 - ⌚ `sessionInfo(package = NULL)`
 - ⌚ Plot and help pane → Packages
- ⚠ Use `library("MASS")` instead of `library(MASS)`



Listing 2.1: example_install.R

```
1 if(interactive()) {  
2   # install from an official CRAN server  
3   install.packages("foreign")  
4   # install from a specific CRAN server  
5   install.packages("foreign", repos="https://ftp.gwdg.de/pub/misc/cran")  
6   # install from GitHub  
7   install.packages("devtools")  
8   library("devtools")  
9   install_github("simecek/additivityTests")  
10 }
```

• `install.packages(pkgs, repos =getOption("repos"))`

• `devtools::install_github(repo)`

Help

- General help
 - ▶ R mailing lists
 - ▶ Contributed documentation
 - ▶ Task views
 - ▶ Quick R
 - ▶ stackoverflow.com
 - ▶ www.cookbook-r.com
- Help about object obj
 - ◀ ?obj
 - ◀ help(obj)

Teator, Paul (2011). *R cookbook*. 1st ed. Beijing ; Sebastopol, CA: O'Reilly. ISBN: 978-0-596-80915-7.

Chang, Winston (2013). *R graphics cookbook: [practical recipes for visualizing data]*. eng. 1st ed. Beijing: O'Reilly. ISBN: 978-1-4493-1695-2 1-4493-1695-6 978-1-4493-1695-2.

- Search help about topic `topic`
 - ☞ `??topic`
 - ☞ `help.search(pattern)`
 - ☞ `RSiteSearch(topic)`
 - Help on package(s)
 - ☞ `library(help=pkg)`
 - ☞ `help(package=pkg)`
 - Viewing an object `obj` in the *Console pane*
 - ▶ for a data object it shows (partly) the data
 - ▶ for a function it shows the source code
- ☞ `obj`

Vector

- vector: collection of elements with the same data type identified by an index
- Data types of a vector
 - ▶ logical - boolean (includes NA)
 - ▶ integer - integer numbers
 - ▶ double - real numbers
 - ▶ complex - complex numbers
 - ▶ character - strings
- Creation
 - ◀ R c(val₁, ..., val_n)
 - ◀ R vector(mode="logical", length=0)
- Length (number of elements)
 - ◀ R length(x)
- Access
 - ◀ R x[index] or x[vector]



Listing 2.2: example_vector.R

```
1 vn <- c(1, 2, 3, 4, 5)
2 vn
3 vn <- 1:10
4 vn
5 vt <- c("a", "b", "c", "d")
6 vt
7 vb <- c(TRUE, FALSE, NA)
8 vb
9 length(vn)
10 length(vb)
11 vt[1]
12 vt[-1]
13 vt[c(1,2,3)]
14 vt[2:3]
15 vn <- c("eins"=1, "zwei"=2, "drei"=3, "vier"=4, "fuenf"=5)
16 vn
17 vn["drei"]
18 names(vn)
```

Matrix/Array

- array: a vector plus a dimension information

- ▶ vector (n)
 - ▶ matrix ($n \times p$)
 - ▶ array ($n_1 \times n_2 \times \dots \times n_d$)
 - ▶ intended for matrix computation

- Creation

⌚ `matrix(data=NA, nrow=1, ncol=1, byrow=F, dimnames=NULL)`
⌚ `array(data=NA, dim=length(data), dimnames=NULL)`

- Length (number of elements)

⌚ `length(x)`

- Dimension

⌚ `dim(x)`

- Access

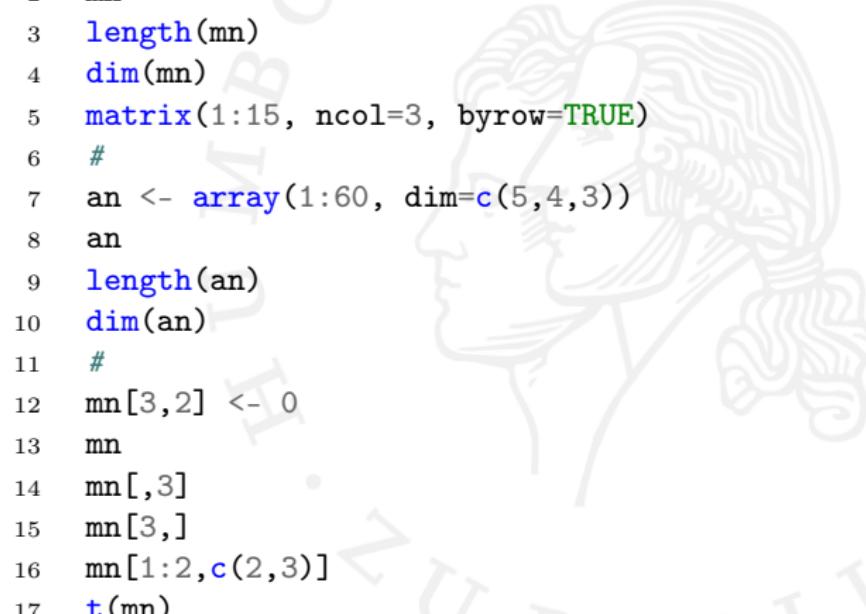
- ▶ names, indices or vectors are possible
 - ▶ mixtures of names, indices or vectors are possible
 - ▶ negative integer values are interpreted as everything except i
- ☞ `x[i1, ..., id] or x[vector1, ..., vectord]`

- Names

- ☞ `rownames(x)`
- ☞ `colnames(x)`
- ☞ `dimnames(x)`

- Special operators and functions

- ☞ `x %*% y` matrix multiplication
- ☞ `x %o% y` outer product xy^T
- ☞ `crossproduct(x, y=x)`
- ☞ `t(x)`

A faint watermark of a classical bust of a person's head, possibly a Greek or Roman figure, is visible in the background of the slide.

R Listing 2.3: example_matrix.R

```
1 mn <- matrix(1:15, ncol=3)
2 mn
3 length(mn)
4 dim(mn)
5 matrix(1:15, ncol=3, byrow=TRUE)
6 #
7 an <- array(1:60, dim=c(5,4,3))
8 an
9 length(an)
10 dim(an)
11 #
12 mn[3,2] <- 0
13 mn
14 mn[,3]
15 mn[3,]
16 mn[1:2,c(2,3)]
17 t(mn)
```

List

- list: collection of objects identified by a name or an index

- Creation

- ☞ `list(name1=obj1, ..., namen=objn)`

- ☞ `vector(mode="list", length=0)`

- Length (number of elements)

- ☞ `length(x)`

- Access

- ☞ `x[[i]]` or `x[[name]]` or `x$name`

- ⚠ `x[i]` or `x[name]` delivers a list with one element, not the element itself!

- Names of list elements

- ☞ `names(x)`



Listing 2.4: example_list.R

```
1 l <- list(a=1, b=TRUE, c="text", NA)
2 l
3 length(l)
4 l$a
5 l[[1]]
6 l[['a']]
7 l[1]
8 names(l)
```

Data frame

- `dataframe`: a list, but all objects have the same length (= data set)
- Creation
 - `R data.frame(name1=obj1, ..., namen=objn)`
- Length (number of elements)
 - `R length(x)`
- Dimension
 - `R dim(x)`
- Access element(s)
 - `R x[i,j] or x[i,name] or x[i,vector]`
- Access row(s) or observation(s)
 - `R x[i,] or x[name,] or x[vector,]`
- Access columns(s) or variable(s)
 - `R x[,i] or x[,name] or x[,vector]`



Listing 2.5: example_access.R

```
1 library("MASS")                      # for Boston Housing data
2 nx <- Boston[-506,]                  # without obs. 506
3 dim(nx)
4 mx <- Boston[,-c(4,8,14)] # without CHAS, RAD and MEDV
5 dim(mx)
6 head(Boston)
7 tail(Boston)
```

- First (six) rows of a data frame

`R head(x)`

- Last (six) rows of a data frame

`R tail(x)`

- Names

`R names(x)`

`R rownames(x)`

`R colnames(x)`

`R dimnames(x)`

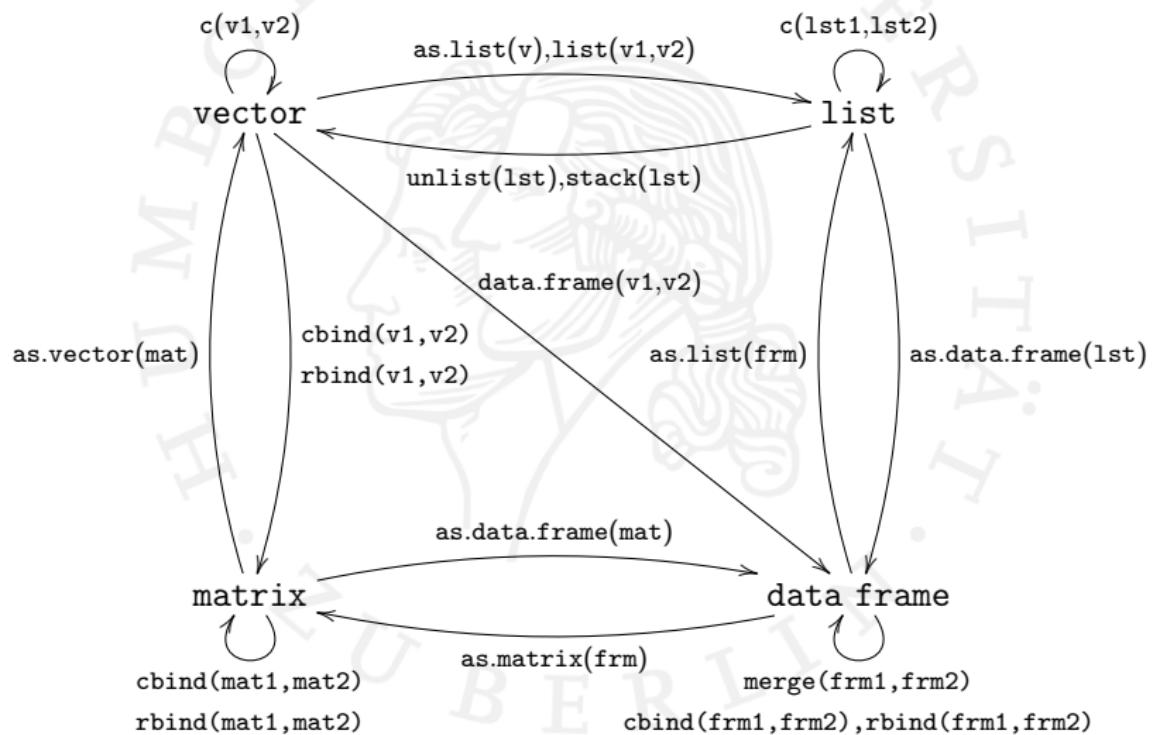
- Using variables as vectors

`R attach(x)`

⚠ `attach` copies(!) all variables in data frame as separate objects!

- Invoke a read-only text editor on a object
  `edit(x)`
- Show structure of an object
  `str(x)`
- Write an object to console (or file)
  `dput(x, file="")`
- Read an object from file
  `dget(file)`
- Summarize object (generic function)
  `summary(x, ...)`
- Plot object (generic function)
  `plot(x, ...)`

Object conversion



Tibble

- Tibbles are a modern take on data frames
- Differences to data frames
 - ▶ printing
 - ▶ subsetting: always a tibble is returned
 - ▶ \$: does not support partial matching
 - ▶ recycling: only for values of length 1
- part of [tidyverse](#)

The tidyverse is a set of packages that work in harmony because they share common data representations and API design

Wickham, Hadley and Grolemund, Garrett (2016). *R for data science: import, tidy, transform, visualize, and model data*. First edition. OCLC: ocn968213225. Sebastopol, CA: O'Reilly. 492 pp. ISBN: 978-1-4919-1039-9 978-1-4919-1036-8.

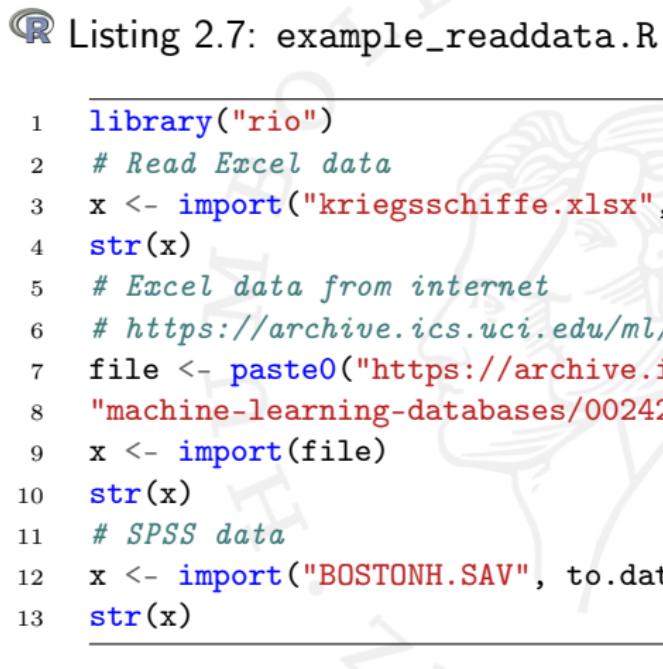


Listing 2.6: example_tibble.R

```
1 library("MASS")
2 # data frame
3 class(Boston)
4 head(Boston)
5 Boston$z
6 class(Boston$z)
7 Boston[506,]
8 # tibble
9 library("tibble")
10 nx <- as_tibble(Boston)
11 nx
12 dim(nx)
13 nx[506,]
14 head(nx)
```

Reading and writing data sets

- offers a lot of packages to read/write data in various formats:
 - ▶ base (R binary formats)
 - ▶ utils, readr (ACSII text files)
 - ▶ foreign, haven (various statistical software)
 - ▶ readxl, xlsx (Excel 97-2016)
 - ▶ feather (Python), jsonlite (JSON)
 - ▶ xml2 (XML, HTML), yaml (YAML)
- rio uses all these packages
 - ◀ `rio::import(file, format, setclass, which, ...)`
 - ◀ `rio::export(x, file, format, ...)`
- ▶ file format is determined by file extension or format
- ▶ ... additional arguments for the underlying export functions
- ▶ see the package vignette [Introduction to rio](#) for further information

R Listing 2.7: example_readdata.R

```
1 library("rio")
2 # Read Excel data
3 x <- import("kriegsschiffe.xlsx", sheet=1)
4 str(x)
5 # Excel data from internet
6 # https://archive.ics.uci.edu/ml/datasets/energy+efficiency
7 file <- paste0("https://archive.ics.uci.edu/ml/",
8 "machine-learning-databases/00242/ENB2012_data.xlsx")
9 x <- import(file)
10 str(x)
11 # SPSS data
12 x <- import("BOSTONH.SAV", to.data.frame=TRUE)
13 str(x)
```

Data management

- sorting observations
 - `R sort(x, partial=NULL, na.last=NA, decreasing=F, index.return=F)`
- index of ordered observations
 - `R order(x, partial=NULL, na.last=T, decreasing=F)`
- merging observations
 - `R merge(x, y, by=intersect(names(x), names(y)), all=F)`
- aggregating
 - `R aggregate(x, by, func, ...)`
 - `R by(x, groups, func, ...)`
 - `R tapply(x, groups, func=NULL)`
- choosing a subset
 - `R subset(x, condition, select)`
 - `R which(condition, arr.ind=F)`
 - `R Filter(func, x)`
- splitting observations
 - `R split(x, factor)`



Listing 2.8: example_attach.R

```
1 library("MASS")      # for Boston Housing data
2 attach(Boston)
3 medv
4 medv <- rep(NA, 506)
5 medv                  # medv changed
6 str(Boston)           # Boston$medv not changed
```



Listing 2.9: example_sort.R

```
1 library("MASS") # for Boston Housing data
2 smedv <- sort(Boston$medv)
3 smedv
4 smedv <- sort(Boston$medv, index.return=T)
5 smedv$x
6 smedv$ix
7 order(Boston$medv) # does the same as smedv$ix
```



Listing 2.10: example_aggregate.R

```
1 library("MASS")      # for Boston Housing data
2
3 aggregate(Boston$medv, list(Boston$rad), mean)
4 aggregate(medv~rad, data=Boston, mean)
5 aggregate(Boston$medv, list(Boston$rad, Boston$chas), mean)
6 aggregate(medv~rad+chas, data=Boston, mean)
7 aggregate(Boston, list(Boston$rad), mean)
8 aggregate(Boston, list(Boston$rad, Boston$chas), mean)
9
10 by(Boston$medv, Boston$rad, mean)
11 by(Boston, Boston$rad, mean) # warning
12 by(Boston$medv, list(Boston$rad, Boston$chas), mean)
13 by(Boston, list(Boston$rad, Boston$chas), mean) # warning
```

Formula interface

$y \sim x$	y grouped by / depends on x
$+x$	include variable x
$-x$	delete variable x
$x:z$	include the interaction between x and z
$x*z$	include the variables x and z and their interactions
x/z	nesting: include z nested within x
$x z$	conditioning: include x given z
$(x+y+z)^3$	include these variables and all interactions up to three way
$\text{poly}(x, 3)$	polynomial inclusion
$I(x*z)$	include a new variable as product of x and z
1	intercept

 Listing 2.11: example_formula.R

```
1 library("MASS")                      # for Boston Housing data
2 par(mfrow=c(2,2))                     # four plots in one
3 # use variables
4 plot(Boston$lstat, Boston$medv, pch=19, cex=0.5)
5 # use with
6 with(Boston, plot(lstat, medv, pch=19, cex=0.5, col="red"))
7 # use attach
8 attach(Boston)
9 plot(lstat, medv, pch=19, cex=0.5, col="green")
10 # use formula
11 plot(medv~lstat, data=Boston, pch=19, cex=0.5, col="blue")
```

Creating graphs

- R supports different graphic systems
- “Standard” graphics: package `graphics`
- Trellis displays (small multiples) (Tufte, 1983 & 1990, Sarkar, 2008): package `lattice`
- Graphics formally (re-)defined (Wilkinson et al., Wickham): package `ggplot2`

Tufte, Edward (1983). *The Visual Display of Quantitative Information*.

<http://vis.cs.brown.edu/results/bibtex/Tufte-1983-VDQ.bib>(bibtex:
Tufte-1983-VDQ). Graphics Press.

— (1990). *Envisioning Information*. Graphics Press Cheshire.

Sarkar, Deepayan (2008). *Lattice: multivariate data visualization with R*. Use R! New York:
Springer. ISBN: 978-0-387-75968-5 978-0-387-75969-2.

Wickham, Hadley (2009). *Ggplot2: elegant graphics for data analysis*. Use R! New York:
Springer. 212 pp. ISBN: 978-0-387-98140-6.

Wilkinson, Leland and Wills, Graham (2011). *The grammar of graphics*. 2. ed., softcover reprint
of the hardcover 2. ed., 2005. Statistics and computing. New York, NY: Springer. 690 pp.
ISBN: 978-1-4419-2033-1.

R Listing 2.12: example_scatterplot_plot.R

```
1 library("MASS") # for Boston Housing data
2 # standard graphics
3 plot(Boston$lstat, Boston$medv)
```

R Listing 2.13: example_scatterplot_lattice.R

```
1 library("MASS") # for Boston Housing data
2 # lattice
3 library("lattice")
4 xyplot(medv~lstat, data=Boston)
```

R Listing 2.14: example_scatterplot_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 # ggplot2
3 library("ggplot2")
4 ggplot(Boston, aes(x=lstat, y=medv)) + geom_point(shape=1)
```

Parameters for plot

- type graphic type, e.g.
 - ▶ "p" only points
 - ▶ "l" only lines
 - ▶ "h" needle plot (\rightarrow bar chart)
 - ▶ "n" no plotting
- main title
- sub subtitle
- xlab label for x-axis
- ylab label for y-axis
- Text in plots may contain greek symbols etc.

R Listing 2.15: example_annotation.R

```
1 x <- (-300:300)/100
2 plot(x, dnorm(x), type="l",
3       main=expression(paste("Density function of X~N(", mu, "=0,", sigma
```

- axes logical for drawing axes
- ann logical for plot annotations
- xlim x limits of plotting window
- ylim y limits of plotting window
- col color for points and lines
 - ▶ col.main, col.sub, ...
- cex size of text and plotting symbols
 - ▶ cex.main, cex.sub, ...
- pch plotting symbol for points
- cex size of text and plotting symbols
- lty line type
- lwd line width

Saving graphs

- Usually R plots to the screen (default device)
- Screen output can be saved by the context menu (right mouse click)
- Devices to plot to a file are available
- Plotting to the device is closed by `dev.off()`
- Several plots will create “pages” in the file or separate files (see `onefile`)

```
❷ pdf(file, width=7, height=7, bg="transparent", onefile=T)
❷ svg(filename, width=7, height=7, bg="white", onefile=T)
❷ png(filename, width=480, height=480, bg="white", onefile=T)
❷ dev.off()
```

Basics and data preparation

November 3, 2022

- Statistics • Level of measurement • Stevens typology • Data quality •
- Classification of Data Accuracy & Reliability • Political influence • Data cleansing • Tableplot • Sampling • The Literary Digest Desaster • Random sampling • Estimator properties • Other sampling methods • Estimator properties • Representative vs. random samples

Statistics

- Statistics: science of the
 - ▶ collection,
 - ▶ analysis,
 - ▶ interpretation,
 - ▶ presentation, and
 - ▶ organization of data
- Types of statistics
 - ▶ descriptive: describe data
 - ▶ explorative: explore data to generate hypotheses
 - ▶ inferential: draw conclusions about a population from a sample

- Analysis methods
 - ▶ coefficients: summarize information
 - ▶ graphics and table: visualize information
 - ▶ confidence intervals: find probable “values”
 - ▶ tests: validate hypotheses about the population
 - ▶ models: predict
- Analysis objects
 - ▶ distribution(s)
 - ▶ parameters of distribution(s)
- Dataset parameters
 - ▶ number of variables
 - ▶ number of observations

On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte.

One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it.

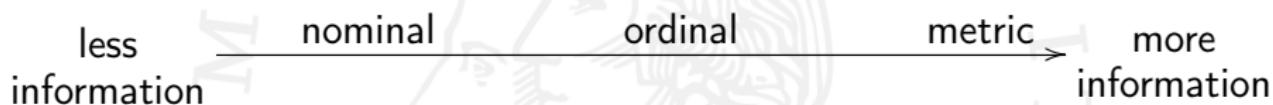
Laplace, Pierre Simon de (1812). "Théorie analytique des probabilités". In: DOI: 10.3931/e-rara-9457. URL: <http://dx.doi.org/10.3931/e-rara-9457> (visited on 08/16/2015).

Level of measurement

nominal values are either equal or unequal, e.g. hair color, eye color

ordinal values can be ranked/sorted, e.g. grades in exam

metric values can be used for calculation



Warning

- do not apply inappropriate methods
 - ▶ but if you do then make sure you know what you do
- if unsure then apply low-information methods
 - ▶ if you do not use all information then your conclusions will be too conservative
 - ▶ if you assume information which is not true then your conclusions will be wrong!

What has been done

- up to six samples were taken from cancer patients
- to each sample marker were applied
- color strength was judged on a scale from 0 to 3
- Intensity of color was averaged over the samples

Is it really ok?

- patient 1: all six samples with strength 1
- patient 2: two samples with strength 3, four with strength 0
- average values on both cases are 1

What has been done

- two teachers examined the same students (grades from 1 to 5)
- the grades are averaged to judge the students

Is it really ok?

- teacher A used all five grades, teacher B only the grades from 2 to 5
- student 1: grade A = 3 and grade B = 4
- student 2: grade A = 4 and grade B = 3
- average values on both cases are 3,5

Stevens typology

- Stevens (1946) presented a hierarchy of data scales
- based on invariance of specific transformation
- s is a *scale* (map) which assigns to the elements of a set S real numbers ($s : S \rightarrow \mathbb{R}$)
 - ▶ a scale s is called *ordinal* if the order is preserved

$$x > y \iff s(x) > s(y) \text{ for all } x, y \in S$$

- ▶ a transformation f is called *permissible* for an ordinal scale if

$$s(x) > s(y) \iff f(s(x)) > f(s(y))$$

- ▶ permissible transformations: \log , $\sqrt{\bullet}$, positive affine transformations

Example 3.4

variable: grade		transformations	
$x_i \in S$	$s(x_i)$	$\log(s(x_i))$	fusion
very good	5.0	1.609	3.0
good	4.0	1.386	3.0
satisfactory	3.0	1.099	2.0
sufficient	2.0	0.693	2.0
not sufficient	1.0	0.000	1.0

- ‘Grade’ is scaled ordinally: $x_i < x_j \Leftrightarrow s(x_i) < s(x_j)$
- log is a permissible transformation
- fusion is a non-permissible transformation

- *interval* scale preserves relative distances

$$s(x) - s(y) = c(f(s(x)) - f(s(y)))$$

- ▶ a positive affine transformation is an interval scale
- ▶ non permissible transformations are \log , $\sqrt{\bullet}$

- *ratio* scale preserves relative ratios

$$\frac{s(x)}{s(y)} = c \frac{f(s(x))}{f(s(y))}$$

- ▶ a permissible operation is the multiplication with a constant
- ▶ non permissible transformations are \log , $\sqrt{\bullet}$, general linear transformation

- *nominal* scale requires unique identifiers
 - ▶ almost all transformations are permissible
 - ▶ non permissible transformations are fusion of identifiers

Scale	Permissible Statistics	Operations
nominal	absolute frequency, mode, contingency coefficients	Equality: $=, \neq$
ordinal	median, percentiles	Greater/less: $<, \leq, >, \geq$
interval	mean, standard deviation, rank-order coefficient, product-moment correlation	Equality of differences: $ax + b$
ratio	coefficient of variation	Equality of ratios: ax

- Problems
 - ▶ Taxonomy is incomplete:
Which scale has “wind direction” ?
 - ▶ Data analyst may be able to interpret the result of “non-permissible” method:
Why is the scale of “temperature” in °K or °C different?
- Scales are too strict
 - ▶ Leads too often to degrading data to rank-based methods
- Statistical theory often relies on
 - ▶ distributional assumptions or
 - ▶ assumptions about distribution parameters/properties

Stevens, S. S. (June 1946). "On the Theory of Scales of Measurement". en. In: *Science* 103.2684, pp. 677–680. ISSN: 0036-8075, 1095-9203. DOI: [10.1126/science.103.2684.677](https://doi.org/10.1126/science.103.2684.677). URL: <http://www.sciencemag.org/cgi/doi/10.1126/science.103.2684.677> (visited on 08/08/2015).

Example 3.5

At a reception consecutively numbered tickets, starting with “1” were given to the people entered for a lottery. As a winning number, 126, was selected and announced, one participant compared it to her ticket to see if she had won, thus interpreting the “126” correctly as a nominal value. She then looked around the room and remarked that, “It doesn’t look like there are 126 people here”, now interpreting the same value, again correctly as a ratio-scale value. One of the authors compared his ticket number (56) to the winning value and realized that he had arrived too soon to win the prize, thus interpreting the values ordinally. If additional data about the rate and regularity of arrivals had been available, he might have tried to estimate by how much longer he should have delayed his arrival from the 70-ticket difference between his ticket and the winner, thus treating the ticket number as an interval-scale value.

Velleman, Paul F. and Wilkinson, Leland (Feb. 1993). “Nominal, Ordinal, Interval, and Ratio Typologies are Misleading”. en. In: *The American Statistician* 47.1, pp. 65–72. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.1993.10475938. URL: <http://www.tandfonline.com/doi/abs/10.1080/00031305.1993.10475938> (visited on 08/08/2015).

- nominal

- `R factor(x, levels, labels=levels, exclude=NA)`
 - `R as.factor(x)`
 - `R is.factor(x)`

- ordinal

- `R ordered(x, levels, labels=levels, exclude=NA)`
 - `R as.ordered(x)`
 - `R is.ordered(x)`

- metric

- `R as.numeric(x)`
 - `R is.numeric(x)`



Listing 3.1: example_scales.R

```
1 x <- 1:5
2 x
3 f <- factor(x)
4 f
5 g <- factor(x, levels=1:6)
6 g
7 o <- ordered(x)
8 o
```

Data quality

- Relevance: How relevant is the data? How fine is the data granulated?
- Accuracy & Reliability: How accurately is data measured? How precise are estimates?
- Missingness: Proportion of missing values per variable or per observation
- Completeness & coverage: How well is the population is covered? Data sampling process?
- Error of sampling: Standard errors & confidence intervals
- Validity: Does data have valid values, e.g. can a be gestation last for 99 weeks?
- Consistency: Does data fit together, e.g. does “student age” and “number of terms” fit?

- Uniformity: Do all observations and variables use the same “unit of measure”?
- Coherence & comparability: Do we expect changes in the future? Can we compare data from different locations? From different times?
- Timeliness & punctuality: How old is the data?
- Accessibility & clarity: Is the data accessible? Is the data well described?
- **Eurostat** requires
 - ▶ relevance, accuracy & reliability, timeliness & punctuality, coherence & comparability, accessibility & clarity
 - ▶ Document: Quality Assurance Framework of the European Statistical System
 - ▶ Document: European Statistics: Code of practice

Classification of Data Accuracy & Reliability

1. Numbers that we can believe in
 - ▶ data collected by legal obligation (official statistics)
 - ▶ data measured automatically
2. Numbers that are reasonably accurate
 - ▶ data from well conducted (and described) surveys
3. Numbers that could be out by quite a long way
 - ▶ data from “representative” surveys
4. Numbers that are unreliable
 - ▶ biased surveys
 - ▶ undocumented surveys
5. Numbers that have just made been up

Spiegelhalter, David (Aug. 2015). "Sex-rated statistics". In: *Significance* 12.4, pp. 21–25. ISSN: 17409705. DOI: 10.1111/j.1740-9713.2015.00840.x. URL: <http://doi.wiley.com/10.1111/j.1740-9713.2015.00840.x> (visited on 10/05/2015).

Example 3.6

1. Numbers that we can believe in
 - ▶ number of formally illegitimate births
2. Numbers that are reasonably accurate
 - ▶ age which at which a (average) women/men had sex for the first time
3. Numbers that could be out by quite a long way
 - ▶ percentage of husbands which had extramarital sex (50%)
4. Numbers that are unreliable
 - ▶ percentage of women having affairs in the first five years after marriage (70%)
5. Numbers that have just made been up
 - ▶ number of trafficked “sex slaves” in UK (25.000)

Political influence

Example 3.7 (Graciela Bevacqua, Argentina)

- 1984 she joined INDEC (Instituto National de Estadistica y Censos)
- involved in the calculation of the consumer price index (CPI)
- 2002 she became Associated Director of CPI group at INDEC
- 2005 the CPI was estimated 12,3% with upward trend
- Minister of Economics starts to challenge the figure
- challenging of data, methodology and results

The demands from Moreno (Secretary of Domestic Trade) were daily. Phone calls from Moreno were 40-minute-long shouting demands, ...

- ▶ round e.g. 2,599 always to 2,5
- ▶ change basket e.g. to “cheap clothes”, capped prices
- ▶ pressure on shops to tell wrong prices

- 2007 she was suspended after vacations
- 2009 she quitted
- prior to 2006, Argentinas methodology was exemplary in Latin America.
- since 2012 "The Economist" refuses to take INDEC CPI figures
- Bevacqua and other companies tried to provide independent estimates of CPI
- in 2011 seven companies were fined with \approx 100.000 EUR for not using "appropriate methodological requirements"
- in 2012 a charge was filed against Bevacqua on "false technical information for distort the market" (2-6y prison)
- charges has been dropped by court

Carriquiry, Alicia (Dec. 2012). "Graciela Bevacqua". In: *Significance* 9.6, pp. 34–36. ISSN: 17409705. DOI: 10.1111/j.1740-9713.2012.00621.x. URL: <http://doi.wiley.com/10.1111/j.1740-9713.2012.00621.x> (visited on 08/16/2015).

Die Deutsche Statistische Gesellschaft fordert Unabhängigkeit der amtlichen Statistik

Leitungen statistischer Behörden sind ausschließlich nach fachlichen, nicht nach politischen Gesichtspunkten zu ernennen!

Diesen Appell richtet die Deutsche Statistische Gesellschaft angesichts der seit geraumer Zeit vakanten Leitungsposten der Statistischen Landesämter Berlin-Brandenburg und Sachsen an die zuständigen politischen Instanzen.

Die Deutsche Statistische Gesellschaft befürchtet beträchtliche Defizite, falls es zu Ernennungen von Führungskräften nicht nach den Kriterien Eignung, Leistung und Befähigung, sondern nach politischen Motiven käme.

Das Bundesstatistikgesetz schreibt die Grundsätze der Neutralität, Objektivität und wissenschaftlichen Unabhängigkeit vor. Auch die Europäische Union fordert in ihrer Statistik-Verordnung Unabhängigkeit, Unparteilichkeit und Objektivität.

Die amtliche Statistik stellt der Politik zwar die für deren Arbeit nötigen Daten bereit, muss aber bei der Erhebung, Aufbereitung und Auswertung der Zahlen unabhängig, unparteiisch und objektiv sein, um weiterhin glaubwürdig zu bleiben.

Quelle: [Pressemitteilung der Dt. Statistischen Gesellschaft vom 13.08.2014](#)

Europäische Prüfung - Zweifel an Unabhängigkeit deutscher Statistikämter

Sind Statistiker vollkommen unabhängig? In Deutschland sind sie das nicht ganz - das ist zumindest der Schluss eines europäischen Kontrollgremiums.

Quelle: [faz.net vom 12.10.2014](#)

Example 3.8

- 2009 raise of budget deficit from 6% to 12.7% of GDP
- 2010 Eurostat had "reservations" about Greek budget deficit figure provided by the National Statistical Service of Greece (ESYE) (13.8% of GDP)

"lack of quality of the Greek fiscal statistics"

"substantial number of unanswered questions and pending issues still remain in some key areas"

"problems related to statistical weaknesses and failures of the relevant Greek institutions in a broad sense"

"unsatisfactory technical procedures in the Greek statistical institute"

"inappropriate governance, with poor cooperation and lack of clear responsibilities ... which leave the quality of fiscal statistics subject to political pressures and electoral cycle"

- 2010

- ▶ ESYE, a part of Ministry of National Economy, was dissolved and replaced by ELSTAT as advised by Eurostat
- ▶ Andreas Georgiou, a greek economist, appointed as first President of the Hellenic Statistical Authority (ELSTAT), Octobre the deficit was estimated to 15.4%

"Hacker had been entering multiple times a day into my account from day one of my work at ELSTAT and had accessed and downloaded thousands of my e-mails" (hacker=vice-president of ELSTAT).

- 2011

- ▶ investigation for artificially inflating the 2009 government deficit figure
- ▶ Board members wanted more direct involvement in producing numbers
- ▶ finance minister dismissed the ELSTAT board

- 2013

- ▶ he was charged with falsifying official data: Lies, damned lies and Greek statistics - Financial Times
- ▶ judge dropped charges

- 2014 Prosecutor reopened the case, again dropped
- 2015
 - ▶ new Prosecutor reopened the case, again dropped
 - ▶ Andreas Georgiou resigned
 - ▶ new charge for violation of duty (not reporting fiscal number to the board first)
- 2016
 - ▶ acceptance of annulment proposal of 2013 charge
 - ▶ convicted for defaming the former Director of National Accounts (the court did not find that what he said was untrue but that he should not have said it): one- year prison sentence, suspended for two years
- 2017
 - ▶ appeal against the defamation conviction was rejected
 - ▶ found guilty in civil lawsuit by the former Director of National Accounts
 - ▶ the 2013 charge again dropped and annulled again
 - ▶ court imposed a two-year suspended prison sentence for second charge

Data cleansing

Garbage in → garbage out

- formatting data
- descriptive statistics
- transforming data, e.g. standardization
- deletion or replacement of implausible observation values
- duplicate deletion
- outlier detection
- deletion or replacement of missing values

Tableplot

- Assumptions:
 - ▶ variables in a dataset are mutually dependent
 - ▶ unusual values in one variable may result in unusual values in another variable
- Plot all variables in parallel
- Sort observations by one variable

Gribov, Alexander, Unwin, Antony, and Hofmann, Heike (2006). "About Glyphs and Small Multiples: Gauguin and the Expo". In: *ASA Sections on Statistical Computing and Statistical Graphics* 17.1.

Malik, Waqas Ahmed, Unwin, Antony, and Gribov, Alexander (2010). "An Interactive Graphical System for Visualizing Data Quality—Tableplot Graphics". In: *Classification as a Tool for Research*. Ed. by Hermann Locarek-Junge and Claus Weihs. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 331–339. ISBN: 978-3-642-10744-3 978-3-642-10745-0. URL: http://link.springer.com/10.1007/978-3-642-10745-0_36 (visited on 08/08/2015).

⌚ Listing 3.2: example_tableplot.R

```
1 library("tabplot")
2 library("MASS")
3 Boston$chas <- factor(Boston$chas)
4 Boston$rad <- ordered(Boston$rad)
5 tableplot(Boston)
```

⌚ `tabplot::tableplot(dat, select, subset=NULL, sortCol=1,
decreasing=TRUE)`

Sampling

- Population: well defined set (in space, time, etc.) of (finite) analysis objects
- Sample: a subset of the population
- Sample (random) variable: a random variable for each analysis object in the sample describing a property
- Estimator: a function build from sample random variables, usually used for estimation:
 - ▶ parameters as mean, variance, proportion incl. distribution
 - ▶ confidence interval
 - ▶ test statistics

The Literary Digest Desaster

- Presidential election in 1936 (Roosevelt vs. Landon)
- *The Literary Digest* magazine started a large questionnaire to predict the result: 2,3 Mio. respondents
- Election result
- *The Literary Digest* predicted a opposite result
- The Gallup institute predict the poll correctly with 50 thousand respondents
- *The Literary Digest* was discontinued soon

What can we learn from the Literary Digest Desaster?

- Several biases (which can not be corrected!)
 - ▶ Telephone, subscription and car register were biased (selection bias)
 - ▶ Motivation to answer is biased (self selection bias)
- A large number of respondents (2,3 Mio.) does not protect against a bias!

Random sampling

- Simple random sampling:
 - ▶ each object has prob. > 0 to get in the sample
 - ▶ all objects have the same prob. to get in the sample
 - ▶ drawing of an object is *independent* of former draws
 - ▶ Example: Sampling *with replacement*
- Random sampling:
 - ▶ each object has prob. > 0 to get in the sample
 - ▶ all objects have the same prob. to get in the sample
 - ▶ drawing of an object is *dependent* of former draws
 - ▶ Example: Sampling *without replacement*
- Why random sampling is important?
 - ▶ X_i : random sample variable: prob. for values of the i th observation
 - ▶ X is *abc* distributed in the population $\Rightarrow X_i$ is also *abc* distributed with the same parameters (!)
 - ▶ statistical inference is done with the (empirical) distribution of X_i

1. if (simple) random sampling has been done then statistical inference (tests, confidence intervals, etc.) give reliable results about the population
2. if **NO** random sampling has been done then statistical inference (tests, confidence intervals, etc.) is unreliable ⇒ **NO tests, confidence intervals, etc.**
3. if the data are the population (census) then statistical inference is unnecessary ⇒ **NO tests, confidence intervals, etc.**

 Listing 3.3: example_sample.R

```
1 data(Boston, package="MASS")
2 mean(Boston$medv) # true mean in the population
3 #
4 library("DescTools")
5 # random sample of size 100
6 ind <- sample(1:nrow(Boston), size=100)
7 MeanCI(Boston$medv[ind])
8 # biased sample of size 100
9 ind <- sample(1:nrow(Boston), size=100, prob=Boston$rm>6.5)
10 MeanCI(Boston$medv[ind])
11 #
12 # There is a population for each sample,
13 # but most likely it is not the population of interest!
```

- ☞ `sample(x, size, replace=FALSE, prob=NULL)`
- ☞ `sampling::srswr(n, N)`

Estimator properties

$\hat{\pi} = X/n$ with $X = \sum X_i$ and $X_i \sim \text{Bernoulli}(\pi)$

- Sampling *with* replacement ($X \sim B(n; \pi)$)

$$\text{Var}_{\text{with}}(\hat{\pi}) = \frac{1}{n}\pi(1 - \pi)$$

- Sampling *without* replacement ($X \sim \text{Hyp}(N; M = N\pi; n)$)

$$\text{Var}_{\text{without}}(\hat{\pi}) = \underbrace{\frac{N-n}{N-1}}_{\text{finity correction}} \frac{1}{n}\pi(1 - \pi)$$

- Obviously holds:

$$\text{Var}_{\text{without}}(\hat{\pi}) \leq \text{Var}_{\text{with}}(\hat{\pi})$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, E(X_i) = \mu, \text{Var}(X_i) = \sigma_X^2$$

- Sampling *with* replacement

$$\text{Var}_{\text{with}}(\bar{X}) = \frac{1}{n} \underbrace{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}_{\text{unbiased estimator for } \sigma_X^2} = \sigma_X^2/n$$

- Sampling *without* replacement

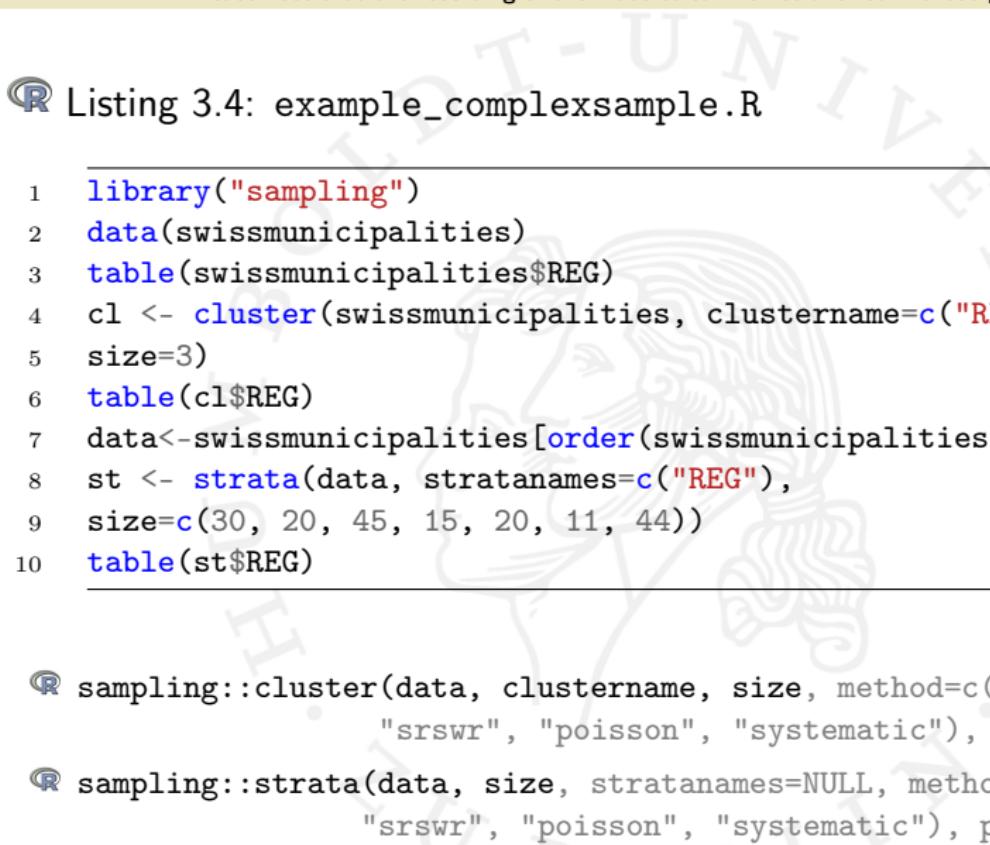
$$\text{Var}_{\text{without}}(\bar{X}) = \frac{1}{n} \underbrace{\frac{N-n}{N-1}}_{\text{finity correction}} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Obviously holds:

$$\text{Var}_{\text{without}}(\bar{X}) \leq \text{Var}_{\text{with}}(\bar{X})$$

Other sampling methods

- Stratified sampling
 - ▶ Decompose the population into L strata, e.g. male/female
 - ▶ In each strata perform (simple) random sampling
- Clustered sampling
 - ▶ Decompose the population into L cluster, e.g. counties of a city
 - ▶ For some clusters perform full sampling
- Judgement sampling methods
 - ▶ Select *typical* cases
 - ▶ Select *large* contributors → official statistics
 - ▶ Select according to quotas → most often used in marketing, opinion research

R Listing 3.4: example_complexsample.R

```
1 library("sampling")
2 data(swissmunicipalities)
3 table(swissmunicipalities$REG)
4 cl <- cluster(swissmunicipalities, clustername=c("REG"),
5 size=3)
6 table(cl$REG)
7 data<-swissmunicipalities[order(swissmunicipalities$REG),]
8 st <- strata(data, stratanames=c("REG"),
9 size=c(30, 20, 45, 15, 20, 11, 44))
10 table(st$REG)
```

R sampling::cluster(data, clustername, size, method=c("srswor",
 "srswr", "poisson", "systematic"), pik)
R sampling::strata(data, size, stratanames=NULL, method=c("srswor",
 "srswr", "poisson", "systematic"), pik)

Estimator properties

- In stratified independent sampling without replacement in L strata it holds

$$\text{Var}(\bar{X}) = \frac{1}{N^2} \sum_{h=1}^L N_h^2 \underbrace{\frac{N_h - n_h}{N_h} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_{ih} - \bar{x}_h)^2}_{\text{as in sampling without replacement}}$$

- It holds (stratification effect)

$$\text{Var}_{\text{stratified}}(\cdot) \leq \text{Var}_{\text{simple}}(\cdot) \leq \text{Var}_{\text{clustered}}(\cdot)$$

Representative vs. random samples

- Random sample
 - ▶ does *not* use object properties for sampling
 - ▶ a random sample is not a representative sample
- Representative sample
 - ▶ does use object properties for sampling
 - ▶ Problem: how do we know that is representative?
 - ▶ usually only for a subset of variables guaranteed
 - ▶ a representative sample is not a random sample

	Knowlegde about population	Statistics applicable?	
		Exploratory	Inferential
Simple Random	Low	Yes	Yes
Random	Low	Yes	Yes
Stratified	Medium	Yes	Yes
Cluster	Medium	Yes	Maybe
Judgement	High	Yes	No
Representative	Very High	Yes	No

Test and estimation theory

November 3, 2022

- Steps in testing • Test types • p -value • Our own test • Bootstrap •
- Bootstrap for confidence intervals • Tests and confidence intervals • Monte Carlo test • Exact tests • Problems of tests • Effect sizes vs. tests •
- Specific effect sizes • From the sample to the population • Estimator properties • Least squares estimation • Maximum-Likelihood estimation • Likelihood ratio test • Score test/Lagrange multiplier test • Wald Test •
- Holy Trinity • Kullback-Leibler divergence and entropy • Akaike Information Criterion • Bayesian Information Criterion • Information criteria • Occam's razor

Steps in testing

1. Check assumptions
2. Formulate hypotheses (H_0, H_1)
 - ▶ important hypothesis is H_1 !
 - ★ for one-sided tests exist two possibilities:
 - ★ hypothesis generation: what you want to prove comes in H_1
 - ★ hypothesis generation: the greater risk has to be controlled therefore becomes type I error (" $H_1 | H_0$ ")
3. Determine test statistics under H_0
4. Draw (simple) random sample
5. Determine critical value(s) and test value
6. Accept or reject H_1

Note: If not said differently then we will assume that the sample random variables X_{ji} are independent und identical distributed (iid)

	Result of testing	
	" H_0 "	" H_1 "
H_0 true	correct decision $P("H_0" H_0)$ (true negative)	wrong decision type I error $P("H_1" H_0)$ (false positive)
H_1 true	wrong decision type II error $P("H_0" H_1)$ (false negative)	correct decision test power $P("H_1" H_1)$ (true positive)

- a good test has a large test power $P("H_1" | H_1)$

Test types

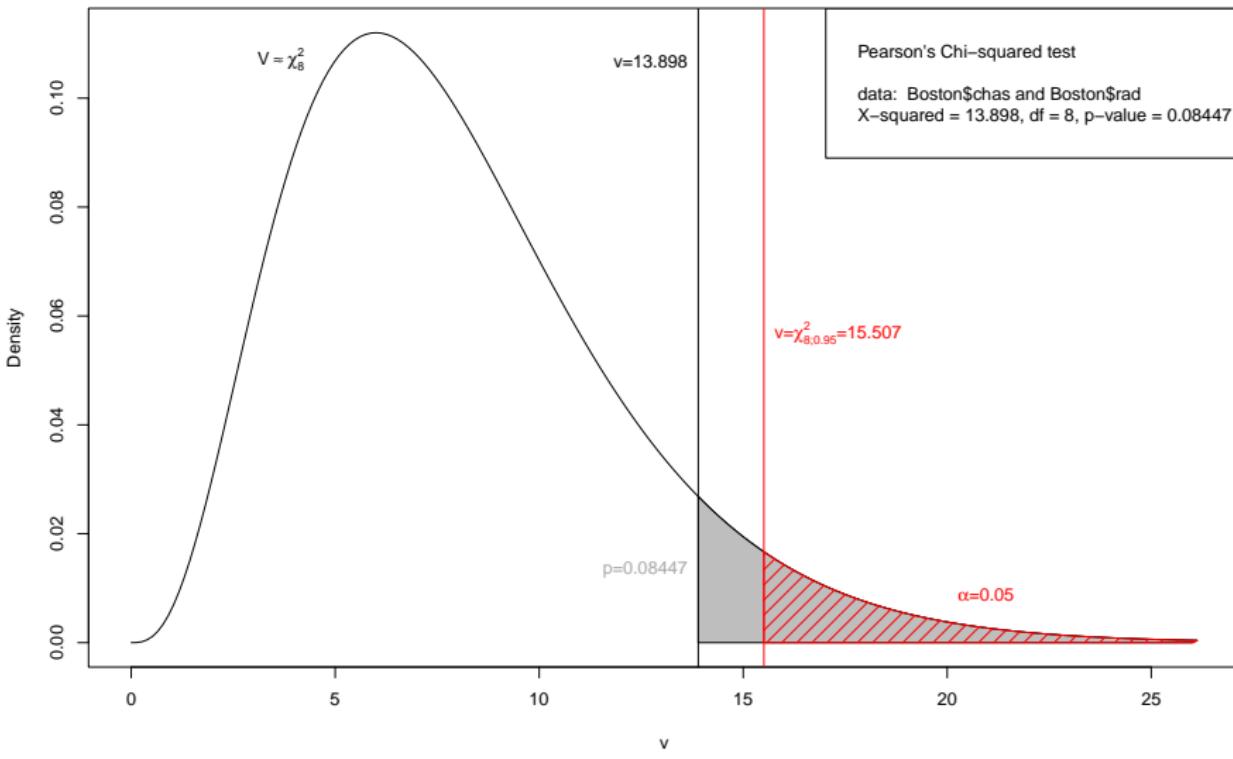
- Parametric test: distribution of population fixed, usually normal
- Nonparametric test: distribution of population free, usually based on ranks

Examples	Parametric	Nonparametric
Tests on mean (median)	Gauss test, t-test, ANOVA	Mann-Whitney U test, Kruskal-Wallis test
Tests on variance	F-test	Levene test
Test on distributions - arbitrary: - normal:	...	χ^2 , Kolmogorov-Smirnov Jarque-Bera
	...	

- Conservative test: Type I error $P("H_1" | H_0) < \alpha$
 - ▶ generally it holds $P("H_1" | H_0) \leq \alpha$ (significance level)
- Exact test: discrete test statistics
 - ▶ Example: Binomial test without normal approximation
- Permutation test: Distribution of the discrete test statistics is computed for each sample size
- Monte Carlo test: Distribution of the test statistics is simulated from the specific data set
 - ▶ those test statistics do not follow a known distribution
 - ▶ we usually use the reproducibility properties or CLT to derive the distribution of the test statistics

p-value

- Let $V = V(X_1, \dots, X_n)$ a test statistics and $v = V(x_1, \dots, x_n)$ the realised test statistics
- The *p*-value is defined as
 - ▶ for a two-tailed test: $p = P(V < v) + P(V > v) = P(|V| > v)$
 - ▶ for a left-tailed test: $p = P(V < v)$
 - ▶ for a right-tailed test: $p = P(V > v)$
- If $p < \alpha$ (significance level) then reject H_0
- If $p \geq \alpha$ then do not reject H_0
- Software does not need to know about the significance level



Our own test

Problem: Is the population symmetric distributed?

Hypothesis: $H_0 : X$ symmetric vs. $H_1 : X$ is not symmetric

Known: If X is symmetric then $\bar{x} = x_{0.5}$

$$n_- = \#\{x_i < \bar{x}\}, n_+ = \#\{x_i \geq \bar{x}\}$$

Expected frequency for each class: $0.5n$

$$\frac{(n_- - 0.5*n)^2}{0.5n} + \frac{(n_+ - 0.5*n)^2}{0.5n} \approx \chi_1^2$$

Note: Rejection of H_1 does not imply symmetry!

Do not use this test: the power is very low,
better to use the ► sign test

Bootstrap

- Consider the sample as the “best” approximation to the population
- Do inference by drawing with replacement (for independence) from the sample rather than the population
- Problem: computer-intensive
- Size of bootstrap sample is n (may vary)
- Number of bootstrap samples B
 - ▶ Efron & Tibshirani: $50 \leq B \leq 100$
 - ▶ Wikipedia (2007): $1000 \leq B \leq 10000$
 - ▶ Crude rule of thumb: $B = 200$ for SE, $B = 2000$ for CI
(see Davidson & MacKinnon, 2000)

Efron, Bradley and Tibshirani, Robert (1993). *An introduction to the bootstrap*. Monographs on statistics and applied probability 57. New York: Chapman & Hall. 436 pp. ISBN: 978-0-412-04231-7.

Davidson, Russell and MacKinnon, James G. (Jan. 2000). “Bootstrap tests: how many bootstraps?” In: *Econometric Reviews* 19.1, pp. 55–68. ISSN: 0747-4938, 1532-4168. DOI: 10.1080/07474930008800459. URL:
<http://www.tandfonline.com/doi/abs/10.1080/07474930008800459> (visited on 12/02/2015).

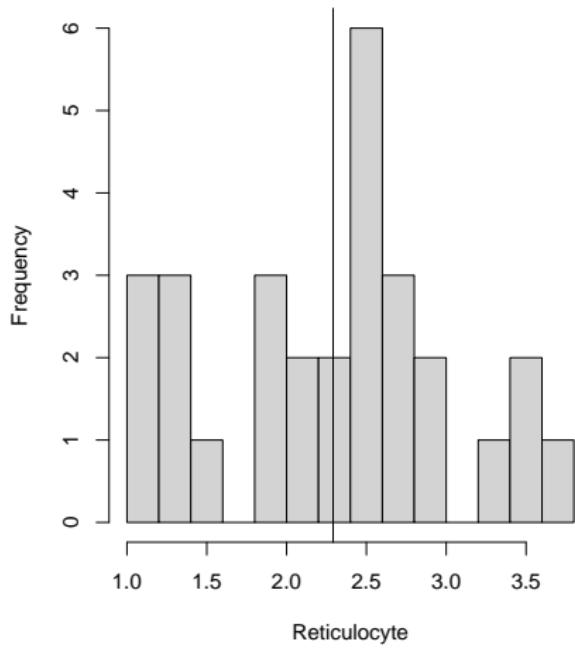
- classical (asymptotical) statistics
 - ▶ start with random sample variables X_1, \dots, X_n
 - ▶ reduce your statistical problem to a (test) statistic $U = u(X_1, \dots, X_n)$
 - ▶ derive the distribution of U , e.g. by
 - ★ making assumptions about the distribution of X_i ($i = 1, \dots, b$)
 - ★ applying CLT or rules from distribution calculus
 - ▶ from the distribution of U derive acceptance region for H_0 or confidence interval
- bootstrapped statistics
 - ▶ generate bootstrap samples with random sample variables X_1^*, \dots, X_n^*
 - ★ the distribution of X_i^* is *only* “near” to the (true) distribution of X_i
 - ▶ the distribution of $U^* = u(X_1^*, \dots, X_n^*)$ is near to the distribution of U
 - ▶ to simulate the distribution of U^* generate B bootstrap samples
 - ▶ pitfall: bootstrap “works” only if the distribution of U and U^* is same (at least asymptotically for $n \rightarrow \infty$)

R Listing 4.1: example_boot.R

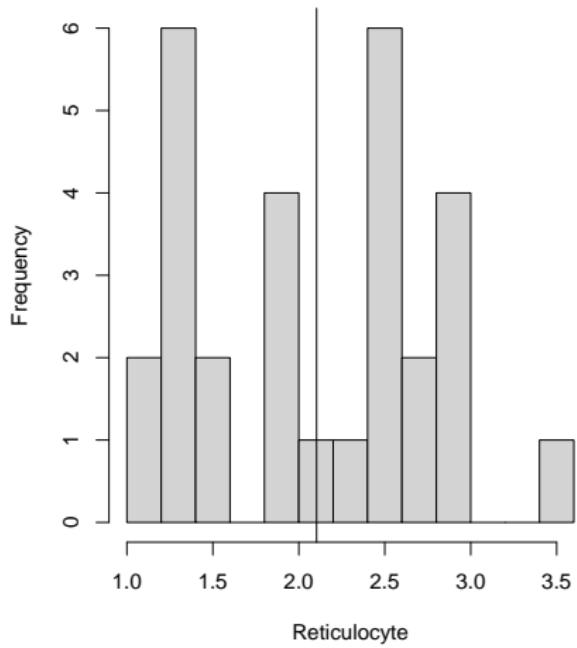
```
1 library("boot")
2 meanboot <- function (x, ind) { return(mean(x[ind])); }
3 #
4 set.seed(24961970)
5 data("pechstein", package="mmstat4")
6 boot(pechstein$Retikulozyten, meanboot, 999)
```

R `boot::boot(data, statistic, B, sim=c("ordinary", "parametric",
 "balanced", "permutation", "antithetic"),
 parallel=c("no", "multicore", "snow"))`

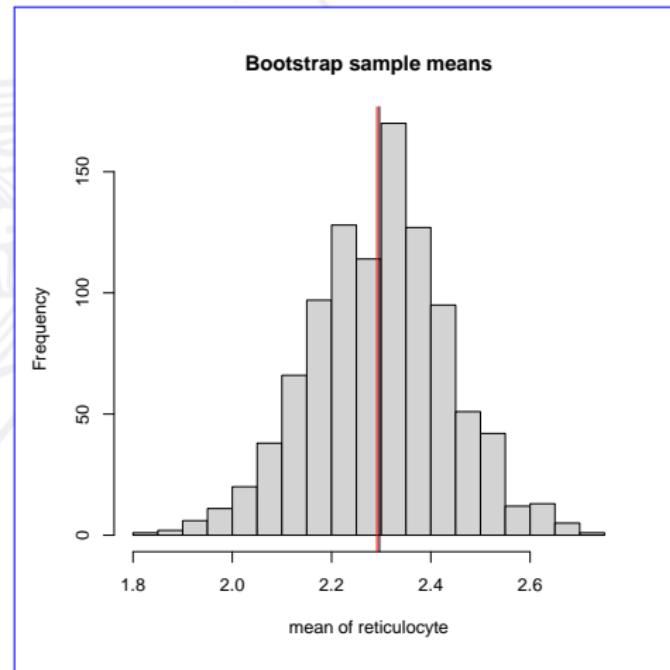
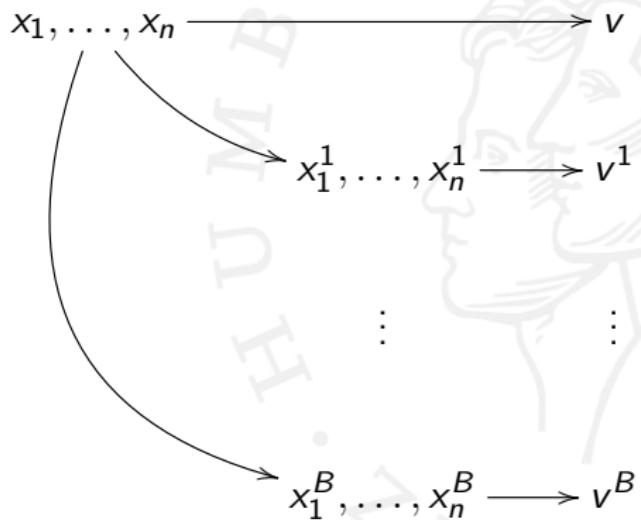
Original sample



Bootstrap sample



Bootstrap for confidence intervals



95% confidence intervall for a mean

- Asymptotical: $P\left(\bar{X} - c \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + c \frac{s}{\sqrt{n}}\right) = 1 - \alpha$

$$n = 29, \bar{x} = 2.29, s = 0.76, c = t_{1-\alpha/2; n-1}$$

$$[2.001358, 2.581401]$$

- Chebyshev's inequality: $P(|X - E(X)| \geq k \cdot Var(X)) \leq \frac{1}{k^2} = \alpha$

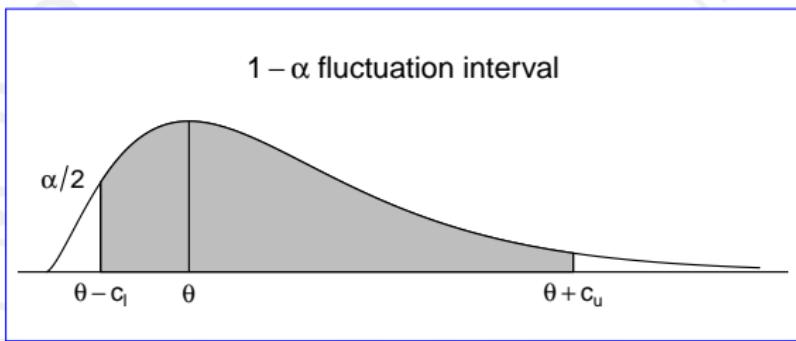
$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}, \quad 1 - \alpha = 95\% \Leftrightarrow k = 4.472136$$

$$[1.658197, 2.924562]$$

- Bootstrap (asymmetric)

$$[2.005172, 2.578966]$$

Tests and confidence intervals



- $1 - \alpha$ fluctuation interval: $P(\theta - c_l \leq \hat{\theta} \leq \theta + c_u) \geq 1 - \alpha$
- $1 - \alpha$ confidence interval: $P(\hat{\theta} - c_l \leq \theta \leq \hat{\theta} + c_u) \geq 1 - \alpha$
- Two-sided test under H_0 : $P(\theta_0 - c_l \leq \hat{\theta} \leq \theta_0 + c_u) \geq 1 - \alpha$
- One-sided tests: union of (one-sided) intervals for all $\theta \in H_0$

Example 4.9

- Average netincome in Germany based on ALLBUS 2010 data:
ca. 1379 EUR with $n = 2000$
- 95% confidence interval: $\underbrace{[1333; 1424]}_{=1379 \pm 44.459}$
- Could be the average netincome 2000 EUR in the population?
 - ▶ Test: $H_0 : \mu = 2000$ vs. $H_1 : \mu \neq 2000$, $\alpha = 5\%$
 - ▶ Test statistics: $T \sim t_{1999} \approx N(0; 1)$, $t = -27.364$
 - ▶ Reject H_0 if $\bar{x} \notin 2000 \pm \underbrace{1.96 * 22,683}_{=44.459} = [1955; 2045]$

Monte Carlo test

- Monte Carlo tests are based on bootstrap idea
- Problem: we need to know the distribution of a test statistics V if H_0 holds
- Bootstrap samples can not be used since we do not know if H_0 in the data holds
- “Bootstrap” samples must be linked to
 - ▶ the data *and*
 - ▶ the null hypothesis
- Note: Bootstrapping in SPSS relates to the stability of a result

Problem: Are two categorical variables X and Y independent?

Hypothesis: $H_0 : X, Y$ independent vs. $H_1 : X; Y$ dependent

Test-statistics: $V = \sum_{i,j} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} \approx \chi^2_{(I-1)(J-1)}$

Problem: It must hold $e_{ij} > 5$

Solution: Bootstrap the test statistics under H_0

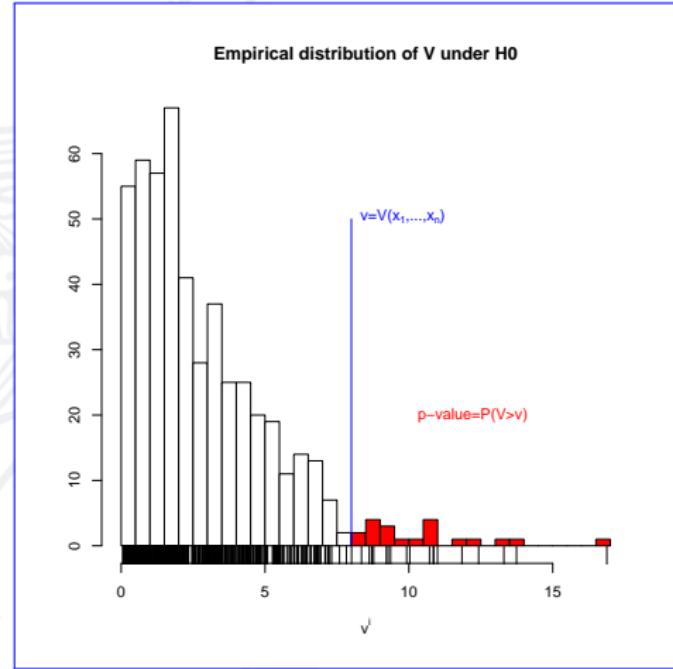
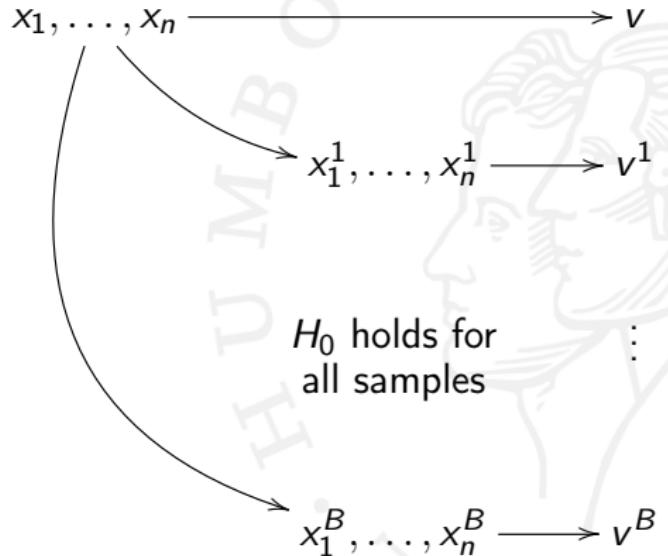
- draw independently from the marginal distributions
- create a bootstrap sample of size n
- compute then V
- use the distribution of V rather than χ^2

$X \setminus Y$	y_1	y_2	\dots	y_K	
x_1	h_{11}	h_{11}	\dots	h_{1K}	$h_{1\bullet}$
x_2	h_{21}	h_{22}	\dots	h_{2K}	$h_{2\bullet}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_J	h_{J1}	h_{J2}	\dots	h_{JK}	$h_{J\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$	\dots	$h_{\bullet K}$	n

- Common distribution of X and Y : (x_i, y_j) with frequency h_{ij}
- Marginal distribution of X : x_i with frequency $h_{i\bullet} = \sum_j h_{ij}$
- Marginal distribution of Y : y_j with frequency $h_{\bullet j} = \sum_i h_{ij}$
- X and Y are independent if for all i and j holds

$$h_{ij} = \hat{e}_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n}$$

- Let (x_i, y_i) a (random) sample ($i = 1, \dots, n$)
- Draw $b = 1, \dots, B$ Bootstrap samples (x_i^b, y_i^b) of size n
 - ▶ Draw x_i^b from x_1, \dots, x_n (or according to the rel. freq. of X)
 - ▶ Draw y_i^b from y_1, \dots, y_n (or according to the rel. freq. of Y)
 - ▶ x_i^b and y_i^b drawn independently according to H_0 !
- Compute the test statistics v^b based on the bootstrap sample (x_i^b, y_i^b) with $i = 1, \dots, n$
- The values v^b simulate the distribution of V^*
- Compute the test statistic v based on (x_i, y_i)
- Compare v and the v^b 's



⌚ Listing 4.2: example_chisq.R

```
1 library("MASS")
2 ctab <- table(Boston$chas, Boston$rad)
3 ctab
4 chisq.test(ctab)
5 chisq.test(ctab, simulate.p.value=T)
```

⌚ chisq.test(x, simulate.p.value=T, B=2000)

Exact tests

Example 4.10

- Assume the following problem: Are X and Y independent?

Y/X	observed		expected	
	0	1	2.5	2.5
0	3	2	5	2.5
1	2	3	5	2.5
	5	5	10	

- List of all 2×2 tables which have the same marginal distributions

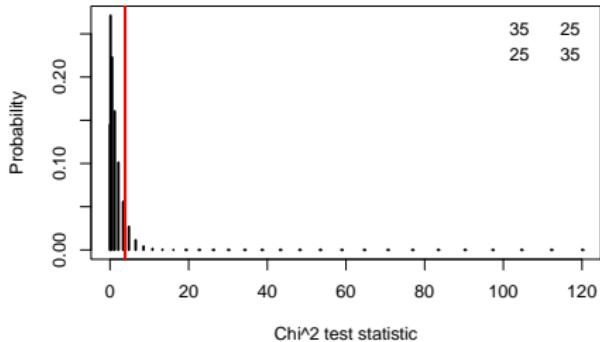
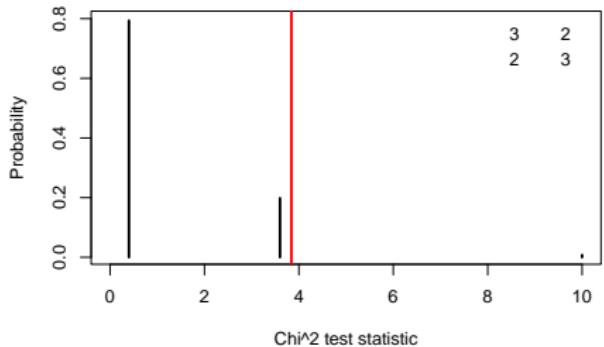
Tables	0 5	1 4	2 3	3 2	4 1	5 0
	5 0	4 1	3 2	2 3	1 4	0 5
χ^2	10.0	3.6	0.4	0.4	3.6	10.0
P(table)	0.4%	9.9%	39.7%	39.7%	9.9%	0.4%

- Generate all possible tables with the same marginals

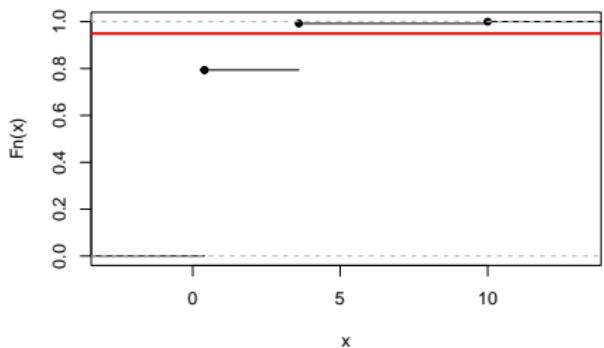
a	b		a+b
c	d		c+d
$a+c$	$b+d$	$n=a+b+c+d$	

$$P(\text{table}) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

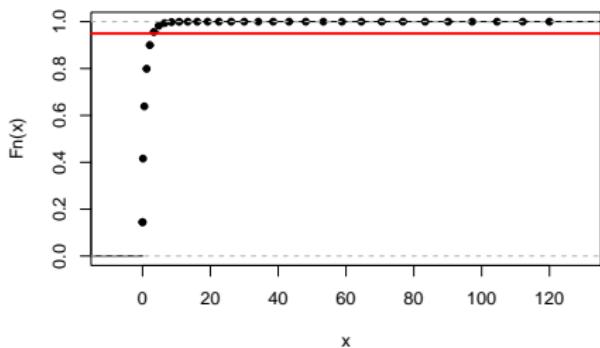
- Compute the χ^2 test statistics for all tables
- Test decision
 - ▶ if the test statistics of the observed table is larger than the $1 - \alpha$ quantile of the test statistics all possible tables then reject the null hypothesis
- Problem: the number of all possible tables may grow exponentially
 - ▶ Monte Carlo test is a “special case” of an exact test: generate a random subset of all possible tables



Cumulative distribution function



Cumulative distribution function



Problems of tests

- the test hypotheses relate to the *population* and not to the *sample*
- test application requires that the sample is *randomly* drawn from the population
 - ▶ in case of a total population survey a test is useless
 - ▶ in case of non-random drawing a test is useless
- tests are sensitive to the sample size
 - ▶ in case of small sample sizes even large “effects” may not lead to a rejection of the null hypothesis
 - ▶ in case of large sample sizes even small “effects” lead to a rejection of the null hypothesis with high significance
- we reduce a complex problem into a 0/1 decision
 - ▶ loss of information?
- lack of adequate statistical power ($= 1 - \text{type II error}$)
 - ▶ when will be the null rejected?

- what does the result of a test mean?
 - ▶ if we reject H_0 then H_0 can not be true (up to α)
 - ▶ but if we can not reject H_0 then what?
 - ▶ we do not know if for any parameter/distribution in the H_1 the distribution of the test statistics is near to the distribution under H_0
 - ▶ in practice: act as if the null hypothesis is true
- the null hypothesis is nearly always false
 - ▶ why do we want to reject the null hypothesis?
 - ▶ in case of large sample sizes the null will be always rejected \Rightarrow what about the type I error?
- misunderstanding and missinterpretation of a p -value
 - ▶ probability that a result is due to mere chance
 - ▶ a $p < 0.05$ finding is one that is 95% likely to replicate
 - ▶ a $p < 0.05$ there is a 1 in 20 chance of making a type I error

- a $p < 0.05$ provides statistical evidence for substantive hypothesis
- Classical test theory does not answer researchers question

$$P(H|E) = P(E|H) \frac{P(H)}{P(E)}$$

- ▶ H null hypothesis true, E evidence based on our data
- ▶ $P(H|E)$ probability that the hypothesis H is true given the evidence E
- ▶ $P(E|H)$ probability that the evidence E appears given the hypothesis H is true
- to alleviate the problems:
 - ▶ have non-statistical theory/evidence available (!)
 - ▶ choose smaller α 's (increases β)
 - ▶ report effect sizes
 - ▶ report confidence intervals
 - ▶ avoid tests

Effect sizes vs. tests

- Problem of tests
 - ▶ test statistics depends on the sample size
 - ▶ Example: χ^2 independence test

$$\chi^2 = n \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - p_{ij})^2}{p_{ij}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(h_{ij} - e_{ij})^2}{e_{ij}}$$

- if the sample size grows then even small deviation of the null hypothesis become highly significant
- Cohen (1988) defined a set of coefficients, effect sizes, which should fulfill:
 - ▶ dimensionless
 - ▶ does not depend from measurement unit of the data
 - ▶ is independent from the sample size
 - ▶ if the null hypothesis of the associated test can not be rejected then the effect size should be near 0

- Cohen was interested to find sample sizes where “small”, “medium” and “large” effect size are statistical significant based on
 - ▶ significance level α
 - ▶ effect size
 - ▶ power $(1 - \beta)$
- Criticism ([Lenth, 2006](#))

This is an elaborate way to arrive at the same sample size that has been used in past social science studies of large, medium, and small size (respectively). The method uses a standardized effect size as the goal. Think about it: for a “medium” effect size, you’ll choose the same n regardless of the accuracy or reliability of your instrument, or the narrowness or diversity of your subjects. Clearly, important considerations are being ignored here. “Medium” is definitely not the message!

- Criticism (Ellis, 2010)

In an ideal world scholars would normally interpret the practical significance of their research results by grounding them in a meaningful context or by assessing their contribution to knowledge. When this is problematic, Cohen's benchmarks may serve as a last resort.

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates. ISBN: 978-0-8058-0283-2.

Lenth, Russ (2006). *Russ Lenth's power and sample-size page*. URL:
<http://homepage.stat.uiowa.edu/~rlenth/Power/> (visited on 08/08/2015).

Ellis, Paul D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge ; New York: Cambridge University Press.
ISBN: 978-0-521-19423-5 978-0-521-14246-5.

Specific effect sizes

- Association of two continuous variables
 - ▶ (absolute) Bravais-Pearson correlation r
 - ▶ $r = 0, 1$ small effect, $r = 0, 3$ medium effect, $r = 0, 5$ large effect
- Association of two discrete variables
 - ▶ Cohens w
 - ▶ $w = \sqrt{\frac{\chi^2}{n}}$
 - ▶ $w = 0, 1$ small effect, $w = 0, 3$ medium effect, $w = 0, 5$ large effect
 - ▶ Note: w can become larger than 1

- Cohens d compares mean differences with the standard deviation

$$D = \frac{\mu_1 - \mu_2}{\sigma_D}$$

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

- ▶ $d = 0, 2$ small effect, $d = 0, 5$ medium effect, $d = 0, 8$ strong effect
- ▶ Note: d can become larger than 1
- Correlation between a binary group variable and the observation

$$r = \frac{d}{\sqrt{d^2 + \frac{(n_1+n_2)^2}{n_1 n_2}}}$$

- ▶ $r = 0, 1$ small effect, $r = 0, 3$ medium effect, $r = 0, 5$ large effect

- Regression

$$f^2 = \frac{R_{\text{included}}^2 - R_{\text{excluded}}^2}{1 - R_{\text{included}}^2}$$

$$f^2 = \frac{R^2}{1 - R^2} \quad (\text{all variables excluded})$$

- ▶ $f^2 = 0,02$ small effect, $f^2 = 0,15$ medium effect, $f^2 = 0,35$ strong effect

- ANOVA

$$f = \frac{\sqrt{\frac{1}{k} \sum_{i=1}^k (\bar{x}_i - \bar{x})^2}}{s}$$

- ▶ $f = 0,10$ small effect, $f = 0,25$ medium effect, $f = 0,40$ strong effect

- Modification of existing effect sizes
 - ▶ d is not an unbiased estimator of D (Hedges' g)

$$g = J(n_1 + n_2 + 2)d \text{ with } J(a) = \frac{\Gamma(a/2)}{\Gamma\left(\frac{a-1}{2}\right)\sqrt{a/2}}$$

- ▶ Special situation, e.g. group 2 is control group

$$\delta = \frac{\bar{x}_1 - \bar{x}_2}{s_2}$$

- Other effect sizes
 - ▶ R^2 , η^2 , contingency coefficient (strength of association)
 - ▶ Odds ratio, Relative risk

Hedges, Larry V. (1981). "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators". In: *Journal of Educational Statistics* 6.2, p. 107. ISSN: 03629791. DOI: 10.2307/1164588. URL: <https://www.jstor.org/stable/1164588?origin=crossref> (visited on 05/15/2018).

 Listing 4.3: example_effectsize.R

```
1 library("effsize")
2 library("MASS")
3 # Cohens d
4 cohen.d(survey$Height, survey$Sex, na.rm=TRUE)
5 cohen.d(survey$Wr.Hnd, survey$W.Hnd, na.rm=TRUE)
6 # Hedges g
7 cohen.d(survey$Height, survey$Sex, na.rm=TRUE, hedges.correction=TRUE)
8 cohen.d(survey$Wr.Hnd, survey$W.Hnd, na.rm=TRUE, hedges.correction=TRUE)
```

 effsize::cohen.d(d, f, na.rm=FALSE, hedges.correction=FALSE,
conf.level=0.95)

- Modification of existing effect sizes
 - ▶ d is not an unbiased estimator of D (Hedges' g)

$$g = J(n_1 + n_2 + 2)d \text{ with } J(a) = \frac{\Gamma(a/2)}{\Gamma\left(\frac{a-1}{2}\right)\sqrt{a/2}}$$

- ▶ Special situation, e.g. group 2 is control group

$$\delta = \frac{\bar{x}_1 - \bar{x}_2}{s_2}$$

- Other effect sizes
 - ▶ R^2 , η^2 , contingency coefficient (strength of association)
 - ▶ Odds ratio, Relative risk

Hedges, Larry V. (1981). "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators". In: *Journal of Educational Statistics* 6.2, p. 107. ISSN: 03629791. DOI: 10.2307/1164588. URL: <https://www.jstor.org/stable/1164588?origin=crossref> (visited on 05/15/2018).

From the sample to the population

- A variable X has a distribution in the population
- X_i is a random variable describing the i -th draw from the population
- Each observation x_i is considered as a realisation of X_i
- If (simple) random sampling is used then the distribution of X_i is the same as X
 - ▶ in simple random sampling the X_i 's are independent
- N population size, n sample size, often holds $g = \tilde{g}$

$$\text{population parameter } \vartheta = g(x_1, \dots, x_N)$$

$$\text{sample parameter } v = g(x_1, \dots, x_n)$$

$$\text{estimator } \hat{\theta} = \tilde{g}(X_1, \dots, X_n)$$

$$\text{estimated parameter } \hat{\vartheta} = \tilde{g}(x_1, \dots, x_n)$$

- $\hat{\theta}$ is a random variable

	population parameter	sample parameter
mean	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$ estimator $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ estimated parameter $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$
variance	population parameter $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ estimator $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$	sample parameter $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ estimated parameter $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
$g = \tilde{g}$		
$g \neq \tilde{g}$		

Estimator properties

If $\hat{\theta}_1$ is an estimator for θ then we would like to have

- Unbiasedness:

$$E(\hat{\theta}_1) = \vartheta$$

- Efficiency: For any other estimator $\hat{\theta}_2$ it holds

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

Mean-squared error:

$$MSE(\hat{\theta}_1) = \underbrace{(E(\hat{\theta}_1) - \vartheta)^2}_{\text{Bias }^2} + \underbrace{\text{Var}(\hat{\theta}_1)}_{\text{Variance}}$$

Consistency:

$$\lim_{n \rightarrow \infty} MSE(\hat{\theta}_1) = 0$$

Least squares estimation

$$\sum_{i=1}^n (X_i - E(X_i|\theta))^2 \rightarrow \min.$$

- Minimize the sum of squared error between x_i and expected value of x_i depending on θ

Gauss, Carl Friedrich (1823). "Theoria combinationis observationum erroribus minimis obnoxiae". In: DOI: 10.3931/e-rara-2848. URL:
<http://dx.doi.org/10.3931/e-rara-2848> (visited on 08/16/2015).

Gauss, Carl Friedrich (1887). *Abhandlungen zur Methode der kleinsten Quadrate*. ger. Berlin: Stankiewicz. URL: <http://eudml.org/doc/203261>.

Example 4.11

- Unknown parameter: probability p of loss or profit in a game
- We know that

Profit	-1 EUR	0 EUR	1 EUR
Prob.	p	p	$1 - 2p$

$$E(X_i) = -1 \times p + 0 \times p + 1 \times (1 - 2p) = 1 - 3p$$

- Minimize the “distance” between observed and expected results
- If we observe $-1, 0, 0, 0, 1, 1$ then

$$((-1) - (1 - 3p))^2 + 3 \times (0 - (1 - 3p))^2 + 2 \times (1 - (1 - 3p))^2 \Rightarrow \hat{p} = \frac{5}{18}$$

Maximum-Likelihood estimation

$$L(x_1, \dots, x_n | \theta) \rightarrow \max.$$

- if X_i discrete

$$L(x_1, \dots, x_n | \theta) \stackrel{X_i \text{i.d.}}{=} \prod_{i=1}^n P(X_i = x_i | \theta)$$

- if X_i continuous

$$L(x_1, \dots, x_n | \theta) \stackrel{X_i \text{i.d.}}{=} \prod_{i=1}^n f_i(x_i | \theta)$$

- For technical simplification use the log-likelihood for maximization

$$\log(L(x_1, \dots, x_n | \theta)) \rightarrow \max.$$

$$\prod_{i=1}^n \dots \mapsto \sum_{i=1}^n \dots$$

Example 4.12

- Unknown parameter: probability p of loss or profit in a game
- We know that

Profit	-1 EUR	0 EUR	1 EUR
Prob.	p	p	$1 - 2p$

- If we observe $-1, 0, 0, 0, 1, 1$ then

$$L(p) = p^1 \times p^3 \times (1 - 2p)^2$$

- Log-likelihood

$$\log(L(p)) = 4 \times \log(p) + 2 \times \log(1 - 2p) \Rightarrow \hat{p} = \frac{1}{3}$$

- If assumptions for ML valid then
 - ▶ we get efficient and asymptotically consistent estimates
 - ▶ and it holds
- Moreover ML-theory provides the “Holy Trinity”:
 - ▶ Likelihood ratio test (nested model test)
 - ▶ Wald test (coefficient test)
 - ▶ Lagrange Multiplier (Score) test
- Basically the three test do the same, but use different information
 - ▶ Likelihood ratio test uses two likelihoods
 - ▶ Wald and Score test uses only one likelihood

Likelihood ratio test

Assumption(s): ML estimation used

Hypotheses:
(simple)

$$\begin{aligned} H_0 : \theta &\in \Theta_0 \\ H_1 : \theta &\in \Theta_1 = \overline{\Theta}_0 \end{aligned}$$

Hypotheses:
(composite)

$$\begin{aligned} H_0 : \theta &\in \Theta_0 \subset \Theta_1 \text{ (reduced model)} \\ H_1 : \theta &\in \Theta_1 \text{ (full model)} \end{aligned}$$

Test statistics:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L_0(x_1, \dots, x_n; \theta)}{\sup_{\theta \in \Theta_1} L_1(x_1, \dots, x_n; \theta)}$$

Reject H_0 (simple): $\lambda < k_\alpha$ with $\sup_{\theta \in \Theta_0} P(\Lambda < k_\alpha) = \alpha$
 Note: Generally k_α is difficult to determine

Reject H_0 (composite): $-2 \log(\Lambda) \approx \chi^2_{\dim(\Theta_1) - \dim(\Theta_0)}$

Score test/Lagrange multiplier test

- Score vector (find $\hat{\theta}$ by $S(\theta) = 0$)

$$S(\theta) = \frac{\partial}{\partial \theta} \log(L(\theta)) = \frac{1}{L(\theta)} \frac{\partial}{\partial \theta} L(\theta)$$

- Under certain regularity conditions it holds
 - $E(S(\theta)) = 0$
 - $Var(S(\theta)) = I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} \log(L(\theta))\right)$
($I(\theta)$ Fisher information)
- Score test
 - $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
 - under H_0 holds:

$$\lim_{n \rightarrow \infty} \frac{S(\theta) - E(S(\theta))}{\sqrt{Var(S(\theta))}} = \lim_{n \rightarrow \infty} \frac{S(\theta)}{\sqrt{I(\theta)}} \approx N(0, 1) \Rightarrow \lim_{n \rightarrow \infty} \frac{S^2(\theta)}{I(\theta)} \approx \chi_1^2$$

Wald Test

- Wald test (coefficient test)

- ▶ $\lim_{n \rightarrow \infty} \hat{\theta} \longrightarrow N(\theta, \sigma_{\hat{\theta}}^2)$
- ▶ $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
- ▶ under H_0 holds:

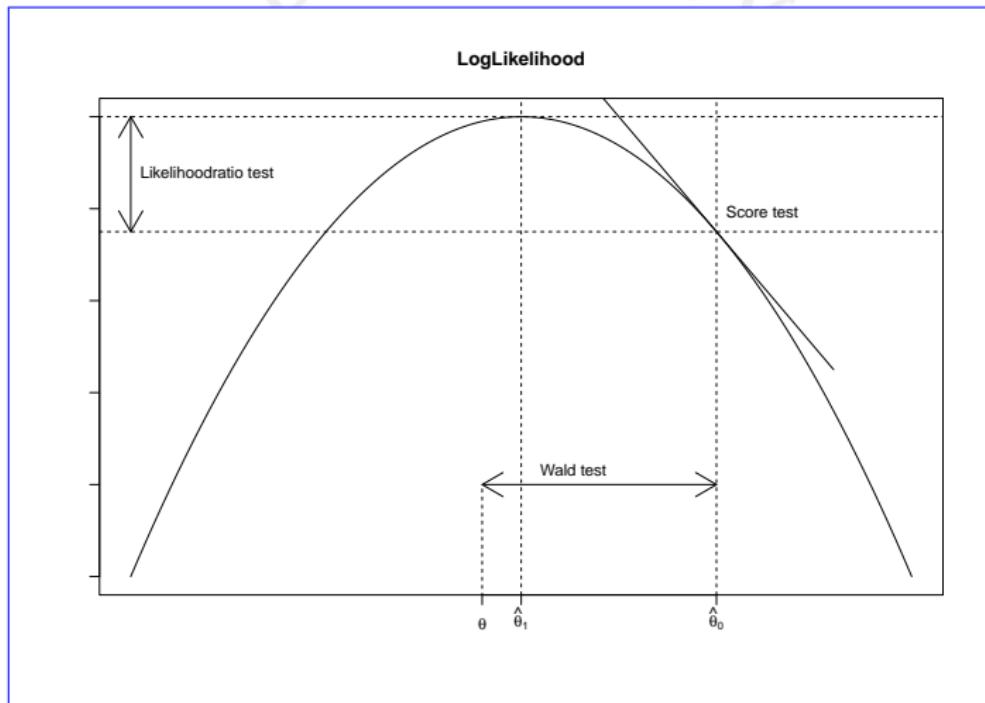
$$\lim_{n \rightarrow \infty} \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} \approx N(0, 1) \implies \lim_{n \rightarrow \infty} \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} \approx \chi_1^2$$

- ▶ Problem: if $|\vartheta|$ is large then $\text{Var}(\hat{\theta})$ will be overestimated
consider changes in log-likelihood instead

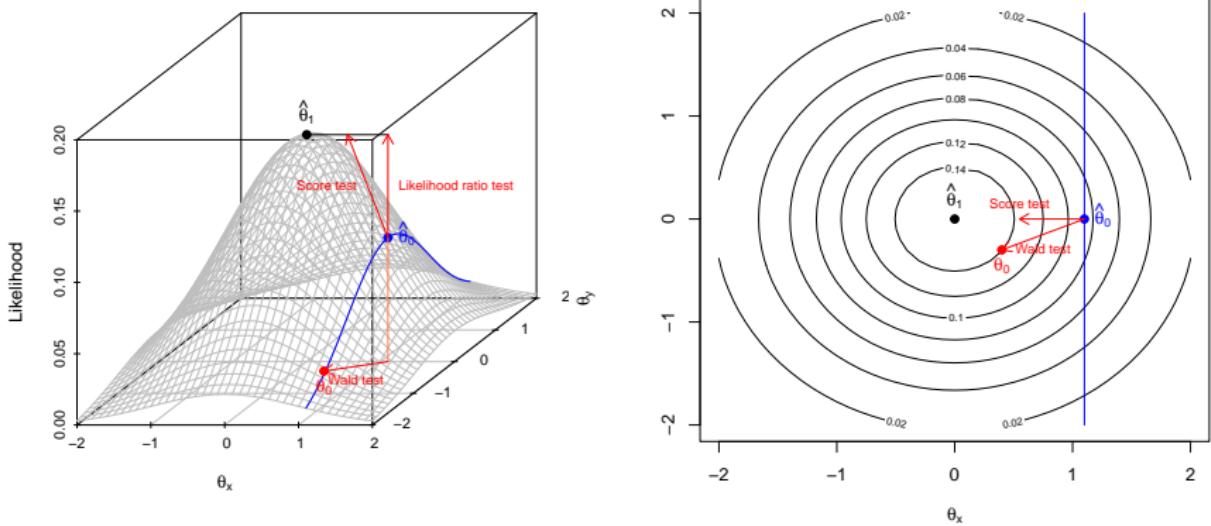
- The *Holy Trinity*

- ▶ Likelihood-ratio test: needs to compute two likelihoods
- ▶ Wald and score test: needs to compute only one likelihood
- ▶ For $n \rightarrow \infty$ the tests are equivalent, but answers may differ in finite samples

HolyTrinity

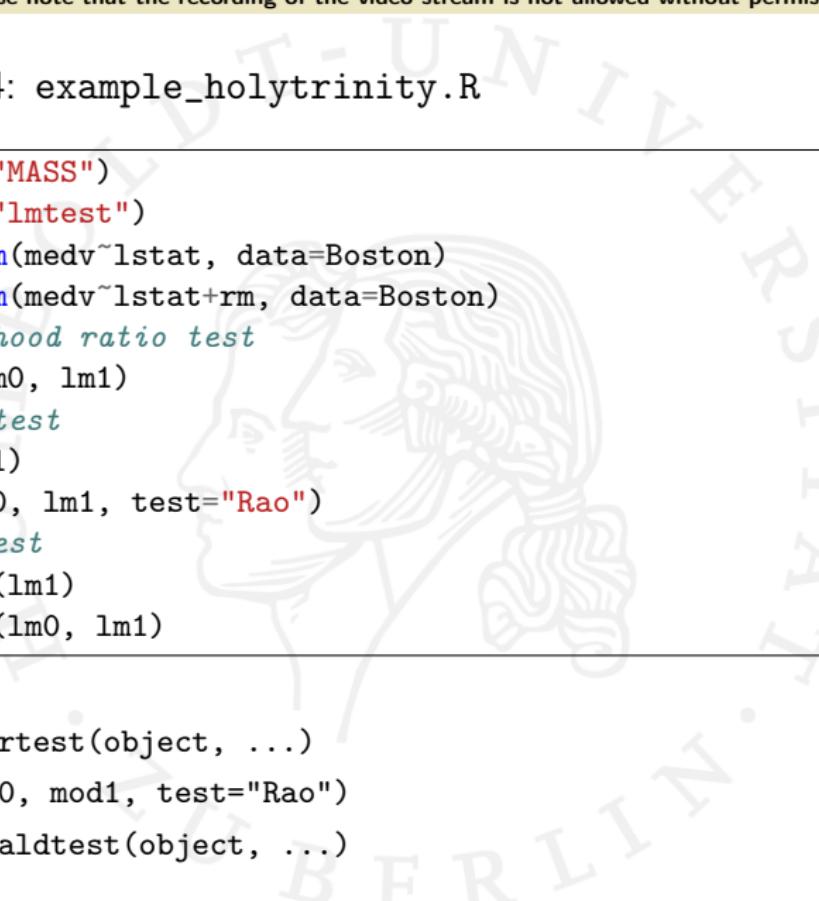


Fox, John (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, Calif: Sage Publications. 597 pp. ISBN: 978-0-8039-4540-1.



3D and contour plot of two nested likelihoods with two parameters
 gray: unrestricted model, blue: restricted model, red: “true” parameter

- Neyman, J. and Pearson, E. S. (July 1928). "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I". In: *Biometrika* 20A.1, p. 175. ISSN: 00063444. DOI: 10.2307/2331945. URL: <http://www.jstor.org/stable/2331945?origin=crossref> (visited on 08/16/2015).
- Wald, Abraham (Nov. 1943). "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large". In: *Transactions of the American Mathematical Society* 54.3, p. 426. ISSN: 00029947. DOI: 10.2307/1990256. URL: <http://www.jstor.org/stable/1990256?origin=crossref> (visited on 08/16/2015).
- Radhakrishna Rao, C. and Bartlett, M. S. (Jan. 1948). "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 44.1, p. 50. ISSN: 0305-0041, 1469-8064. DOI: 10.1017/S0305004100023987. URL: http://www.journals.cambridge.org/abstract_S0305004100023987 (visited on 08/16/2015).

A faint watermark of a classical building with four columns and a triangular pediment is visible in the background.

R Listing 4.4: example_holytrinity.R

```
1 library("MASS")
2 library("lmtest")
3 lm0 <- lm(medv~lstat, data=Boston)
4 lm1 <- lm(medv~lstat+rm, data=Boston)
5 # likelihood ratio test
6 lrtest(lm0, lm1)
7 # score test
8 anova(lm1)
9 anova(lm0, lm1, test="Rao")
10 # wald test
11 waldtest(lm1)
12 waldtest(lm0, lm1)
```

- R lmtest::lrtest(object, ...)
- R anova(mod0, mod1, test="Rao")
- R lmtest::waldtest(object, ...)

Linear regression model (2 variables):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \text{ with } U \sim N(0; \sigma_U^2)$$

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2; \sigma_U^2)$$

$$Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}; \sigma_U^2)$$

Likelihood:

$$\begin{aligned} L((y_1, x_{11}, x_{12}), \dots, (y_n, x_{n1}, x_{n2}) \mid \beta_2, \beta_1, \beta_0, \sigma_U^2) &= \\ \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_U^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2}{\sigma_U^2}\right) \end{aligned}$$

Kullback-Leibler divergence and entropy

$$d_{KL}(f\|g) = \begin{cases} \int_{-\infty}^{\infty} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx & \text{if } f \text{ and } g \text{ continuous} \\ \sum_i f_i \log \left(\frac{f_i}{g_i} \right) & \text{if } f \text{ and } g \text{ categorical} \end{cases}$$

- KL divergence measures how close the distributions are
 - ▶ $d_{KL}(f\|g) \geq 0$, $d_{KL}(f\|g) = 0 \Leftrightarrow f = g$
 - ▶ not symmetric: $d_{KL}(f\|g) \neq d_{KL}(g\|f)$
- For a maximum likelihood estimator $\hat{\theta}_{ML}$ it holds for all θ :

$$d_{KL}(f\|f(\hat{\theta}_{ML})) \leq d_{KL}(f\|f(\theta))$$

- ▶ f the true density of the random sample variables X_i
- ▶ $f(\theta)$ the density induced by θ for X_i

$$\text{Entropy}(f) = \begin{cases} - \int_{-\infty}^{\infty} f(x) \log(f(x)) dx & \text{if } f \text{ is continuous} \\ - \sum_i f_i \log(f_i) & \text{if } f \text{ is categorical} \end{cases}$$

- entropy is a measure of the uncertainty in a random variable
- It holds

$$E(-\log(f(\theta))) = \underbrace{\text{Entropy}(f)}_{=const.} + d_{KL}(f \| f(\theta))$$

- ▶ maximizing the loglikelihood is equivalent to minimizing the negative loglikelihood
- ▶ minimizing the negative loglikelihood is equivalent to minimizing the KL divergence

Space of distributions (infinite dimensional !)

Normal distributions

$$\bullet N(\mu, \sigma^2) = f(\theta)$$

$$d_{KL}(N(\mu, \sigma^2) || f)$$

$$N(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML}) = f(\hat{\theta}_{ML})$$

• true data distribution f

Akaike Information Criterion

- Akaike (1974) showed
 - ▶ the negative maximized loglikelihood is an biased estimator for the KL divergence
 - ▶ the bias is asymptotically k , the number of estimated parameters in the model
- Akaike information criterion

$$AIC = -2 \log(L(x_1, \dots, x_n | \hat{\theta}_{ML})) + 2k$$

- Small Sample AIC (AICc) (Hurvich & Tsai, 1989)

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- ▶ normal errors assumed, but not sensitive to departures from normality

Bayesian Information Criterion

- If the sample size becomes larger then smaller effects become more relevant
- AIC does not depend on sample size and therefore chooses larger models
- Schwarz (1978) suggested the Bayesian Information Criterion (BIC)

$$BIC = -2 \log(L(x_1, \dots, x_n | \hat{\theta}_{ML})) + k \log(n)$$

- ▶ assumption: data come from a distribution of the exponential family
- BIC vs. AIC
 - ▶ BIC prefers models with less parameters than AIC
 - ▶ AIC model better for reducing the prediction error
 - ▶ BIC assumes that the true model is under the candidate models

Information criteria

- Shannon, C. E. (July 1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 00058580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6773024> (visited on 08/16/2015).
- Kullback, S. and Leibler, R. A. (Mar. 1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729694. URL: <http://projecteuclid.org/euclid.aoms/1177729694> (visited on 08/16/2015).
- Akaike, H. (Dec. 1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1100705> (visited on 08/16/2015).
- Schwarz, Gideon (Mar. 1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464. ISSN: 0090-5364. DOI: 10.1214/aos/1176344136. URL: <http://projecteuclid.org/euclid.aos/1176344136> (visited on 08/16/2015).
- Hurvich, Clifford M. and Tsai, Chih-Ling (1989). "Regression and time series model selection in small samples". In: *Biometrika* 76.2, pp. 297–307. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/76.2.297. URL: <http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/76.2.297> (visited on 08/16/2015).

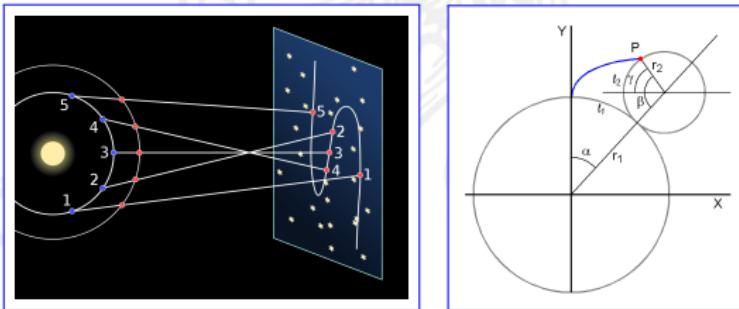
Occam's razor

- William of Ockham (1287-1347):

Entia non sunt multiplicanda praeter necessitatem
(entities must not be multiplied beyond necessity)
- The exact phrase can not be found in his work
- Occam's razor: "When you have two competing theories that make exactly the same predictions, the simpler one is the better."
- Combines explanatory power and complexity of a model
- Information criteria formalize this
 - ▶ large likelihood: model fits well the data
 - ▶ complexity: number of parameters

Example 4.13 (Geocentric vs. heliocentric model)

- Main problem: easter date (julian year 11 minutes too long)
- Ptolemy (~ 95-168) supported a geocentric model
- Copernicus (1473-1543) began the Copernican Revolution with heliocentric model



Copernicus, Nicolaus (1543). *De revolutionibus orbium coelestium*. Nuremberg, Holy Roman Empire: Printed by Johannes Petreius.

Ptolemy, Claudius (around 150). *Almagest (Mathematike Syntaxis)*.

Parameter of distributions

November 3, 2022

- Central tendency
- One sample Gauss test
- One sample t-test
- One sample Binomial test
- One sample sign test
- One sample median test
- Confidence interval
- Quantile
- Quantile confidence interval
- Dispersion
- Entropy
- Gini impurity index
- Dispersion summary
- Skewness
- Excess and kurtosis
- Higher moments
- Summarize continuous variables
- Summarize categorical variables
- Empirical cumulative distribution function

Central tendency

- Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Median

$$\tilde{x} = \begin{cases} n \text{ odd} & x\left(\frac{n+1}{2}\right) \\ n \text{ even} & \frac{1}{2} \left(x\left(\frac{n}{2}\right) + x\left(\frac{n}{2}+1\right) \right) \end{cases}$$

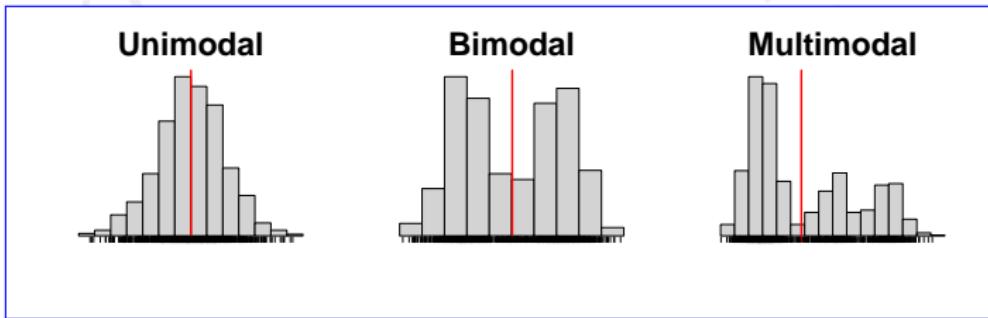
- Mode

$$\hat{x} = \{x_i \mid \max_i h(x_i)\}$$

- Required measurement level

parameter	nominal	ordinal	metric discrete	metric continuous
mean	no	no	yes	yes
median	no	maybe	maybe	yes
modus	yes	yes	yes	maybe

- Distribution should be unimodal otherwise the central tendency might be misleading



- rule-of-thumb: coefficient of variation (s/\bar{x}) greater than 0.5 is an indication that the average is not a suitable representative of the individual values

Eckstein, Peter P. (2014). *Repetitorium Statistik*. Wiesbaden: Springer Fachmedien Wiesbaden.
ISBN: 978-3-658-05747-3 978-3-658-05748-0. DOI: 10.1007/978-3-658-05748-0. URL:
<http://link.springer.com/10.1007/978-3-658-05748-0> (visited on 09/01/2021).

 Listing 5.1: example_location.R

```
1 data(Boston, package="MASS")
2 # mean
3 mean(Boston$medv)
4 # median
5 median(Boston$medv)
6 # mode, see http://stackoverflow.com/questions/2547402
7 Mode <- function(x) {
8   ux <- unique(x)
9   ux[which.max(tabulate(match(x, ux)))]
10 }
11 Mode(Boston$rad)
```

 `mean(x, na.rm=F)`

 `median(x, na.rm=F)`

 `DescTools::Mode(x, na.rm = FALSE)`

One sample Gauss test

Assumptions: σ known and either $X_i \sim N(\mu; \sigma)$ or
 $X_i \sim (\mu; \sigma)$ and $n > 30$ (central limit theorem)

Hypotheses: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

Test statistics: $T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \approx N(0; 1)$

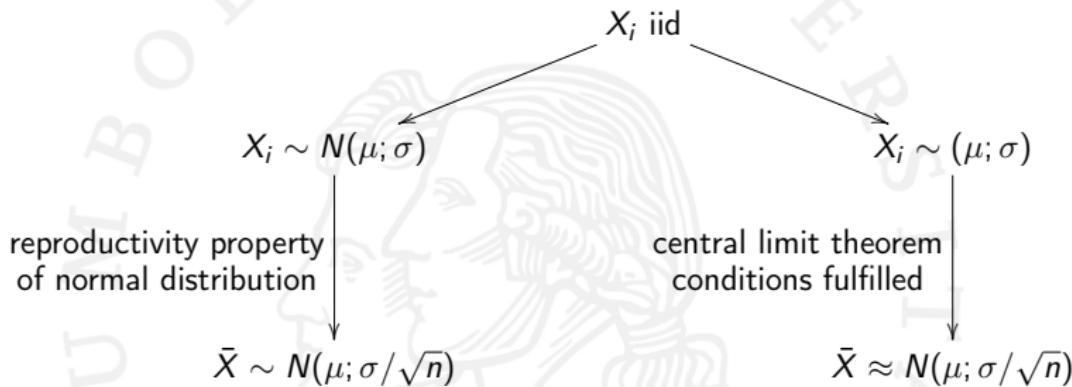
Reject H_0 : $|t| > z_{1-\alpha/2}$

$t > +z_{1-\alpha}$ if $H_0 : \mu \leq \mu_0$

$t < -z_{1-\alpha}$ if $H_0 : \mu \geq \mu_0$

Remark(s): other name: Z-test
One-sided tests ($H_0 : \mu \leq \mu_0$, $H_0 : \mu \geq \mu_0$) are preferred

Derivation of test statistics



σ known:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0; 1)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0; 1)$$

σ unknown:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \stackrel{n > 30}{\approx} N(0; 1)$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx t_{n-1} \stackrel{n > 30}{\approx} N(0; 1)$$

One sample t-test

Assumptions: σ unknown and either $X_i \sim N(\mu; \sigma)$ or
 $X_i \sim (\mu; \sigma)$ and $n > 30$ (central limit theorem)

Hypotheses: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

Test statistics: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \stackrel{n > 30}{\approx} N(0; 1)$

Reject H_0 : $|t| > t_{n-1; 1-\alpha/2}$

$t > +t_{n-1; 1-\alpha}$ if $H_0 : \mu \leq \mu_0$

$t < -t_{n-1; 1-\alpha}$ if $H_0 : \mu \geq \mu_0$

Remark Because of CLT the test can be applied in most situations

Student (Mar. 1, 1908). "The probable error of a mean". In: *Biometrika* 6.1, pp. 1–25. ISSN: 0006-3444, 1464-3510. DOI: [10.1093/biomet/6.1.1](https://doi.org/10.1093/biomet/6.1.1). URL: <http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/6.1.1> (visited on 08/14/2015).

⌚ Listing 5.2: example_onettest.R

```
1 data(Boston, package="MASS")
2 t.test(Boston$medv, mu=10)
```

⌚ `t.test(x, alternative=c("two.sided", "less", "greater"), mu=0)`

Consider $X_i \sim N(\mu; \sigma)$ with

- σ unknown, acceptance region of H_0

$$\mu_0 - t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$$

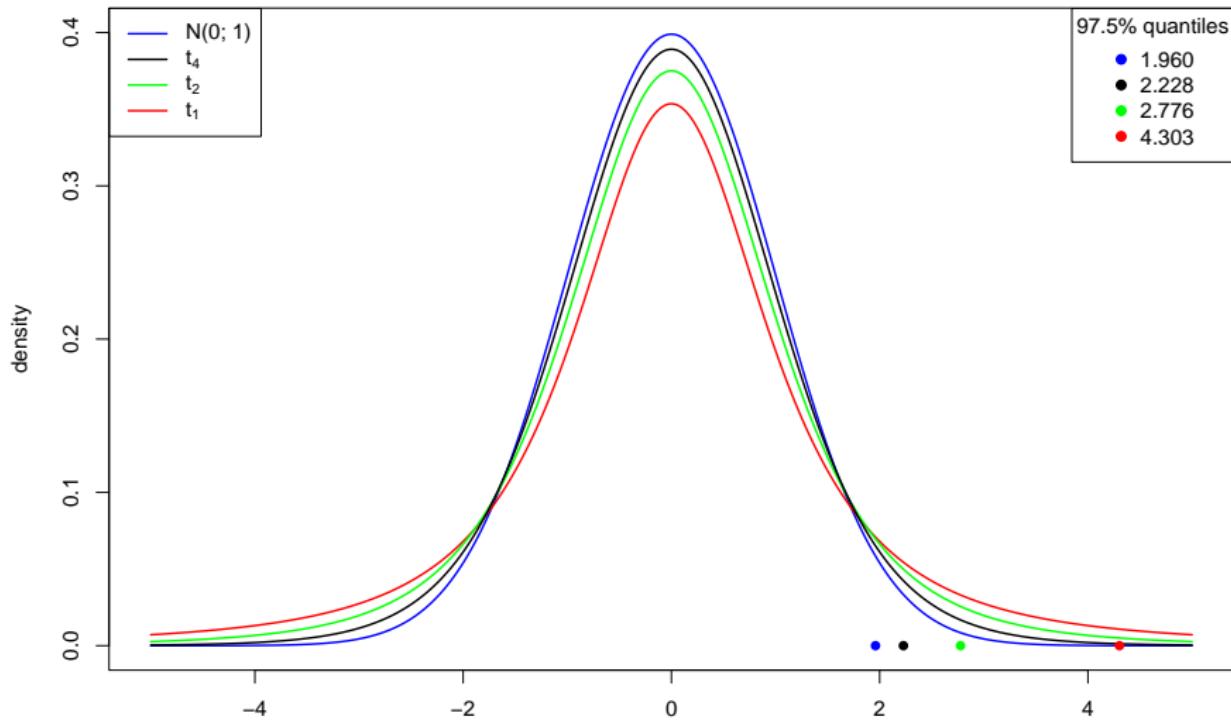
- σ known, acceptance region of H_0

$$\mu_0 - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Even if our estimate $s = \sigma$ then the acceptance region for σ known (more information !) is smaller:

$$t_{n-1;1-\alpha/2} \geq z_{1-\alpha/2}$$

t-distribution with n degrees of freedom: densities and 97.5%-quantiles



One sample Binomial test

Assumptions: Independent random experiments with binary outcome and constant success probability π

Approximation possible if $n\pi_0(1 - \pi_0) > 9$

Hypotheses: $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$

Test statistics: $V = X \sim B(\pi_0; n)$

$$\text{(approximation)} \quad Z = \frac{\frac{X}{n} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \approx N(0; 1)$$

Reject H_0 : see critical values from table

(approximation) $|z| > z_{1-\alpha/2}$

Clopper, C. J. and Pearson, E. S. (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial". In: *Biometrika* 26.4, pp. 404–413. ISSN: 0006-3444, 1464-3510.
 DOI: 10.1093/biomet/26.4.404. URL:
<http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/26.4.404> (visited on 08/14/2015).

 Listing 5.3: example_binomtest.R

```
1 data(Boston, package="MASS")
2 # Sign test
3 # H0: pi = pi0
4 # H1: pi <> pi0
5 binom.test(sum(Boston$chas==1), length(Boston$chas))
6 prop.test(sum(Boston$chas==1), length(Boston$chas))
```

☞ `binom.test(x, n, p=0.5, alternative=c("two.sided", "less",
"greater"), conf.level=0.95)`

☞ `prop.test (x, n, p=0.5, alternative=c("two.sided", "less",
"greater"), conf.level=0.95, correct=T)`

⚠ Uses Yates' continuity correction

One sample sign test

Assumption(s): X_i is continuous metric

Hypotheses: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs. $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$

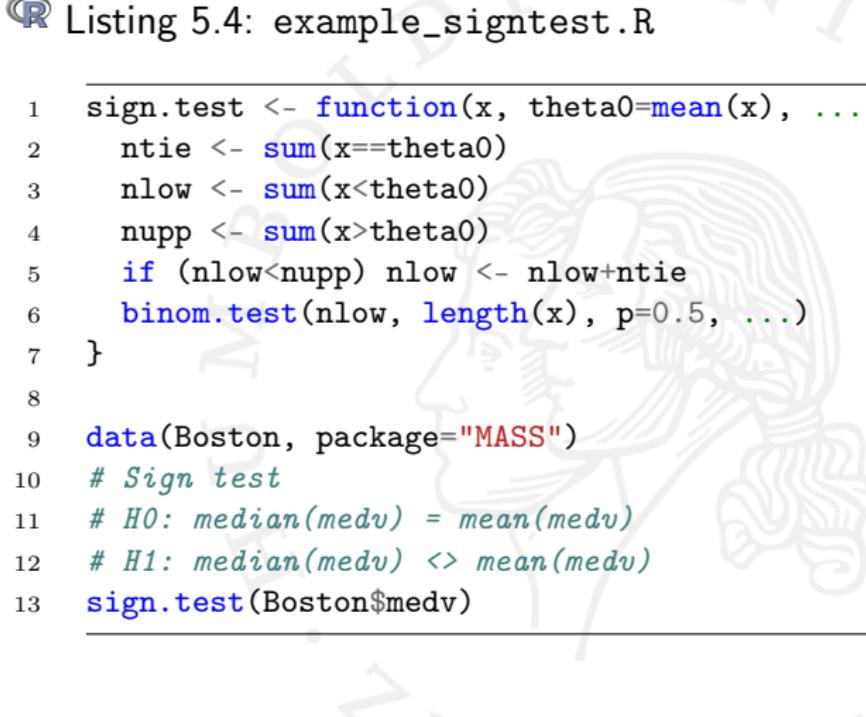
Test statistics: V : "Number of observations smaller than $\tilde{\mu}_0$ "
 $V \sim B(n; p = 0.5) \stackrel{n > 36}{\approx} N(n/2; \sqrt{n}/2)$

Reject H_0 : $|v| > c_{\text{krit}}$ from Table

Problem: $x_i = \tilde{\mu}_0$

- Delete observation with $x_i = \tilde{\mu}_0 \Rightarrow$ reduce n
- Increase v by a $\#\{x_i = \tilde{\mu}_0\}/2$
- With probability 0.5 assign randomly to $x_i < \tilde{\mu}_0$
- Assign x_i to the group $(x_i < \tilde{\mu}_0, x_i > \tilde{\mu}_0)$ with lower numbers of observations

Dixon, W. J. and Mood, A. M. (Dec. 1946). "The Statistical Sign Test". In: *Journal of the American Statistical Association* 41.236, p. 557. ISSN: 01621459. DOI: 10.2307/2280577.
URL: <http://www.jstor.org/stable/2280577?origin=crossref> (visited on 08/14/2015).

A large, faint watermark of a stylized brain in profile, facing left, is centered behind the text.

R Listing 5.4: example_signtest.R

```
1 sign.test <- function(x, theta0=mean(x), ...) {  
2   ntie <- sum(x==theta0)  
3   nlow <- sum(x<theta0)  
4   nupp <- sum(x>theta0)  
5   if (nlow<nupp) nlow <- nlow+ntie  
6   binom.test(nlow, length(x), p=0.5, ...)  
7 }  
8  
9 data(Boston, package="MASS")  
10 # Sign test  
11 # H0: median(medv) = mean(medv)  
12 # H1: median(medv) <> mean(medv)  
13 sign.test(Boston$medv)
```

R `binom.test(x, n, alternative=c("two.sided", "less", "greater"))`

One sample median test

Assumption(s): X_i is continuous metric

Hypotheses: $H_0 : \tilde{\mu} = \tilde{\mu}_0$ vs. $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$

Test statistics: N_0 : "Number of observations smaller than $\tilde{\mu}_0$ "
 N_1 : "Number of observations larger than $\tilde{\mu}_0$ "

$$V = \frac{(N_0 - \frac{n}{2})^2}{n/2} + \frac{(N_1 - \frac{n}{2})^2}{n/2} \stackrel{n > 9}{\approx} \chi^2_1$$

Reject H_0 : $|v| > c\chi^2_{1;1-\alpha/2}$ from Table

Remark based on χ^2 goodness-of-fit test, bad power

Brown, G.W. and Mood, A. M. (1951). "On Median Tests for Linear Hypotheses". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Second Berkeley Symposium on Mathematical Statistics and Probability. Statistical Laboratory of the University of California, Berkeley: University of California Press, pp. 159–166.

Freidlin, Boris and Gastwirth, Joseph L. (Aug. 2000). "Should the Median Test be Retired from General Use?" In: *The American Statistician* 54.3, pp. 161–164. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2000.10474539. URL: <http://www.tandfonline.com/doi/abs/10.1080/00031305.2000.10474539> (visited on 08/14/2015).

⌚ Listing 5.5: example_median_test.R

```
1 data(Boston, package="MASS")
2 #
3 library("UsingR")
4 simple.median.test(Boston$medv, mean(Boston$medv))
```

⌚ UsingR::simple.median.test(x, median=NA)

Confidence interval

- $1 - \alpha$ confidence interval for the mean (with CLT)

aionsa σ unknown

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx t_{n-1} \stackrel{n>30}{\approx} N(0; 1)$$

$$\Rightarrow \left[\bar{x} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right]$$

- Approx. $1 - \alpha$ confidence interval for the median (X roughly normal)

$$\tilde{X} \stackrel{n \rightarrow \infty}{\approx} N(\tilde{\mu}; \sigma_{\tilde{X}})$$

with $\hat{\sigma}_{\tilde{X}} = \frac{1.25 \text{ IQR}}{1.35\sqrt{n}}$

$$\Rightarrow \left[\tilde{x} - z_{1-\alpha/2} \frac{1.25 \text{ IQR}}{1.35\sqrt{n}}; \tilde{x} + z_{1-\alpha/2} \frac{1.25 \text{ IQR}}{1.35\sqrt{n}} \right]$$

McGill, Robert, Tukey, John W., and Larsen, Wayne A. (Feb. 1978). "Variations of Box Plots". In: *The American Statistician* 32.1, p. 12. ISSN: 00031305. DOI: 10.2307/2683468. URL: <http://www.jstor.org/stable/2683468?origin=crossref> (visited on 08/11/2015).

Quantile

- Problem 1: Estimate the p -quantile x_p from observations ($F(x_p) = p$)
- Problem 2: Estimate for an observation x_i a percentile rank p_i ($F(x_i) = p_i$)
- Special quantiles:
 - ▶ quartiles: 25%, 50% and 75% quantiles
 - ▶ percentiles: 1%, 2%, ..., 99% quantiles
 - ▶ percentiles often used as synonym to quantiles
- Quantile estimation is based on order statistics either by
 1. discontinuous estimation ($\gamma = 0, 0.5$, or 1)
$$x_p = (1 - \gamma)x_{(j)} + \gamma x_{(j+1)}$$
 2. or continuous estimation by linear interpolation between
$$(p_j, x_{(j)}) \text{ and } (p_{j+1}, x_{(j+1)})$$

⚠ Finite number of observations and continuous interval for $p \in [0; 1]$

- Methods differ in the choice of m and γ

- ▶ $j = \lfloor np + m \rfloor$ (integer)
 - ▶ $g = np + m - j$ (decimal places)

- Discontinuous estimation ($m = 0$)

1. inverse of empirical distribution function:

$$g = 0 \Rightarrow \gamma = 0 \text{ otherwise } \gamma = 1$$

2. as before but with averaging at discontinuities:

$$g = 0 \Rightarrow \gamma = 0.5 \text{ otherwise } \gamma = 1$$

3. nearest even order statistic (SAS):

$$g = 0 \text{ and } j \text{ even } \Rightarrow \gamma = 0 \text{ otherwise } \gamma = 1$$

- Continuous estimation ($\gamma = g$)

4. linear interpolation of the empirical distribution function:

$$m = 0, p_k = k/n$$

5. Rankit:

$$m = 0.5, p_k = (k - 0.5)/n$$

6. Van der Waerden (SPSS):

$$m = p, p_k = k/(n + 1)$$

7. Default (R):

$$m = 1 - p, p_k = (k - 1)/(n - 1)$$

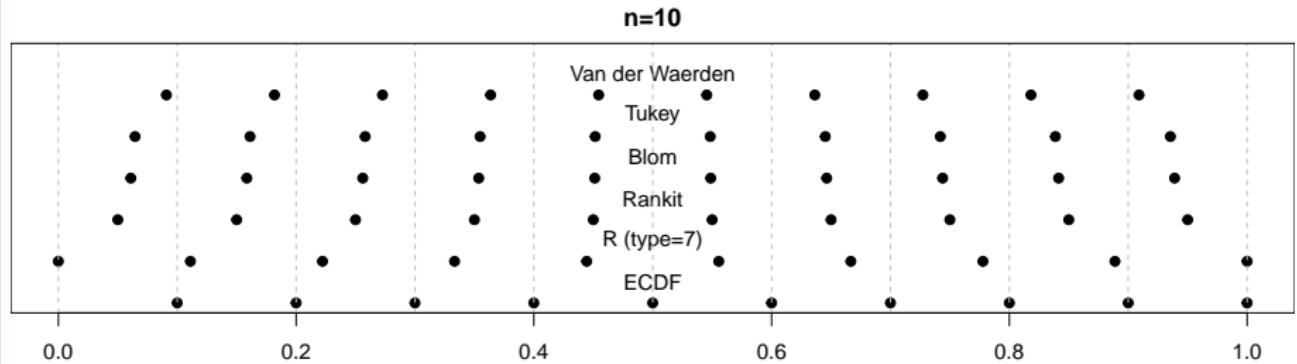
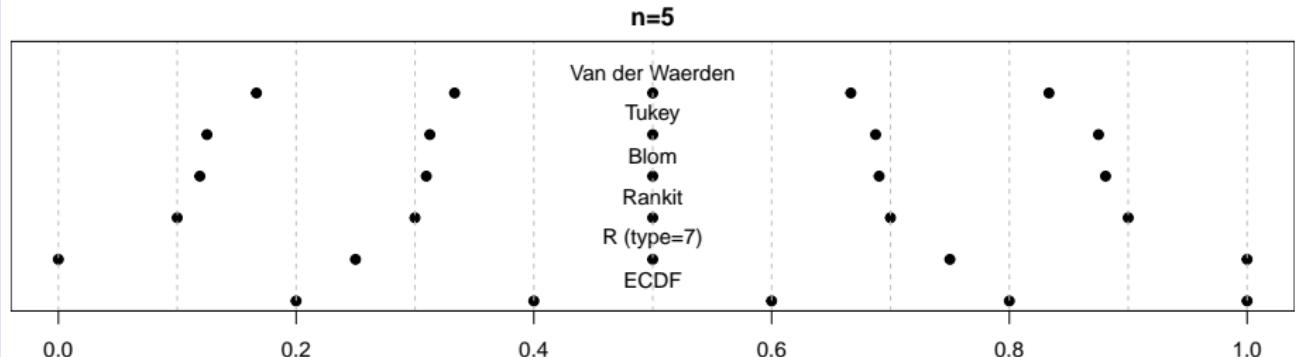
8. Tukey:

$$m = (p + 1)/3, p_k = (k - 1/3)/(n + 1/3)$$

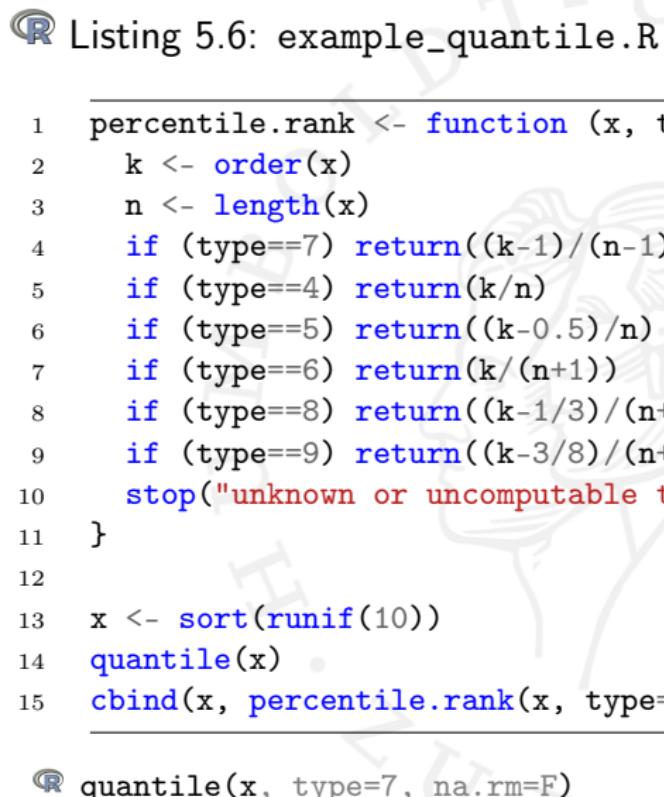
9. Blom:

$$m = p/4 + 3/8, p_k = (k - 3/8)/(n + 1/4)$$

Choice of p_k for $n = 5$ and 10 observations with different methods



```
R quantile(x, probs=seq(0, 1, 0.25), type=7, na.rm=F)
```

R Listing 5.6: example_quantile.R

```
1 percentile.rank <- function (x, type=7) {  
2   k <- order(x)  
3   n <- length(x)  
4   if (type==7) return((k-1)/(n-1))  
5   if (type==4) return(k/n)  
6   if (type==5) return((k-0.5)/n)  
7   if (type==6) return(k/(n+1))  
8   if (type==8) return((k-1/3)/(n+1/3))  
9   if (type==9) return((k-3/8)/(n+1/4))  
10  stop("unknown or uncomputable type")  
11 }  
12  
13 x <- sort(runif(10))  
14 quantile(x)  
15 cbind(x, percentile.rank(x, type=9))
```

R quantile(x, type=7, na.rm=F)

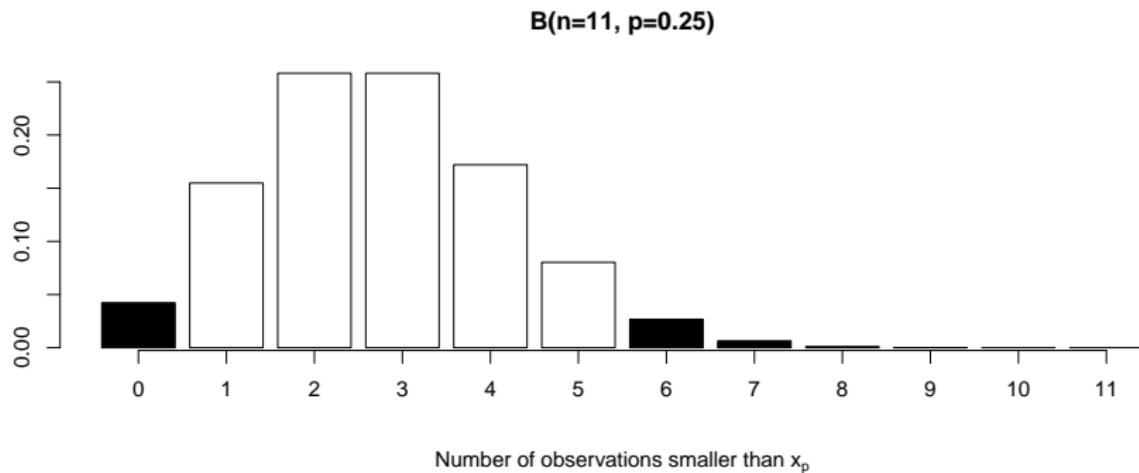
R	Maple	SPSS	Excel	SAS	Minitab
---	-------	------	-------	-----	---------

type	method	PCTLDEF			
1	1			3	
2	2			5	
3				2	
4	3			1	
5	4	Rankit			
6	5	v.d. Waerden*	xxx.EXC	4	Default
7*	6		xxx.INC, xxx*		
8	7	Tukey			
9	8	Blom			

* Default method

Quantile confidence interval

- Find $1 - \alpha$ confidence interval for x_p
- Consider the sample as n independent random draws
- “The number of observations less equal than the p quantile” is $B(n; p)$ distributed
- To find the confidence interval use the binomial distribution function



- It holds, independent of the distribution of X , that

$$P(X_{(l)} \leq x_p \leq X_{(u)}) = \sum_{i=l}^{u-1} \binom{n}{i} p^i (1-p)^{n-i}$$

- and

$$\begin{aligned} P(X_{(l)} \leq x_p \leq X_{(u)}) &= \sum_{i=l}^{u-1} \binom{n}{i} p^i (1-p)^{n-i} \\ &= F_B(u-1) - F_B(l-1) \\ &\geq 1 - \alpha \end{aligned}$$

Scheffe, H. and Tukey, J. W. (June 1945). "Non-Parametric Estimation. I. Validation of Order Statistics". In: *The Annals of Mathematical Statistics* 16.2, pp. 187–192. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731119. URL: <http://projecteuclid.org/euclid.aoms/1177731119> (visited on 08/11/2015).

Example 5.14

- Find for eleven observations the 90% confidence interval of the 25% quantile:

i	1	2	3	4	5	6	7	8	9	10	11
$x(i)$	4.9	5.0	5.1	5.2	5.3	5.4	5.6	5.8	5.9	6.0	6.5

- $F_B(0; n = 11; p = 0.25) = 0.042 < 0.05 \Rightarrow l - 1 = 0$,
 $F_B(5; n = 11; p = 0.25) = 0.966 > 0.95 \Rightarrow u - 1 = 5$
- The 92,4% confidence interval is $[x_{(1)}; x_{(6)}] = [4.9; 5.4]$
- Because of the symmetry of the binomial distribution the 92,4% confidence interval for the 75% quantile is $[x_{(6)}; x_{(11)}] = [5.4; 6.5]$

Dispersion

- Variance in descriptive statistics

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

► average squared deviation from the mean

- Variance in inferential statistics

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

► unbiased estimator for the unknown population variance σ^2

- Standard deviation: square root of variance

- Variance as average squared difference of observations (Gini)

$$\frac{1}{n(n-1)} \sum_{i < j} (x_i - x_j)^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\frac{1}{n^2} \sum_{i \leq j} (x_i - x_j)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Further coefficients

- ▶ Range

$$r = \max_i x_i - \min_i x_i$$

- ▶ Coefficient of variation (relative variability)

$$v_x = \frac{s_x}{\bar{x}}$$

R Listing 5.7: example_variance.R

```
1 data(Boston, package="MASS")
2 # variance an standard deviation
3 var(Boston$medv)
4 sd(Boston$medv)
5 # range
6 diff(range(Boston$medv))
7 # coefficient of variation
8 sd(Boston$medv)/mean(Boston$medv)
```

R var(x, na.rm=F)

R sd(x, na.rm=F)

Entropy

- Measure of concentration
- Entropy (Shannon)

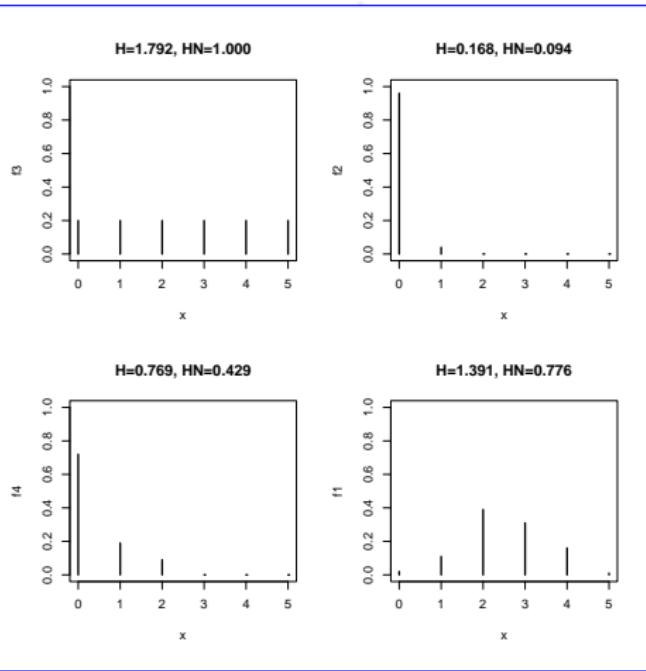
$$H = - \sum_{i=1}^K f_i \log(f_i)$$

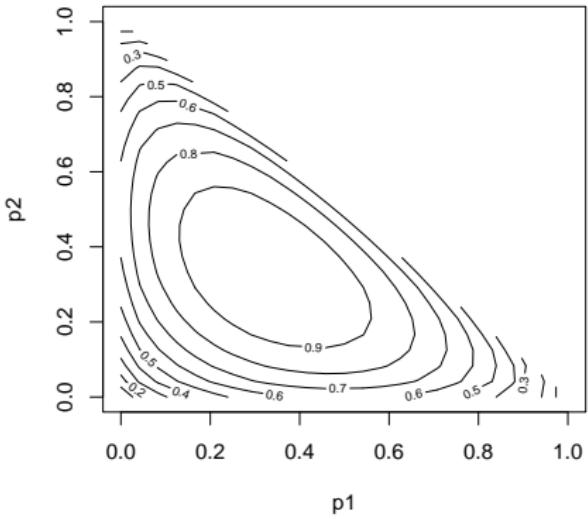
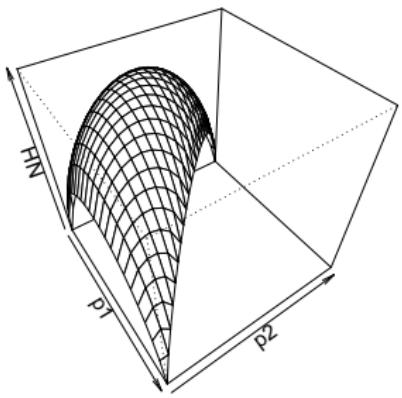
► $0 \leq H \leq \log(K)$

- Normalized entropy

$$HN = -\frac{1}{\log(K)} \sum_{i=1}^K f_i \log(f_i)$$

► $0 \leq HN \leq 1$





Normalized entropy for $K = 3$

R Listing 5.8: example_entropy.R

```
1 data(Boston, package="MASS")
2 library("entropy")
3 ftab <- table(Boston[,9])
4 entropy(ftab)           # estimates the entropy from data
5 entropy(ftab)/length(ftab) # estimates the normalized entropy from da
```

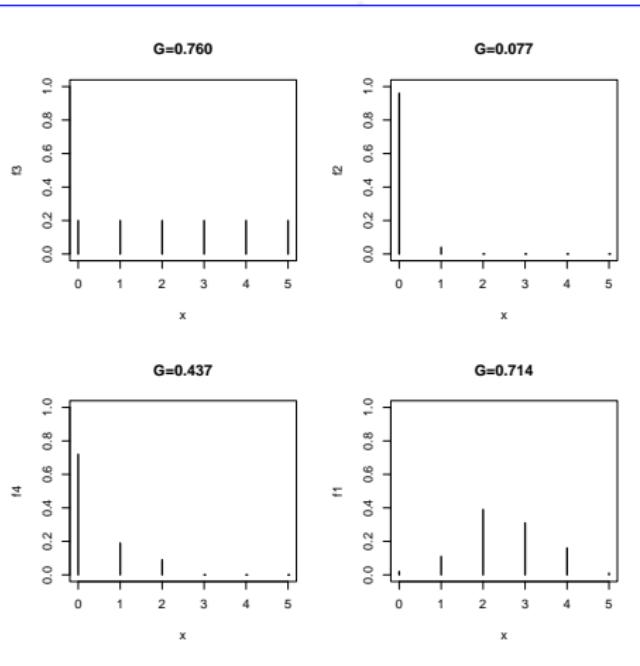
R `entropy::entropy(counts, unit=c("log", "log2", "log10"),
method=c("ML", ...))`

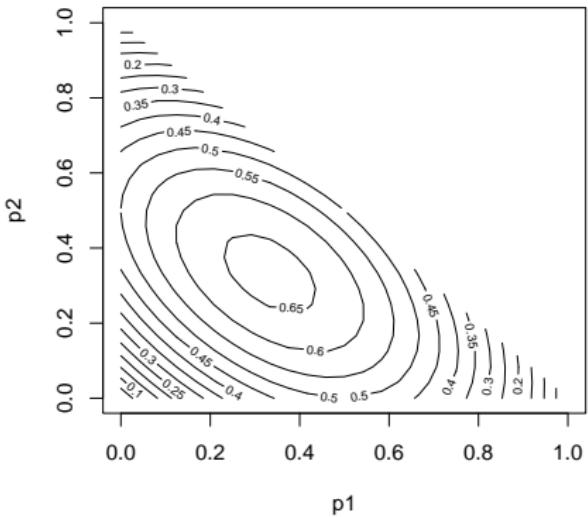
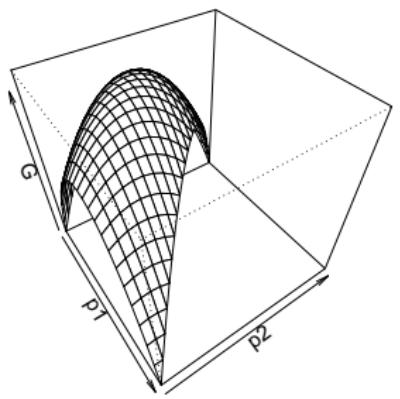
Gini impurity index

- Measure of concentration
- Gini impurity index

$$G = 1 - \sum_{i=1}^K f_i^2 = \sum_{i=1}^K f_i(1 - f_i)$$

- ▶ the chance that we draw (with replacement) two times different values
- Entropy and Gini impurity index are used in Classification and Regression trees (CART)





$$K = 3$$

R Listing 5.9: example_impurity.R

```
1 data(Boston, package="MASS")
2 ftab <- table(Boston$rad)
3 p    <- prop.table(ftab)
4 sum(p*(1-p))                      # gini impurity
```

Dispersion summary

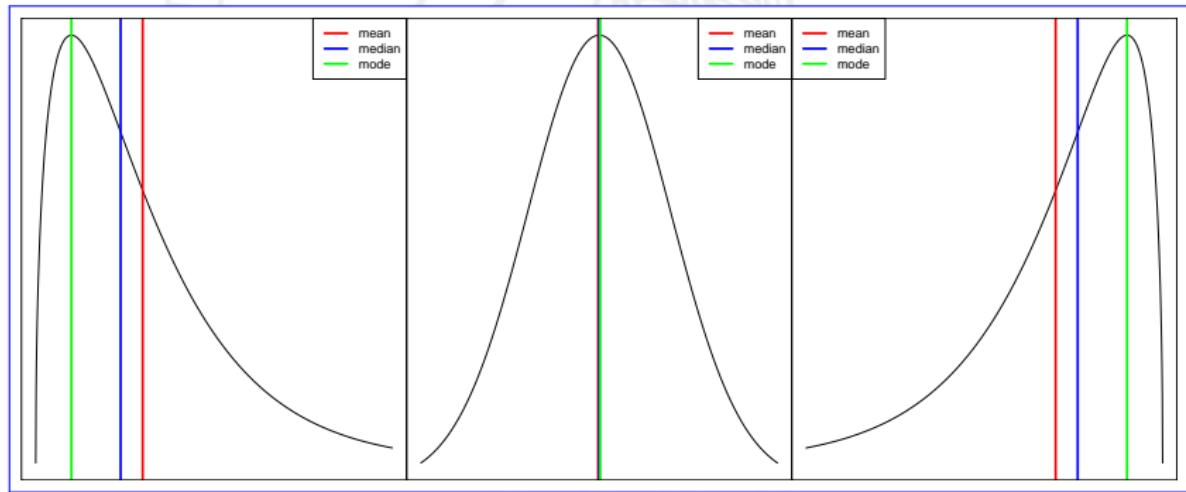
- Required measurement level

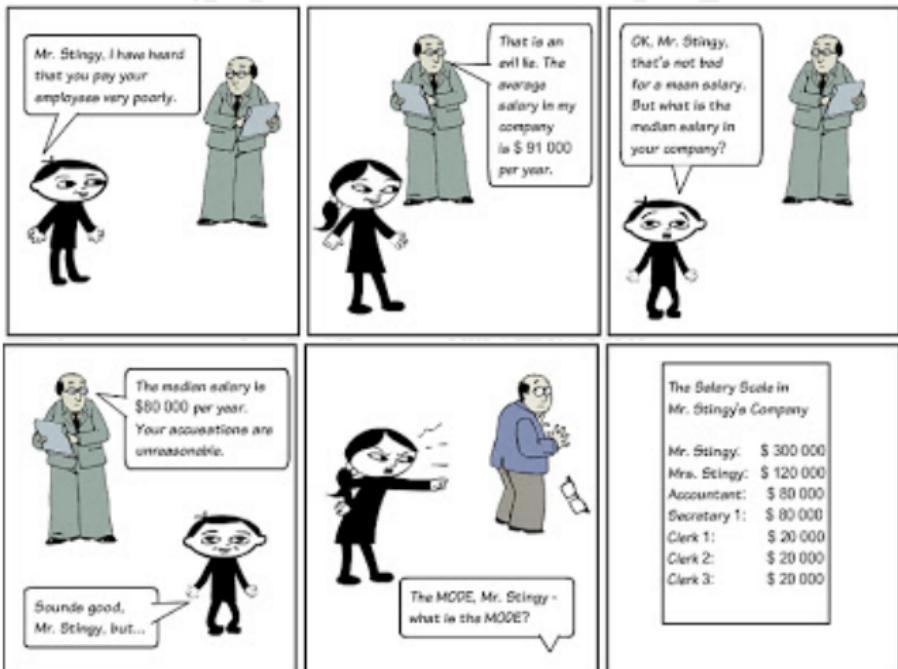
parameter	nominal	ordinal	metric discrete	metric continuous
variance	no	no	yes	yes
standard deviation	no	no	yes	yes
range	no	no	yes	yes
coefficient of variation	no	no	yes	yes
entropy	yes	yes	maybe	no
impurity	yes	yes	maybe	no

Skewness

- Comparison of mean \bar{x} , median \tilde{x} and mode \hat{x}

$\hat{x} < \tilde{x} < \bar{x}$: right skewed $\bar{x} = \tilde{x} = \hat{x}$: symmetric $\hat{x} > \tilde{x} > \bar{x}$: left skewed





Source: [Hanna's blog](#)

right skewed	symmetric	left skewed
$\frac{1}{2}(x_{0.25} + x_{0.75}) > x_{0.5}$	$\frac{1}{2}(x_{0.25} + x_{0.75}) = x_{0.5}$	$\frac{1}{2}(x_{0.25} + x_{0.75}) < x_{0.5}$
$x_{0.25} + IQR/2 > x_{0.5}$	$x_{0.25} + IQR/2 = x_{0.5}$	$x_{0.25} + IQR/2 < x_{0.5}$
$x_{0.75} - IQR/2 > x_{0.5}$	$x_{0.75} - IQR/2 = x_{0.5}$	$x_{0.75} - IQR/2 < x_{0.5}$
$x_{0.75} - x_{0.5} > x_{0.5} - x_{0.25}$	$x_{0.75} - x_{0.5} = x_{0.5} - x_{0.25}$	$x_{0.75} - x_{0.5} < x_{0.5} - x_{0.25}$

- interquartile range $IQR = x_{0.75} - x_{0.25}$, median $x_{0.5} = \tilde{x}$
- skewness (Pearson-Fisher)

$$g = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- ▶ right skewed: $g > 0$
- ▶ symmetric: $g = 0$
- ▶ left skewed: $g < 0$

Pearson, K. (June 1, 1905). "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A rejoinder". In: *Biometrika* 4.1, pp. 169–212. issn: 0006-3444, 1464-3510. doi: 10.1093/biomet/4.1-2.169. url:

<http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/4.1-2.169> (visited on 06/08/2016).

Fisher, R. A. (Jan. 1, 1930). "Moments and Product Moments of Sampling Distributions". In: *Proceedings of the London Mathematical Society* s2-30.1, pp. 199–238. issn: 0024-6115, 1460-244X. doi: 10.1112/plms/s2-30.1.199. url:

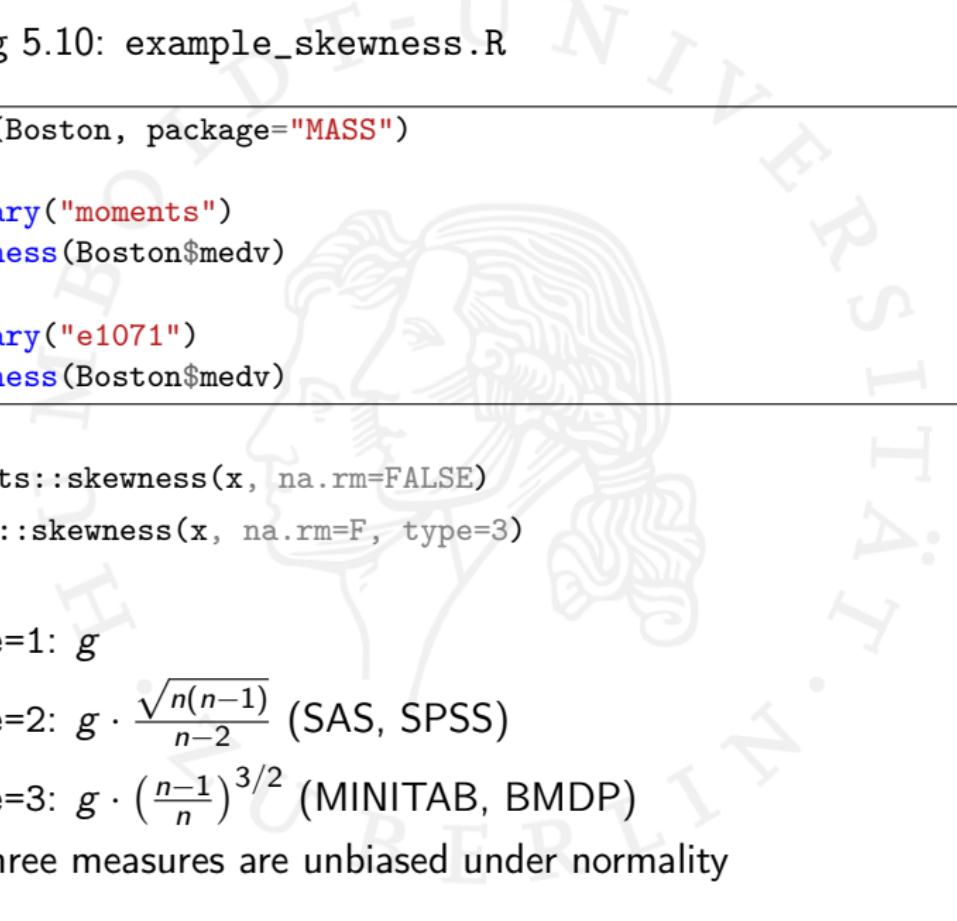
<http://plms.oxfordjournals.org/cgi/doi/10.1112/plms/s2-30.1.199> (visited on 06/08/2016).

Listing 5.10: example_skewness.R

```

1  data(Boston, package="MASS")
2  #
3  library("moments")
4  skewness(Boston$medv)
5  #
6  library("e1071")
7  skewness(Boston$medv)

```


 moments::skewness(x, na.rm=FALSE)
 e1071::skewness(x, na.rm=F, type=3)

- type=1: g
- type=2: $g \cdot \frac{\sqrt{n(n-1)}}{n-2}$ (SAS, SPSS)
- type=3: $g \cdot \left(\frac{n-1}{n}\right)^{3/2}$ (MINITAB, BMDP)
- all three measures are unbiased under normality

Excess and kurtosis

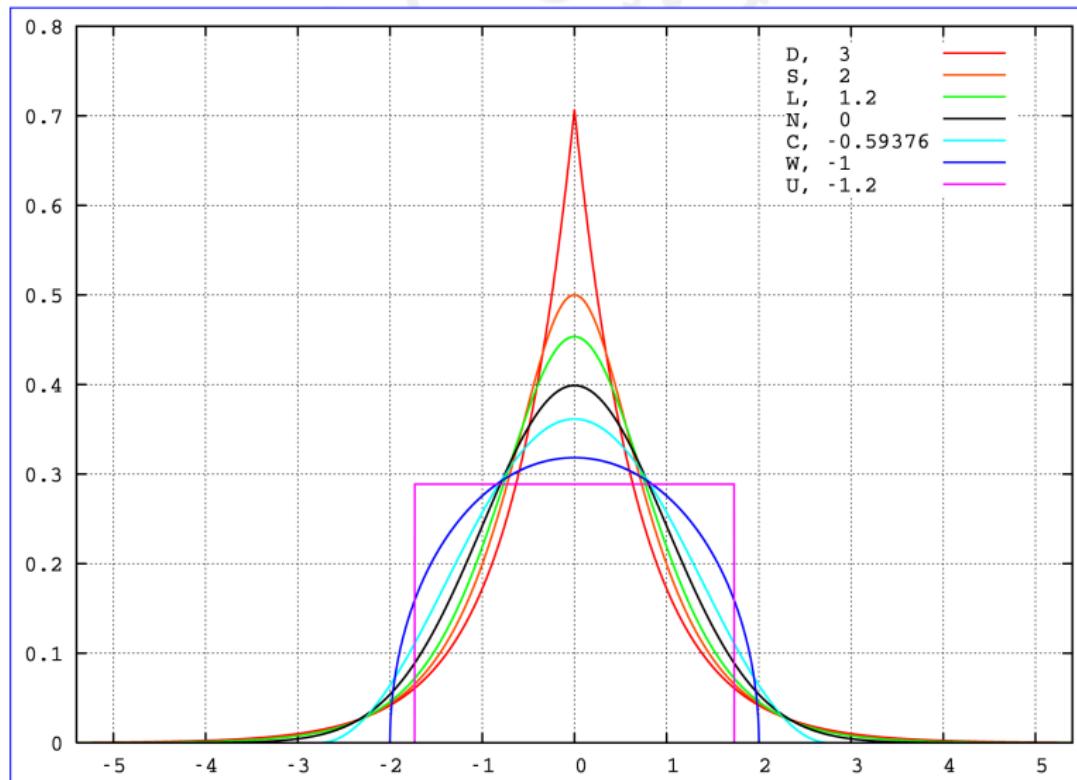
- measure of the peakedness and tail heaviness
- Kurtosis (Pearson, 1905)

$$k = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4$$

- Excess kurtosis

$$e = k - 3$$

- ▶ leptokurtic $e > 0$ (more acute peak around the mean and fatter tails)
- ▶ mesokurtic $e = 0$ (mass distribution as in normal distribution)
- ▶ platykurtic $e < 0$ (lower, wider peak around the mean and thinner tails)
- “poor” measure of peakedness and tail heaviness since interpretation is only valid if $g = 0$



Listing 5.11: example_kurtosis.R

```

1  data(Boston, package="MASS")
2  #
3  library("moments")
4  moments::kurtosis(Boston$medv) # kurtosis
5  #
6  library("e1071")
7  e1071::kurtosis(Boston$medv)   # excess

```

 moments::kurtosis(x, na.rm=FALSE)
 e1071::kurtosis(x, na.rm=F, type=3)

- type=1: e
- type=2: $((n + 1) \cdot e + 6) \cdot \frac{n-1}{(n-2)(n-3)}$ (SAS, SPSS)
- type=3: $(e + 3) \cdot (1 - \frac{1}{n})^2 - 3$ (MINITAB, BMDP)
- only type=2 is unbiased under normality

Higher moments

- Moments

raw $m(r) = \frac{1}{n} \sum_{i=1}^n x_i^r$

absolute $m_a(r) = \frac{1}{n} \sum_{i=1}^n |x_i|^r$

central $m_c(r) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r$

standardized $m_s(r) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^r$

- higher moments are difficult to interpret
- higher moments needs a lot of observations for reliable estimation



Listing 5.12: example_moments.R

```
1 data(Boston, package="MASS")
2 #
3 mean(Boston$medv)
4 var(Boston$medv)
5 #
6 library("moments")
7 all.moments(Boston$medv)
8 all.moments(Boston$medv, central=T)
9 all.moments(Boston$medv, order.max=4)
```

• ZU BERLIN •

```
R moments::all.moments(x, order.max=2, central=F, absolute=F,
                           na.rm=F)
```

Summarize continuous variables

- Tukey's five number summary



- ▶ Tukey used lower and upper hinges instead of quartiles
- Seven number summary
 - ▶ add 10% and 90% quantile to the five number summary
- Bowley's seven number summary
 - ▶ 2%, 9%, 25%, 50%, 75%, 91% and 98% quantile
 - ▶ under a normal distribution the distance between the quantiles is approximately equal

Bowley, Arthur Lyon (1952). *An Elementary Manual of Statistics*. 7th ed. London: Macdonald and Evans. 297 pp.

Tukey, John Wilder (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co. 688 pp. isbn: 978-0-201-07616-5.

R Listing 5.13: example_descriptive_summary.R

```
1 data(Boston, package="MASS")
2 library("psych")
3 #
4 summary(Boston$medv)
5 fivenum(Boston$medv)
6 describe(Boston$medv)
```

R summary(x)

⚠ Generic function - output depends on x!

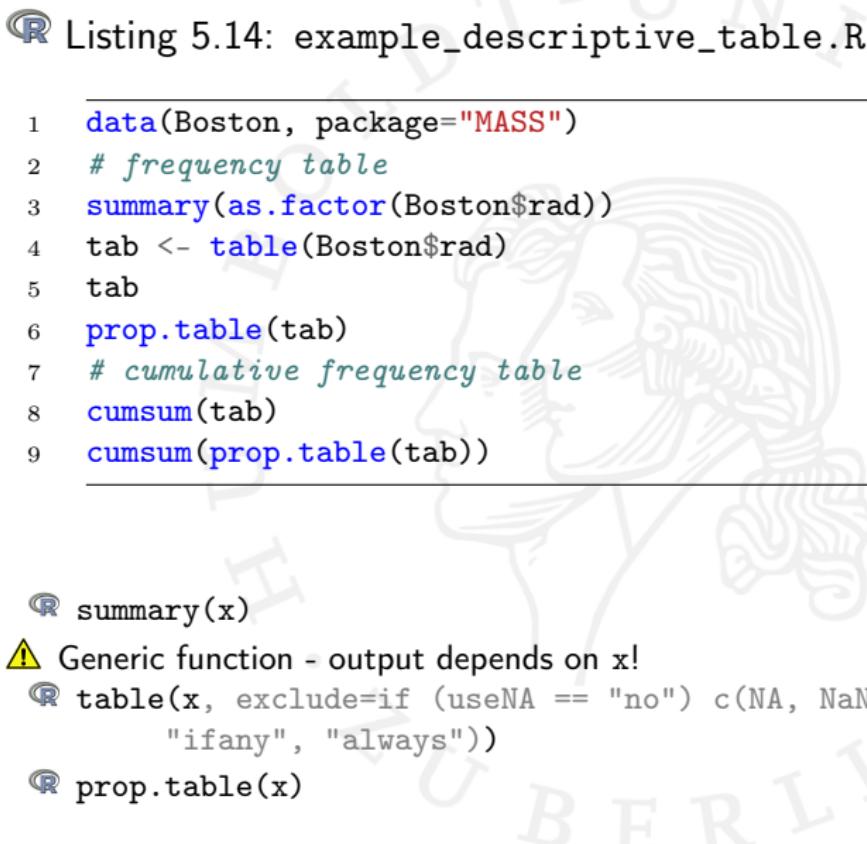
R fivenum(vec, na.rm=F)

⚠ uses the lower and upper hinges instead of quartiles

R psych::describe(x, na.rm=T)

Summarize categorical variables

- frequency table with
 - ▶ absolute or
 - ▶ relative frequencies
- cumulative frequency table with
 - ▶ absolute or
 - ▶ relative values
- cumulative frequency table require at least an ordinal scale

The background of the slide features a faint watermark of the HU Berlin logo, which is a circular emblem with a profile of a head and Latin text around it.

R Listing 5.14: example_descriptive_table.R

```
1 data(Boston, package="MASS")
2 # frequency table
3 summary(as.factor(Boston$rad))
4 tab <- table(Boston$rad)
5 tab
6 prop.table(tab)
7 # cumulative frequency table
8 cumsum(tab)
9 cumsum(prop.table(tab))
```

R summary(x)

⚠ Generic function - output depends on x!

R table(x, exclude = if (useNA == "no") c(NA, NaN), useNA=c("no",
"ifany", "always"))

R prop.table(x)

Empirical cumulative distribution function

- Theoretical definition

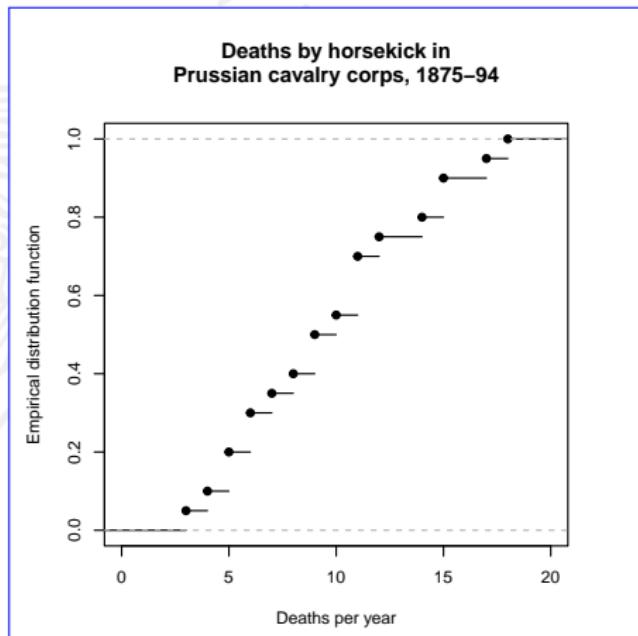
$$F(x) = P(X \leq x)$$

- Empirical definition

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

with

$$I(cond) = \begin{cases} 1 & \text{if } cond \text{ is true} \\ 0 & \text{if } cond \text{ is false} \end{cases}$$

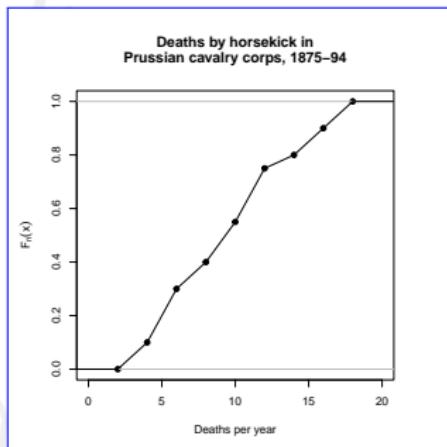


- Ecdf for grouped data

$$F_n(x) = \begin{cases} 0, & \text{if } x < x_j^l \\ F(x_j^l) + \frac{x - x_j^l}{x_j^u - x_j^l} f_j, & \text{if } x_j^l < x \leq x_j^u \\ 1, & \text{if } x_j^u \leq x \end{cases}$$

x_j^l, x_j^u lower and upper class border

f_j relative class frequency, $j = 1, \dots, J$



- The *strong law of large numbers* and the *Glivenko–Cantelli theorem* ensures (uniform) convergence:

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \stackrel{\text{a.s.}}{\equiv} 0$$



Listing 5.15: example_ecdf.R

```
1 data(Boston, package="MASS")
2 ecdf.medv <- ecdf(Boston$medv)
3 # print
4 ecdf.medv
5 # summary
6 summary(ecdf.medv)
7 # plot
8 plot(ecdf.medv)
```

R ecdf(x)

R Hmisc::Ecdf(x, what="F")

Distribution

November 3, 2022

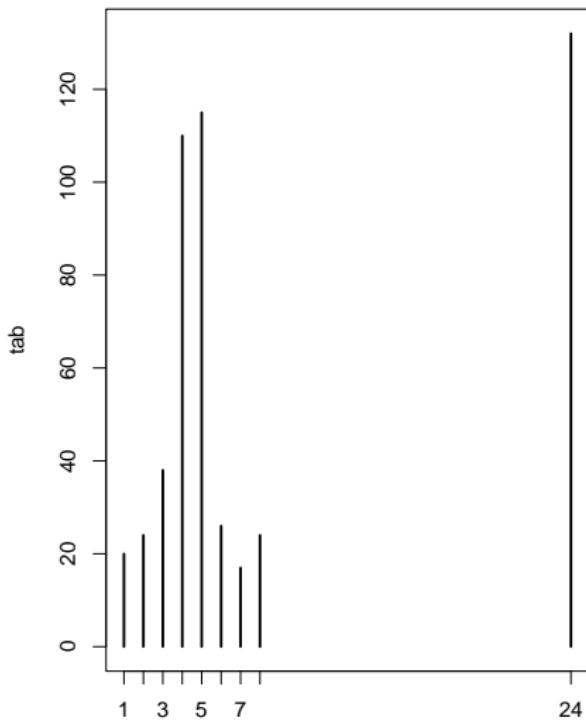
- Bar plot • Pie chart • Strip plot • Stem-and-leaf plot • Histogram •
- Boxplot • Violin plot • Average shifted histogram • Kernel density estimator • Quantile quantile plot • Probability probability plot • χ^2 goodness-of-fit test • Kolmogorov Smirnow test • Cramer-von Mises test •
- Anderson-Darling test • Jarque-Bera test • Shapiro-Wilk test

Bar plot

- Playfair (1786)
- Visualization: distribution
 - ▶ Variables: at least one categorical variable
 - ▶ Observations: unlimited

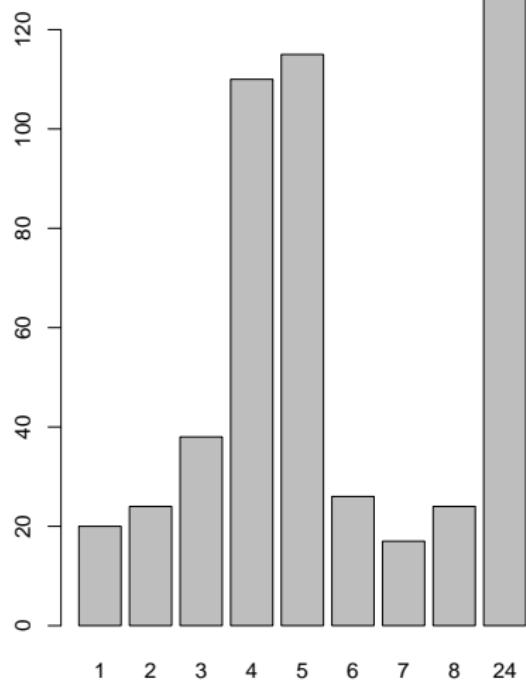
Playfair, William (1786). "Commercial and political atlas: Representing, by copper-plate charts, the progress of the commerce, revenues, expenditure, and debts of England, during the whole of the eighteenth century". In: London: Corry.

plot



metric discrete variable

barplot



nominal or ordinal variable

② Listing 6.1: example_barchart_plot.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$rad)
3 # bar chart as needles (should not be used!)
4 plot(tab)
```

② Listing 6.2: example_barplot.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$rad)
3 # standard barchart with own labels
4 at<-barplot(tab, axes=F)
5 axis(1, at=at, labels=names(tab))
```

② plot(table)

② barplot(table)

 Listing 6.3: example_barchart_lattice.R

```
1 library("MASS") # for Boston Housing da
2 library("lattice")
3 tab2 <- table(Boston$rad, Boston$chas)
4 ldat <- data.frame(rad=rep(rownames(tab2), 2),
5 count=as.vector(tab2),
6 chas=c(rep(colnames(tab2)[1],9),
7 rep(colnames(tab2)[2],9)))
8 barchart(count~rad|chas, data=ldat)
```

 Listing 6.4: example_barchart_ggplot.R

```
1 library("MASS") # for Boston Housing da
2 tab2 <- table(Boston$rad, Boston$chas)
3 library("ggplot2")
4 ggplot(Boston, aes(x=factor(rad))) + geom_bar()
```

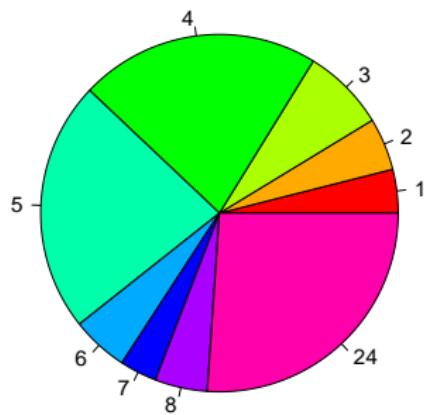
 lattice::barchart(formula, data)

Pie chart

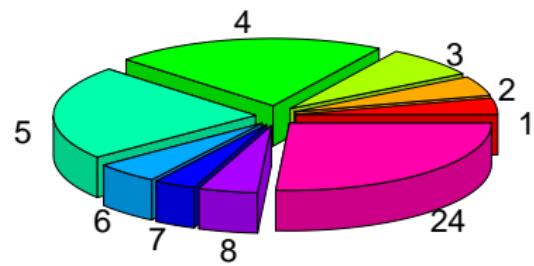
- Playfair (1801)
- Visualization: distribution
 - ▶ Variables: at least one categorical variable
 - ▶ Observations: unlimited
- Problem: comparison of pies

Playfair, W. (1801). *The statistical breviary; shewing, on a principle entirely new, the resources for every state and kingdom in Europe; illustrated with stained copper-plate charts, representing the physical powers of each distinct nation with ease and perspicuity.* London: T. Bensley, J. Wallis.

pie



pie3D



 Listing 6.5: example_piechart_graphics.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$rad)
3 pie(tab)
```

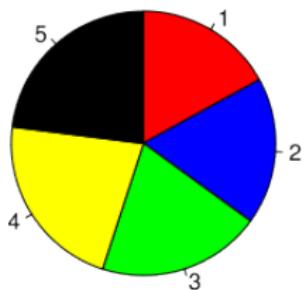
 Listing 6.6: example_piechart_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$rad)
3 library("ggplot2")
4 pie <- ggplot(Boston, aes(x = factor(1), fill = factor(rad)))
5 pie + geom_bar(width = 1) + coord_polar(theta = "y")
```

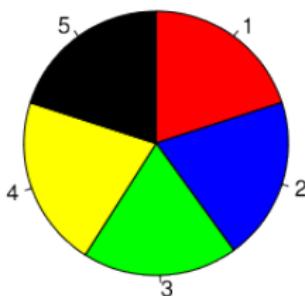
② pie(table, labels=names(x), col=NULL)

② plotrix:::pie3D(table, radius=1, height=0.1, theta=pi/6,
start=0, col=NULL, labels=NULL, labelpos=NULL,
sector.order=NULL, explode=0)

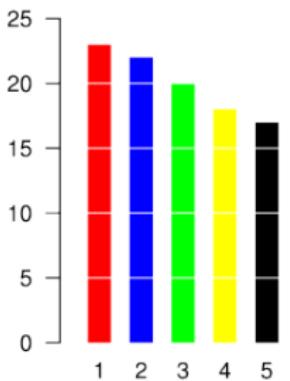
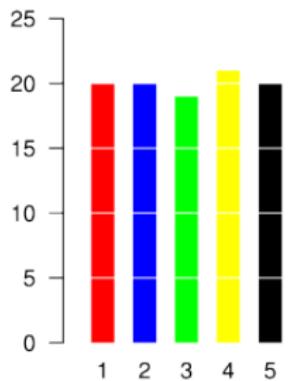
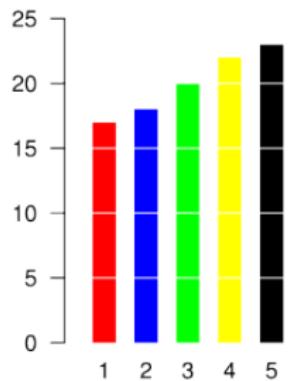
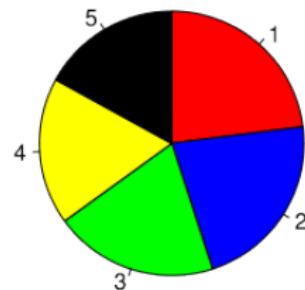
A



B



C



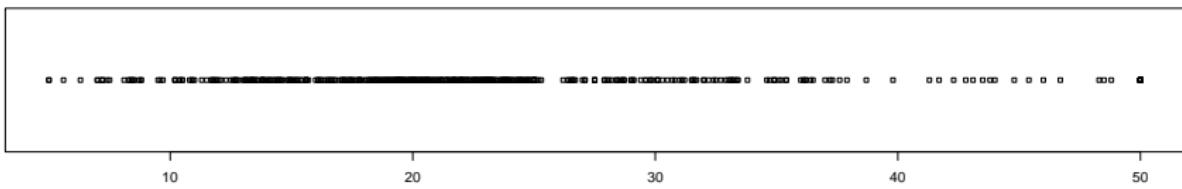
Strip plot

- van Langren (1644), Tukey, Tukey (1990)
- Visualization: observation values
 - ▶ Variables: at least one variable
 - ▶ Observations: medium
- Other names:
 - ▶ one dimensional scatterplot
 - ▶ dotplot
- jittering: move a randomly a point in y direction

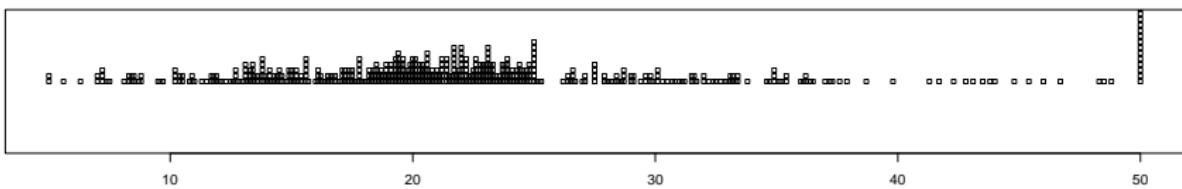
Langren, M.F. van (1644). *La verdadera longitud por mar y tierra*. Antwerp.

Tukey, John and Tukey, Paul (1990). *Strips displaying empirical distributions: I. textured dot strips*. Tech. rep. Bellcore Technical Memorandum.

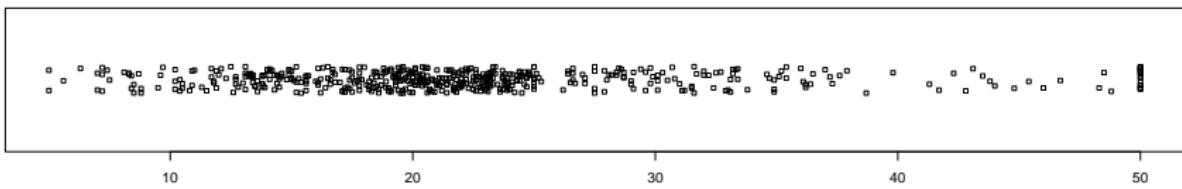
overplot



stack



jitter



R Listing 6.7: example_stripchart_graphics.R

```
1 library("MASS") # for Boston Housing data
2 stripchart(Boston$medv, method="j")
```

R Listing 6.8: example_stripchart_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 stripplot(~medv, data=Boston, jitter.data=T)
```

R Listing 6.9: example_stripchart_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 d <- ggplot(Boston, aes(x=medv, y=0))
4 d + geom_point(position=position_jitter(height=0.1))
```

```
R stripchart(x, method="overplot", jitter=0.1, vertical=F)
```

Stem-and-leaf plot

- Stem-and-leaf-plot: Tukey (1972)
- Visualization: distribution
 - ▶ Variables: one or two (groupable) variable
 - ▶ Observations: unlimited
- Construction
 - ▶ create classes based on (significant) digits
 - ▶ fixed stemwidth: 10^d
 - ▶ for leaves use next digit
 - ▶ show extreme values

Tukey, J.W. (1972). "Some Graphic and Semigraphic Displays". In: *Statistical Papers in Honor of George W. Snedecor*. Ed. by T.A. Bancroft. Ames, IA: Iowa State University Press, pp. 293–316. Presented at the Annual Meeting of the American Statistical Association, August 1969.

The decimal point is at the |

4 | 006
6 | 30022245
8 | 1334455788567
10 | 2224455899035778899
12 | 0135677780111233344455668888899
14 | 0111233445556689990001222344666667
16 | 01112234556677880111222344455567888889
18 | 0122233444555667778899900111122333344444455556666677888999
20 | 000001111223333444455566666778889900011222244444556677777788999
22 | 000000012222334455566667788899900001111112222333344566777788889
24 | 001112333444455566777888800000000123
26 | 24456667011555599
28 | 0124456777011466889
30 | 111357801255667
32 | 0024579011223448
34 | 679991244
36 | 01224502369
38 | 78
40 | 37
42 | 38158
44 | 084
46 | 07
48 | 358
50 | 000000000000000000

④ Listing 6.10: example_stem.R

```
1 library("MASS") # for Boston Housing data
2 stem(Boston$medv)
3 stem(Boston$medv, scale=0.5)
4 #
5 library("aplypack")
6 stem.leaf(Boston$medv)
```

④ stem(x, scale=1)

④ aplpack::stem.leaf(x, unit, m)

④ aplpack::stem.leaf.backback(x, y, unit, m)

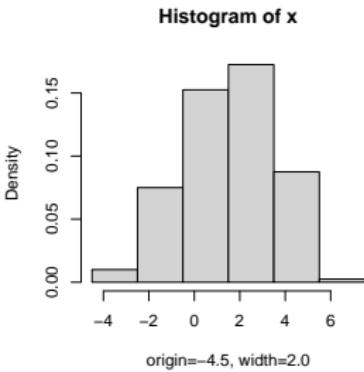
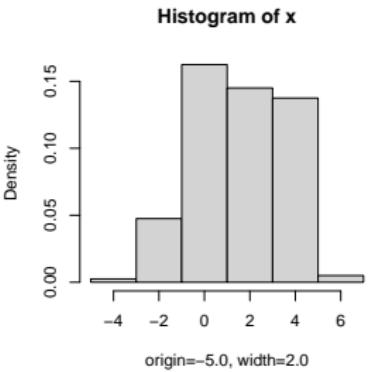
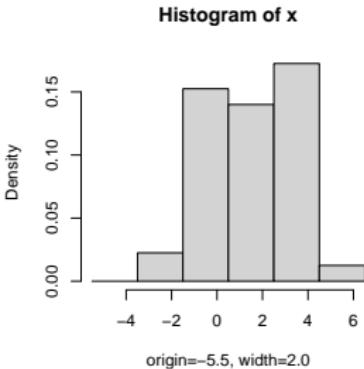
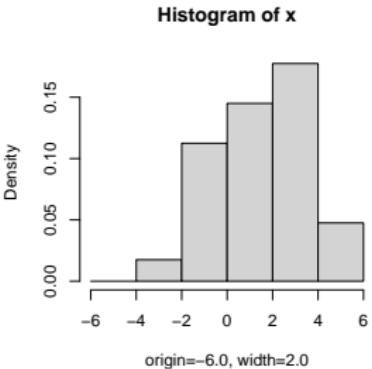
→ aplpack package contains an implementation of Tukey's original stem-and-leaf plot

Histogram

- Guerry (1833), Pearson (1895)
- Visualization: distribution
 - ▶ Variables: at least one continuous (groupable) variable
 - ▶ Observations: unlimited
- Problems:
 - ▶ choice of origin
 - ▶ choice of binwidth h or number of bins K
 - ▶ unequal binwidths
- Area of a bin corresponds to the frequency
- Corresponds to a rough estimate of the true density
 - ▶ add observations

Guerry, A.-M. et al. (1833). *Essai sur la Statistique Morale de la France*. Paris: Crochard.

Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". In: *Philosophical Transactions of the Royal Society of London.(A.)* 186, pp. 343–414.



Rule-of-thumbs for the number of classes K or binwidth h

$K = 1 + \log_2(n)$ approximates a normal distribution with a binomial distribution (Sturges, 1926)

$h = \frac{3,49s}{\sqrt[3]{n}}$ minimizes (approx.) integrated squared error $\int (\hat{f}(x) - f(x))^2 dx \Rightarrow h = \sqrt[3]{C(f')/n}$ with f as a normal density (Scott, 1979)

$h = \frac{2*IQR}{\sqrt[3]{n}}$ robustify Scott's rule (Freedman and Diaconis, 1981)

Sturges, H.A. (1926). "The choice of a class interval". In: *Journal of the American Statistical Association* 21.153, pp. 65–66.

Scott, D.W. (1979). "On optimal and data-based histograms". In: *Biometrika* 66.3, pp. 605–610.

Freedman, David and Diaconis, Persi (1981). "On the histogram as a density estimator: L 2 theory". In: *Probability theory and related fields* 57.4, pp. 453–476.

R Listing 6.11: example_histogram_graphics.R

```
1 library("MASS") # for Boston Housing data
2 # histogram + observations
3 hist(Boston$medv)
4 rug(Boston$medv)
```

- R hist(x, breaks="Sturges", freq=NULL, right=T)
- R nclass.Sturges(x)
- R nclass.scott(x)
- R nclass.FD(x)
- R rug(x, side=1)



Listing 6.12: example_histogram_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 histogram(~medv, data=Boston)
```



Listing 6.13: example_histogram_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 ggplot(Boston, aes(x=medv)) + geom_histogram(binwidth=2)
```

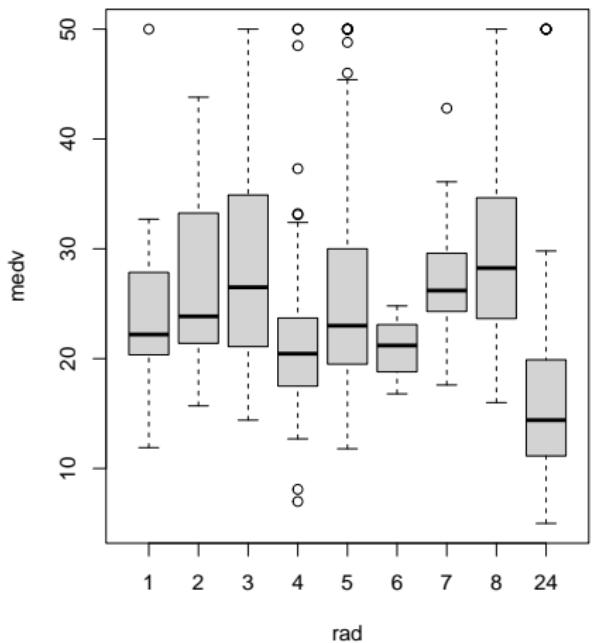
lattice::histogram(formula, data, type, nint, endpoints, breaks,
na.rm)

Boxplot

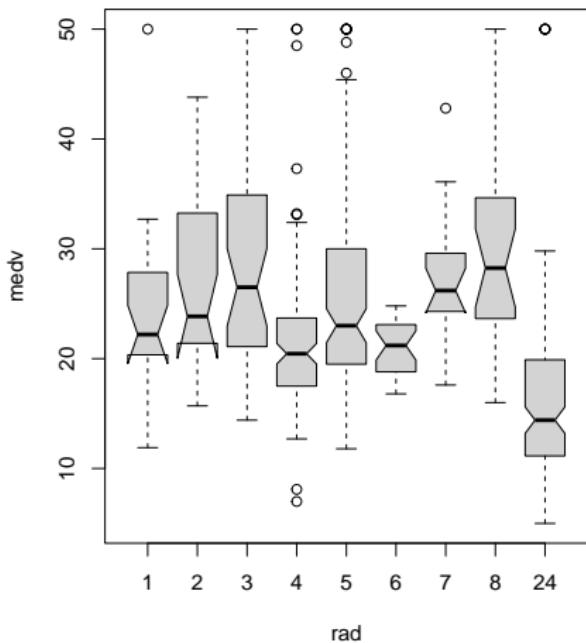
- Tukey (1972)
- Variables: at least one continuous variable
- Observations: unlimited
- Visualization: extreme values and parameters of distribution
- Notches indicate confidence intervals for median
- Problems:
 - ▶ shows only parameters
 - ▶ extreme values are not necessarily outliers

Tukey, J.W. (1972). "Some Graphic and Semigraphic Displays". In: *Statistical Papers in Honor of George W. Snedecor*. Ed. by T.A. Bancroft. Ames, IA: Iowa State University Press, pp. 293–316. Presented at the Annual Meeting of the American Statistical Association, August 1969.

Boxplot



Notched boxplot



 Listing 6.14: example_boxplot_graphics.R

```
1 library("MASS") # for Boston Housing data
2 boxplot(medv~rad, data=Boston)
```

 Listing 6.15: example_boxplot_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 bwplot(~medv|rad, data=Boston)
```

 Listing 6.16: example_boxplot_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 ggplot(Boston, aes(x=factor(rad), y=medv)) + geom_boxplot()
```

☞ `boxplot(x, range=1.5, width=NULL, varwidth=F, notch=F,
horizontal=F)`

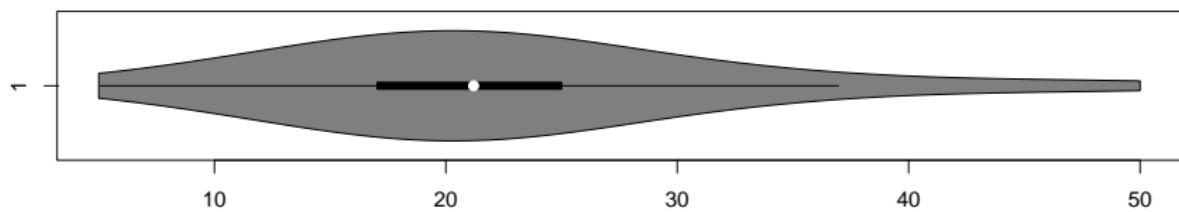
☞ `lattice::bwplot(formula, data)`

Violin plot

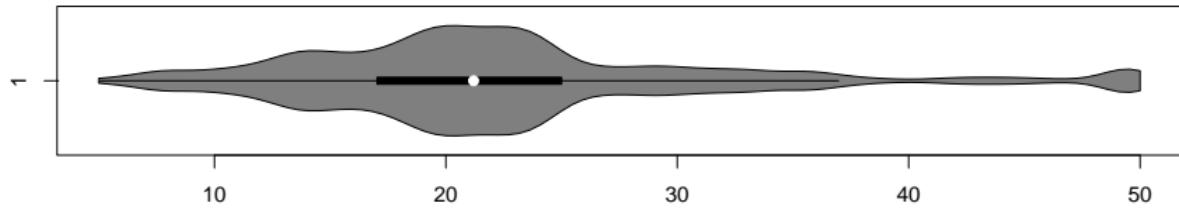
- Hintze, Nelson (1998)
- Variables: at least one continuous variable
- Observations: unlimited
- Visualization: distribution and parameters
- Construction:
 - ▶ marker for the median of the data
 - ▶ box indicating the interquartile range
- Problems:
 - ▶ no extreme values

Hintze, Jerry L and Nelson, Ray D (1998). "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2, pp. 181–184.

medv (default h)



medv (h=1)





Listing 6.17: example_violin.R

```
1 id <- function(x) { return(x); }
2 #
3 library("MASS") # for Boston Housing data
4 library("vioplot")
5 args <- tapply(Boston$medv, Boston$rad, FUN=id)
6 args$names <- names(args)
7 names(args)[1] <- 'x'
8 do.call("vioplot", args)
9 title(main="Violin plot", xlab="rad", ylab="medv")
```

⌚ vioplot::vioplot(x, range=1.5, h, horizontal=F)

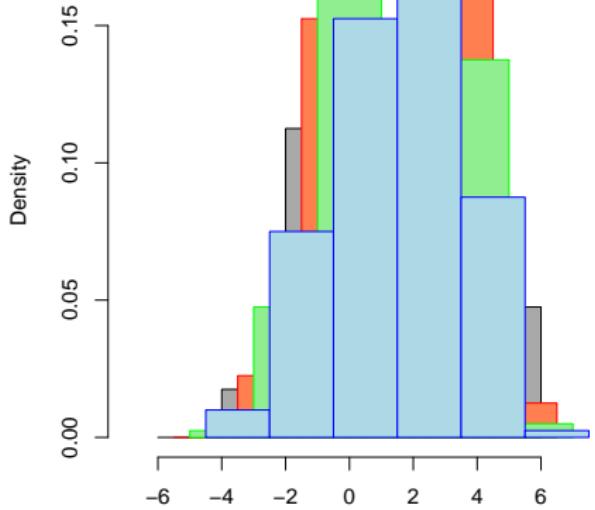
Average shifted histogram

- Visualization: distribution
 - ▶ Variables: at least one groupable variable
 - ▶ Observations: unlimited
- Construction:
 - ▶ create histograms with shifted origins
 - ▶ average the (weighted) histogram heights at each bin
- Problems:
 - ▶ choice of binwidth
 - ▶ unequal binwidths

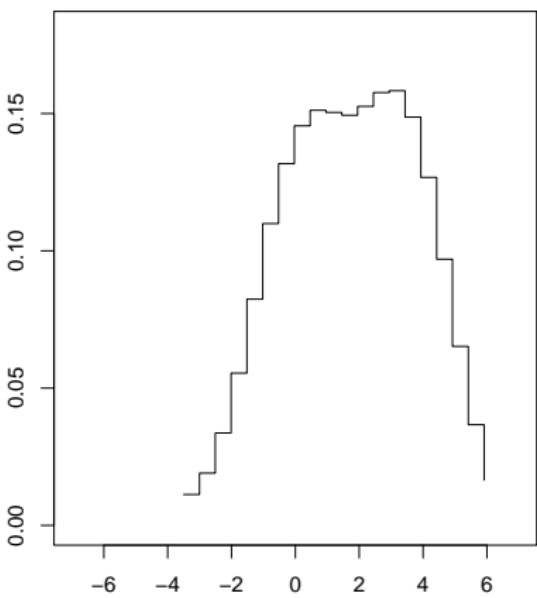
Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*.
Vol. 275. John Wiley & Sons.

Histogram of x



Average shifted histogram



 Listing 6.18: example_ash.R

```
1 library("MASS") # for Boston Housing data
2 library("ash")
3 ash <- ash1(bin1(Boston$medv))
4 plot(ash, type="s")
```

-  ash::ash1(bins, m=5)
-  compute bins with ash::bin1
-  ash::ash2(bins, m=c(5,5))
-  compute bins with ash::bin2

Kernel density estimator

- Rosenblatt (1956), Parzen (1962)
- Visualization: distribution
 - ▶ Variables: at least one continuous variable
 - ▶ Observations: unlimited
- Estimate of the true density function $f(x)$

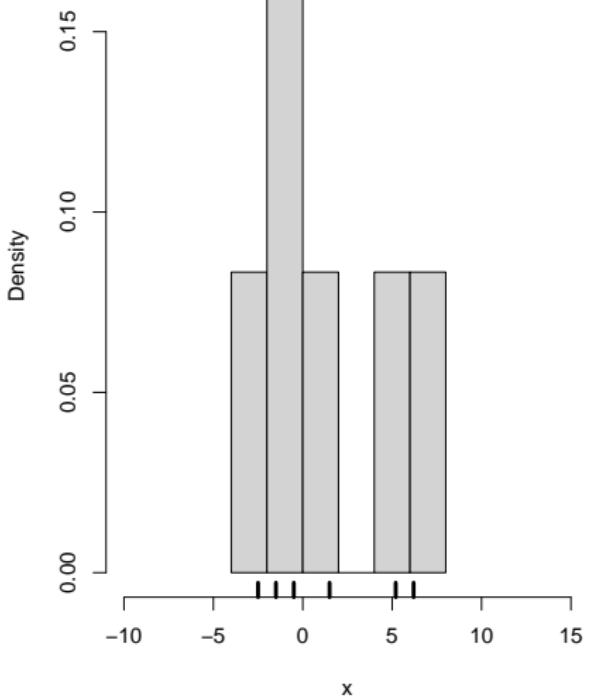
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

- ▶ kernel function $K(x)$
- ▶ bandwidth h (equivalent to binwidth in histograms)

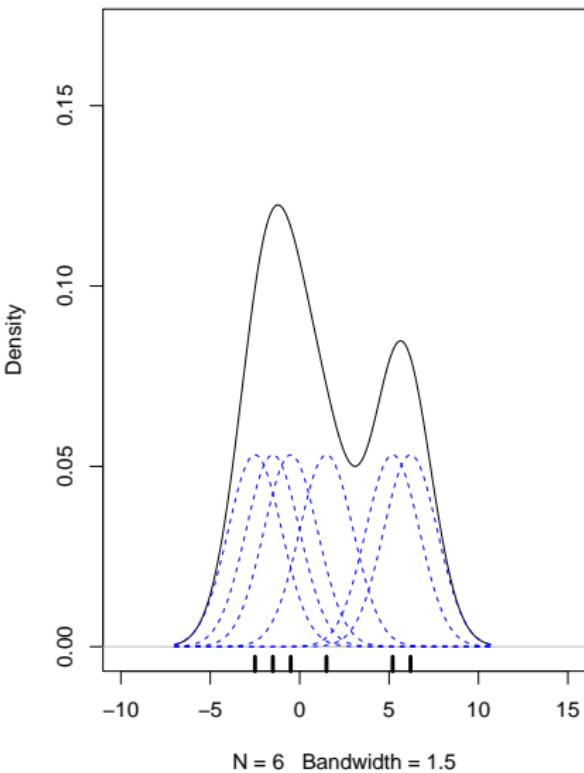
Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function". In: *The Annals of Mathematical Statistics* 27.3, pp. 832–837.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076.

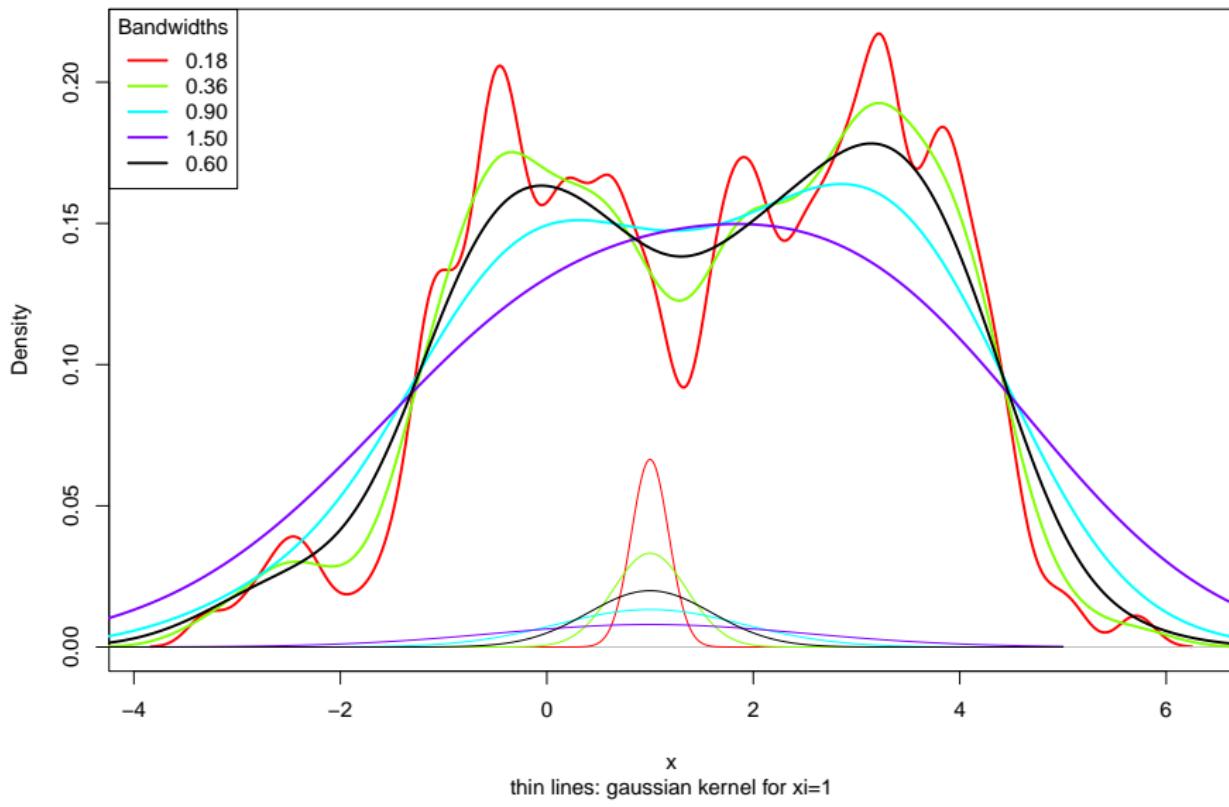
Histogram of x



`density.default(x = x, bw = 1.5)`



Kernel density estimates



④ Listing 6.19: example_kde_graphics.R

```
1 library("MASS") # for Boston Housing data
2 fhat <- density(Boston$medv)
3 plot(fhat)
```

④ Listing 6.20: example_kde_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 densityplot(~medv, data=Boston)
```

④ Listing 6.21: example_kde_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 ggplot(Boston, aes(x=medv)) + geom_density()
```

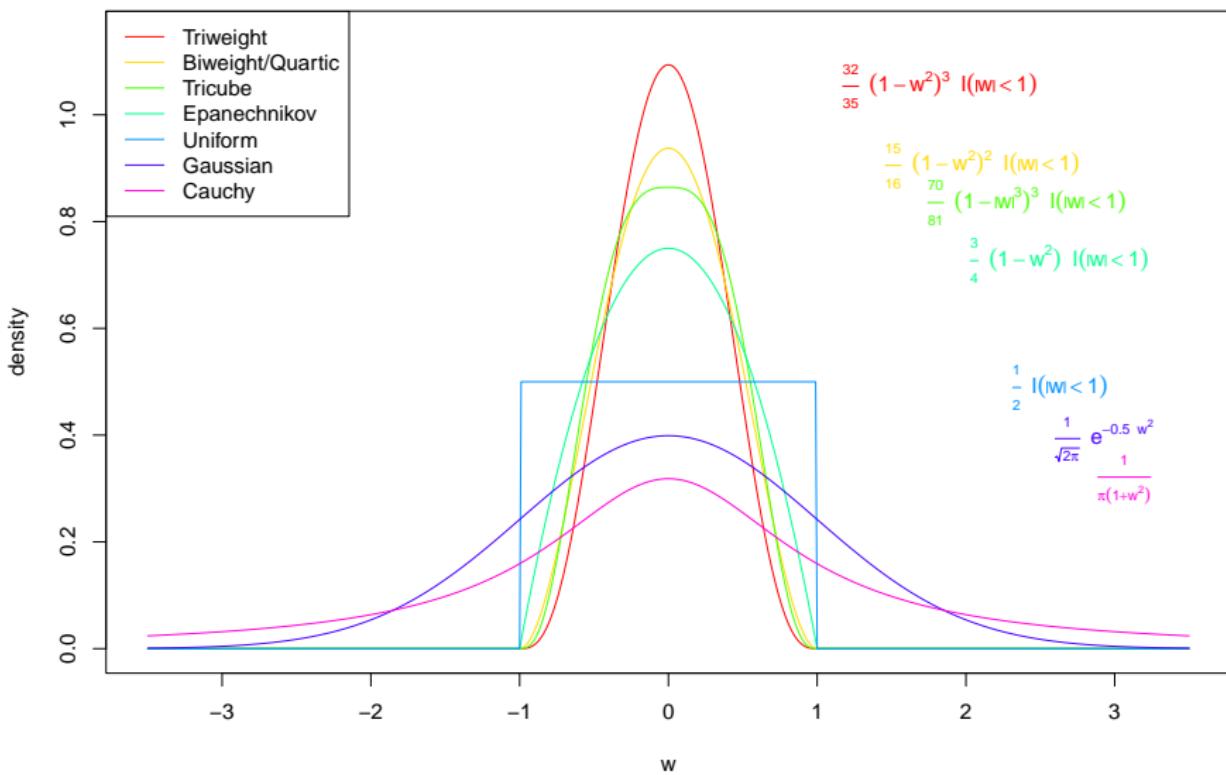
④ `density(x, bw="nrd0", n=512, kernel=c("gaussian",...))`

④ `lattice::densityplot(x, bw="nrd0", n=512, kernel=c("gaussian",...))`

Choice of kernel function K

- it must hold: $\int K(w)dw = 1$ and $K(w) \geq 0 \Rightarrow \hat{f}_h(x)$ is density
- kernel choice can be compensated with bandwidth adjustment
- kernel density estimate inherits smoothness properties of kernel
- computation time can be reduced
 - ▶ by using kernels with finite support
 - ▶ by pre-binning the data

Kernels



- for each x the same bandwidth
 - ▶ Silverman rule-of-thumb: $h_{rot} = \frac{1.06s}{\sqrt[5]{n}}$
 - ▶ nrd0: $h_{nrd0} = \frac{0.9 \min(s, IQR/1.34)}{\sqrt[5]{n}}$ (R, Stata)
 - ▶ Sheather-Jones plug-in method
 - ★ bandwidth choice depending on first and second derivative of true density
 - ★ estimate first and second derivative with rule-of-thumb bandwidth
- for each x a different bandwidth (SPSS)
 - ▶ fixed window: the window contains a fixed percentage of the observations
 - ▶ k nearest neighbour: the window contains the k nearest neighbours

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Sheather, Simon J. and Jones, M. Chris (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, pp. 683–690.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Vol. 275. John Wiley & Sons.

 Listing 6.22: example_kdeh_graphics.R

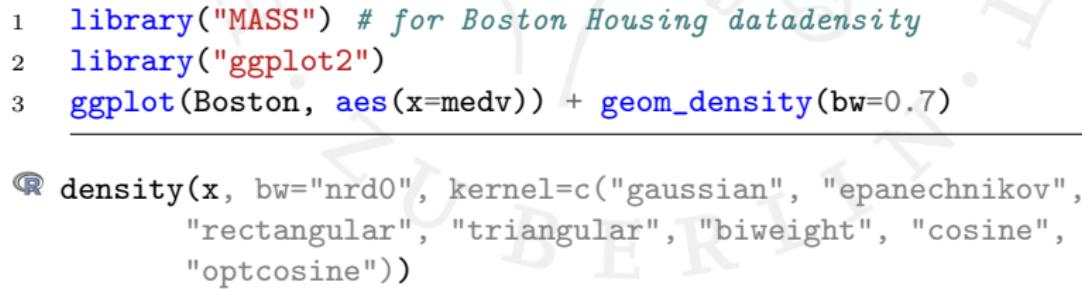
```
1 library("MASS") # for Boston Housing data
2 fhat <- density(Boston$medv, bw=0.7)
3 plot(fhat)
```

 Listing 6.23: example_kdeh_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 densityplot(~medv, data=Boston, bw=0.7)
```

 Listing 6.24: example_kdeh_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 ggplot(Boston, aes(x=medv)) + geom_density(bw=0.7)



A density plot showing the distribution of median value (medv) for Boston Housing data. The x-axis ranges from approximately 5 to 50, and the y-axis ranges from 0 to 0.04. The distribution is roughly bell-shaped, centered around a medv of about 18.

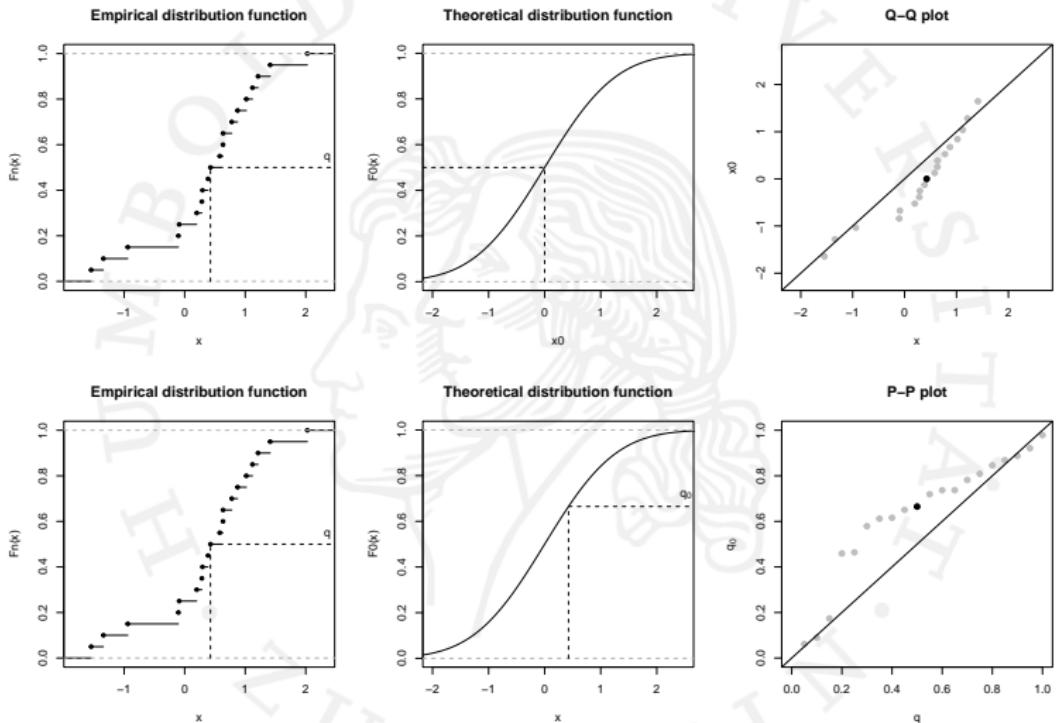

```

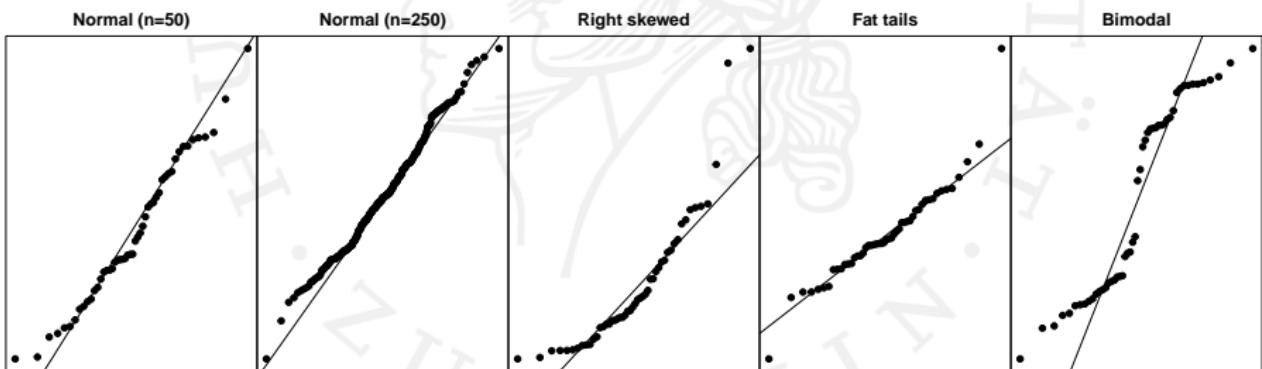
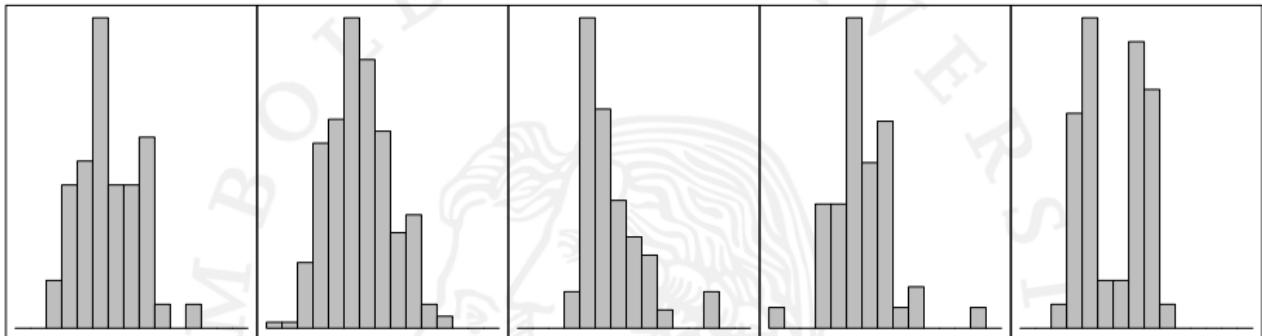
Quantile quantile plot

- Chambers, Cleveland, Tukey (1983)
- Visualization: compare distributions
 - ▶ Variables: at least one (continuous) variable
 - ▶ Observations: unlimited
- Aim: Follows a variable a given distribution?
- Plot empirical quantiles vs. theoretical quantiles in scatterplot
- Look for systematic deviations from the 45° line

Chambers, J.M., Cleveland, W.S., and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks.

- For each observation x_i estimate a quantile $\hat{q}_i = \hat{F}(x_i)$
- Compute for a given distribution F_0 the value $F_0^{-1}(\hat{q}_i)$
 - ▶ parameters for the given distribution must be estimated from the sample
- Plot $(x_i, F_0^{-1}(\hat{F}(x_i)))$
- If \hat{F} and F_0 are identical then all plotted points lie on (x_i, x_i) , the 45° line
- Note:
 - ▶ usually plots are scaled by minimum and maximum x- and y-coordinates of the plotted objects, therefore the 45° line may not look like a 45° line
 - ▶ expect at the left and right side deviations from the 45° line





 Listing 6.25: example_qqplot_graphics.R

```
1 library("MASS") # for Boston Housing data
2 qqnorm(Boston$medv, pch=19, cex=0.5)
3 qqline(Boston$medv)
```

- ¶ qqnorm(x)
- ¶ qqplot(x, y)
- ¶ qqline(x)
- ¶ car::qq.plot(x, distribution="norm")
- ¶ qualityTools::qqPlot(x, distribution="norm", confbounds=T)
- ¶ PerformanceAnalytics::chart.QQPlot(x, distribution="norm")

 Listing 6.26: example_qqplot_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 qqmath(~medv, data=Boston, panel = function(x, ...) {
4   panel.qqmathline(x, ...)
5   panel.qqmath(x, ...)
6 })
```

 Listing 6.27: example_qqplot_ggplot.R

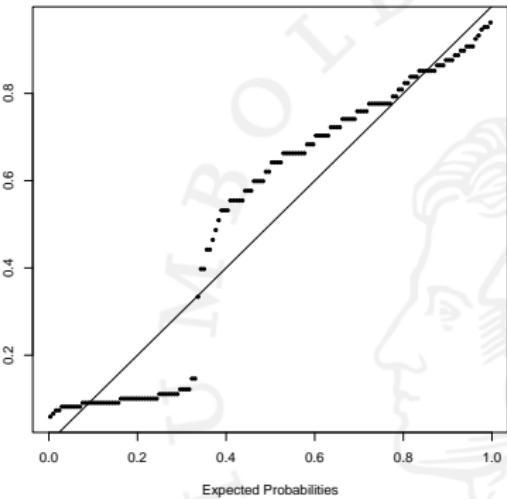
```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 d <- ggplot(Boston, aes(sample=medv)) + stat_qq()
4 d + geom_abline(intercept = mean(Boston$medv),
5 slope = sd(Boston$medv))
```

Probability probability plot

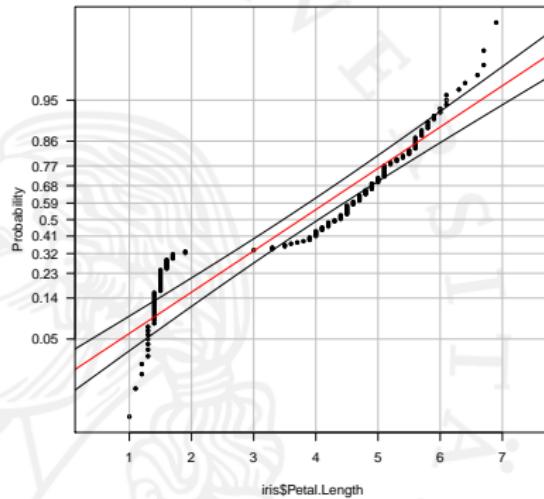
- Q-Q plot
 - ▶ shows sample quantiles vs. theoretical quantiles
 - ▶ plot area: minimum and maximum of observations and quantiles of theoretical distribution
 - ▶ are preferred about P-P plots
- P-P plot
 - ▶ shows the empirical vs. the theoretical distribution function
 - ▶ plot $(\hat{F}(x_i), F_0(x_i))$
 - ▶ plot area: $[0; 1] \times [0; 1]$
 - ▶ better overview about the whole distribution

Normal P-P Plot iris\$Petal.Length

Observed Probabilities



Normal P-P Plot iris\$Petal.Length



```
R p... (x)  
R ppoints(n)  
R qualityTools::ppPlot(x, distribution="norm", confbounds=T)
```

χ^2 goodness-of-fit test

Assumption(s): X_i categorical or grouped, $e_i > 5$ for all cells

Hypotheses: $H_0 : F(X) = F_0(X)$ vs. $H_1 : F(X) \neq F_0(X)$

Test statistics: $V = n \sum_{i=1}^I \frac{(f_i - p_i)^2}{p_i} = \sum_{i=1}^I \frac{(h_i - e_i)^2}{e_i} \approx \chi^2_{I-1-k}$

Reject H_0 : $|v| > \chi^2_{I-1-k; 1-\alpha}$
 h_i observed observations, e_i expected observations
 f_i observed frequency, p_i expected frequency
 k number of distribution parameters estimated from
the sample

- other approximation conditions are possible
 - ▶ all $e_i \geq 5$ or $e_i > 10$
 - ▶ in at least 80% of cells it holds: $e_i \geq 5$
 - ▶ for 2×2 tables: $e_i > 10$
- for continuous distribution the Kolmogorov-Smirnov should be used (more information!)

Pearson, Karl (July 1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *Philosophical Magazine Series 5* 50.302, pp. 157–175. issn: 1941-5982, 1941-5990. doi: 10.1080/14786440009463897. url: <http://www.tandfonline.com/doi/abs/10.1080/14786440009463897> (visited on 12/02/2015).

⌚ Listing 6.28: example_gof.R

```
1 data(Boston, package="MASS")
2 # test on uniform distribution
3 chisq.test(table(Boston$rad))
```

⌚ chisq.test(x, p=rep(1/length(x), length(x)), rescale.p=FALSE,
simulate.p.value=FALSE, B=2000)

Kolmogorov Smirnow test

Assumption(s): $F_0(X)$ is a continuous distribution, parameters known

Hypotheses: $H_0 : F(X) = F_0(X)$ vs. $H_1 : F(X) \neq F_0(X)$

Test statistics: $D_n = \sup |F_n(X) - F_0(X)|$

Reject H_0 : for $n \leq 40$ are critical values tabulated

for $n > 40$ use as critical values $d_\alpha = \sqrt{\frac{\log(2/\alpha)}{2n}}$

Effect size: D_n

- tends to be more sensitive in the center of the distribution than at the tails
- Modifications for unknown parameters, e.g. Lilliefors correction for normal distribution

- Kolmogorov, A. N. (1933). "Sulla Determinazione Empirica di una Legge di Distribuzione". In: *Giornale dell'Istituto Italiano degli Attuari* 4, pp. 83–91.
- Smirnoff, N. (1939). "Sur les écarts de la courbe de distribution empirique". In: *Rec. Math. N.S.* [Mat. Sbornik] 6(48), pp. 3–26.
- Lilliefors, Hubert W. (June 1967). "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". In: *Journal of the American Statistical Association* 62.318, pp. 399–402. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1967.10482916. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1967.10482916> (visited on 06/07/2016).

R Listing 6.29: example_ks.R

```
1 data(Boston, package="MASS")
2 # Kolmogorov-Smirnov test
3 ks.test(Boston$medv, "pnorm", mean=mean(Boston$medv),
4 sd=sd(Boston$medv))
5 # Lilliefors test
6 library("nortest")
7 lillie.test(Boston$medv)
```

- Kolmogorov Smirnow test

 R ks.test(x, pDist, ..., alternative=c("two.sided", "less",
 "greater"), exact=NULL)

- Lilliefors test for a normal distribution

 R fBasics::ksnormTest(x)

⚠ Tests if data follow a standard normal distribution!

 R nortest::lillie.test(x)

 R fBasics::lillieTest(x)

Cramer-von Mises test

Assumption(s): $F_0(X)$ is a continuous distribution, parameters known

Hypotheses: $H_0 : F(X) = F_0(X)$ vs. $H_1 : F(X) \neq F_0(X)$

Test statistics: $T = \frac{1}{12n} + \sum_{i=1}^n \left(\frac{2i-1}{2n} - F_0(X_{(i)}) \right)^2$

Reject H_0 : critical values are tabulated

- compare distances between the empirical cdf and theoretical cdf

$$\int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 dx$$

- the test statistics compares the cont. uniform distribution with $F_0(x_{(i)})$

- Cramér, Harald (Jan. 1928). "On the composition of elementary errors: First paper: Mathematical deductions". In: *Scandinavian Actuarial Journal* 1928.1, pp. 13–74. issn: 0346-1238, 1651-2030. doi: 10.1080/03461238.1928.10416862. url: <http://www.tandfonline.com/doi/abs/10.1080/03461238.1928.10416862> (visited on 06/07/2016).
- Mises, R. von (1931). *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und theoretischen Physik*. Leipzig-Wien: Verlag Franz Deuticke, pages. 574 pp.
- Anderson, T. W. (Sept. 1962). "On the Distribution of the Two-Sample Cramer-von Mises Criterion". In: *The Annals of Mathematical Statistics* 33.3, pp. 1148–1159. issn: 0003-4851. doi: 10.1214/aoms/1177704477. url: <http://projecteuclid.org/euclid.aoms/1177704477> (visited on 06/07/2016).

R Listing 6.30: example_cvm.R

```
1 data(Boston, package="MASS")
2 library("nortest")
3 cvm.test(Boston$medv)
```

R nortest::cvm.test(x)

⚠ Test if data follow a normal distribution!

Anderson-Darling test

Assumption(s): $F_0(X)$ is a continuous distribution
parameters known, $n \geq 8$

Hypotheses: $H_0 : F(X) = F_0(X)$ vs. $H_1 : F(X) \neq F_0(X)$

Test statistics: $A^2 = -n - S$

$$S = \sum_{i=1}^n \frac{2i-1}{n} (\log(F_0(X_{(i)})) + \log(1 - F_0(X_{(n+1-i)})))$$

Reject H_0 : critical values are tabulated

- compare distances between the empirical cdf and theoretical cdf

$$\int_{-\infty}^{+\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dx$$

- differences in the tails are upweighted

Anderson, T. W. and Darling, D. A. (June 1952). "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". In: *The Annals of Mathematical Statistics* 23.2, pp. 193–212. issn: 0003-4851. doi: 10.1214/aoms/1177729437. url: <http://projecteuclid.org/euclid.aoms/1177729437> (visited on 06/08/2016).

⌚ Listing 6.31: example_ad.R

```
1 data(Boston, package="MASS")
2 library("nortest")
3 ad.test(Boston$medv)
```

⌚ nortest::ad.test(x)

⚠ Tests if data follow a normal distribution!

Jarque-Bera test

Assumption(s): X_i continuous

Hypotheses: $H_0 : X$ normal distributed vs.

$H_1 : X$ not normal distributed

Test statistics: $V = \frac{n}{6} \left(Sk^2 + \frac{(Ku - 3)^2}{4} \right) \approx \chi^2_2$

$$Sk = \bar{X}_3/S^3 \text{ and } Ku = \bar{X}_4/S^4$$

$$\bar{X}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

- rejection of H_1 does not imply normality
- the robust version utilizes the mean absolute deviation from the median to estimate skewness and kurtosis

- Jarque, Carlos M. and Bera, Anil K. (Jan. 1980). "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". In: *Economics Letters* 6.3, pp. 255–259.
issn: 01651765. doi: 10.1016/0165-1765(80)90024-5. url:
<http://linkinghub.elsevier.com/retrieve/pii/0165176580900245> (visited on
06/08/2016).
- Gel, Yulia R. and Gastwirth, Joseph L. (Apr. 2008). "A robust modification of the Jarque–Bera test of normality". In: *Economics Letters* 99.1, pp. 30–32. issn: 01651765. doi:
10.1016/j.econlet.2007.05.022. url:
<http://linkinghub.elsevier.com/retrieve/pii/S0165176507001838> (visited on
06/08/2016).

 Listing 6.32: example_jb.R

```
1 library("DescTools")
2 data(Boston, package="MASS")
3 JarqueBeraTest(Boston$medv)
```

```
② DescTools:::JarqueBeraTest(x, robust=TRUE, method=c("chisq", "mc"),
                               N=0, na.rm=FALSE)
② lawstat:::rjb.test(x, option="JB", crit.values=c("chisq.approximation",
                                                 "empirical"), N=0)
② tseries:::jarque.bera.test(x)
② fBasics:::jarqueberaTest(x)
```

Shapiro-Wilk test

Assumption(s): X_i continuous

$$3 \leq n \leq 5000$$

Hypotheses: $H_0 : X$ normal distributed vs.

$H_1 : X$ not normal distributed

Test statistics:

$$W = \frac{B^2}{(n-1)S^2}$$

S^2 variance estimator of the observations

B^2 variance estimator under normality:

$b^2 = \hat{\sigma}^2$ is estimated from $x_{(i)} = \mu + \sigma^2 m_{(i)}$

with $m_{(i)} \approx \Phi^{-1} \left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}} \right)$ (Blom)

Reject H_0 : critical values are tabulated

Remark: Shapiro and Francia (1972) and Royston (1992) extended the test for $n \leq 5000$

- Shapiro, S. S. and Wilk, M. B. (Dec. 1, 1965). "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3, pp. 591–611. issn: 0006-3444, 1464-3510. doi: 10.1093/biomet/52.3-4.591. url: <http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/52.3-4.591> (visited on 06/08/2016).
- Shapiro, S. S. and Francia, R. S. (Mar. 1972). "An Approximate Analysis of Variance Test for Normality". In: *Journal of the American Statistical Association* 67.337, pp. 215–216. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1972.10481232. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481232> (visited on 06/08/2016).
- Royston, Patrick (Sept. 1992). "Approximating the Shapiro-Wilk W-test for non-normality". In: *Statistics and Computing* 2.3, pp. 117–119. issn: 0960-3174, 1573-1375. doi: 10.1007/BF01891203. url: <http://link.springer.com/10.1007/BF01891203> (visited on 06/08/2016).

⌚ Listing 6.33: example_shapiro.R

```
1 data(Boston, package="MASS")
2 shapiro.test(Boston$medv)
3 #
4 library("DescTools")
5 ShapiroFranciaTest(Boston$medv)
```

⌚ shapiro.test(x)
⌚ DescTools::ShapiroFranciaTest(x)
⌚ fBasics::shapiroTest(x)

Transformations

November 3, 2022

Simple transformations • Ladder of transformations • Box-Cox transformation • Yeo-Johnson transformation • Further transformations

Simple transformations

- Aims
 - ▶ reduce the range of a variable
 - ▶ reduce skewness towards a symmetric distribution
 - ▶ reduce heterogeneity of variances
- Assumptions
 - ▶ range large enough, rule-of-thumb: $\frac{\max_i x_i}{\min_i x_i} \geq 20$
 - ▶ interpretation of transformed values is possible
- a well known special transformation

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

- ▶ Standardization: $\bar{z} = 0, s_z = 1$

- Linear transformation

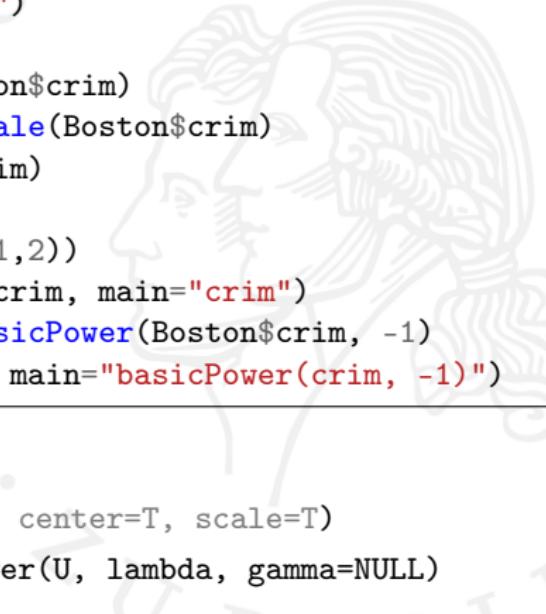
$$y_i = ax_i + b$$

- ▶ effect on coefficients: $\bar{y} = a\bar{x} + b$, $s_y = |b|s_x$, ...
- ▶ Standardization is a special linear transformation

- Power transformation

$$y_p(x) = \begin{cases} (x + c)^p & p \neq 0 \\ \log(x + c) & p = 0 \end{cases}$$

- ▶ $c > \min_i x_i$
- ▶ Note: if r small then it holds $\log(1 + r) \approx r$
- ▶ small changes in the natural log of a variable are directly interpretable as percentage changes $\log(x(1 + r)) \approx \log(x) + r$

A large, faint watermark of a classical bust of a man's head, facing slightly left, is centered behind the text.

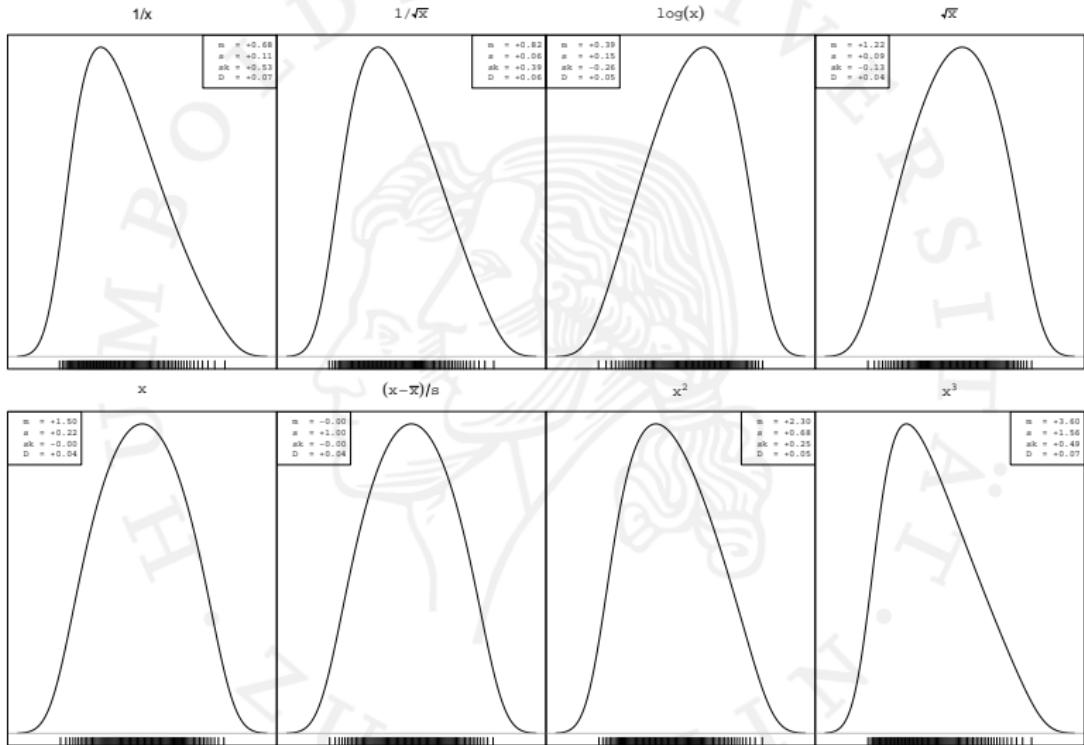
R Listing 7.1: example_scale_power.R

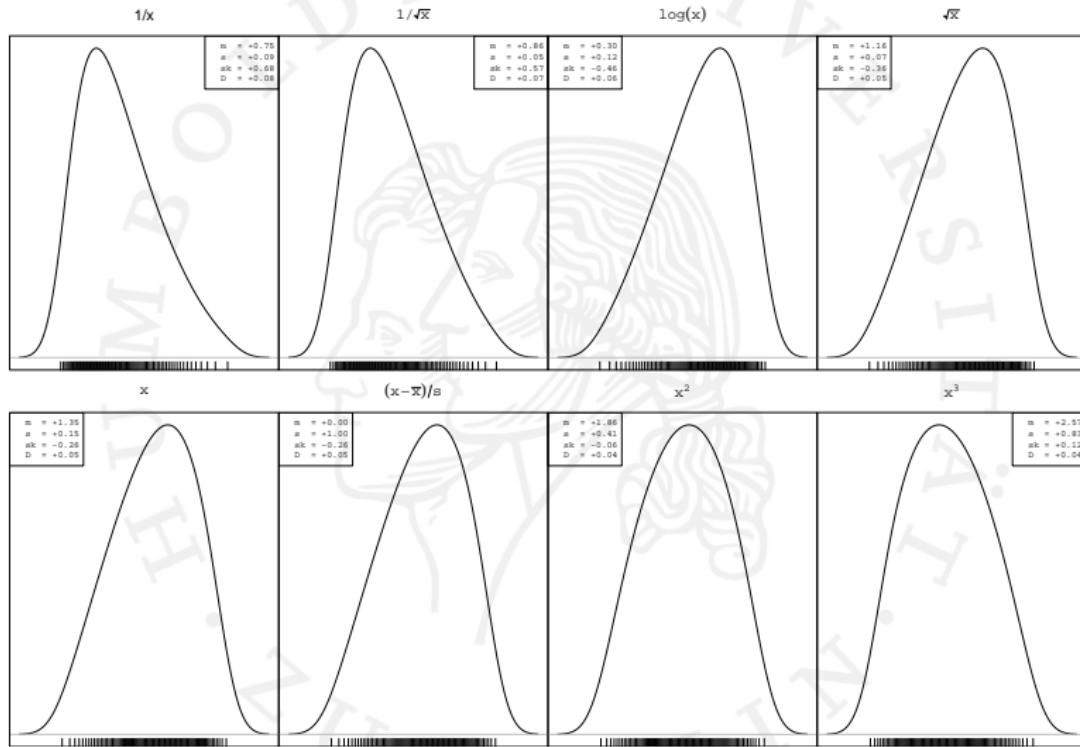
```
1 data(Boston, package="MASS")
2 library("car")
3 # scale
4 summary(Boston$crim)
5 sccrim <- scale(Boston$crim)
6 summary(sccrim)
7 # Power
8 par(mfrow=c(1,2))
9 hist(Boston$crim, main="crim")
10 spcrim <- basicPower(Boston$crim, -1)
11 hist(spcrim, main="basicPower(crim, -1)")
```

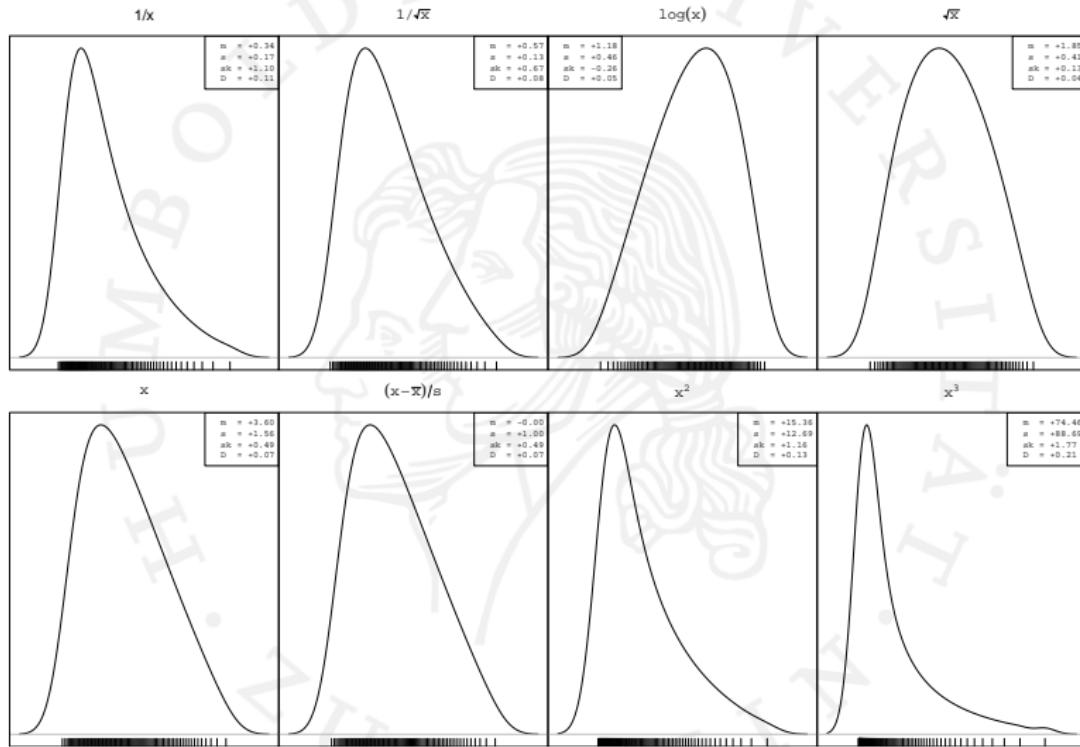
R scale(matrix, center=T, scale=T)
R car::basicPower(U, lambda, gamma=NULL)

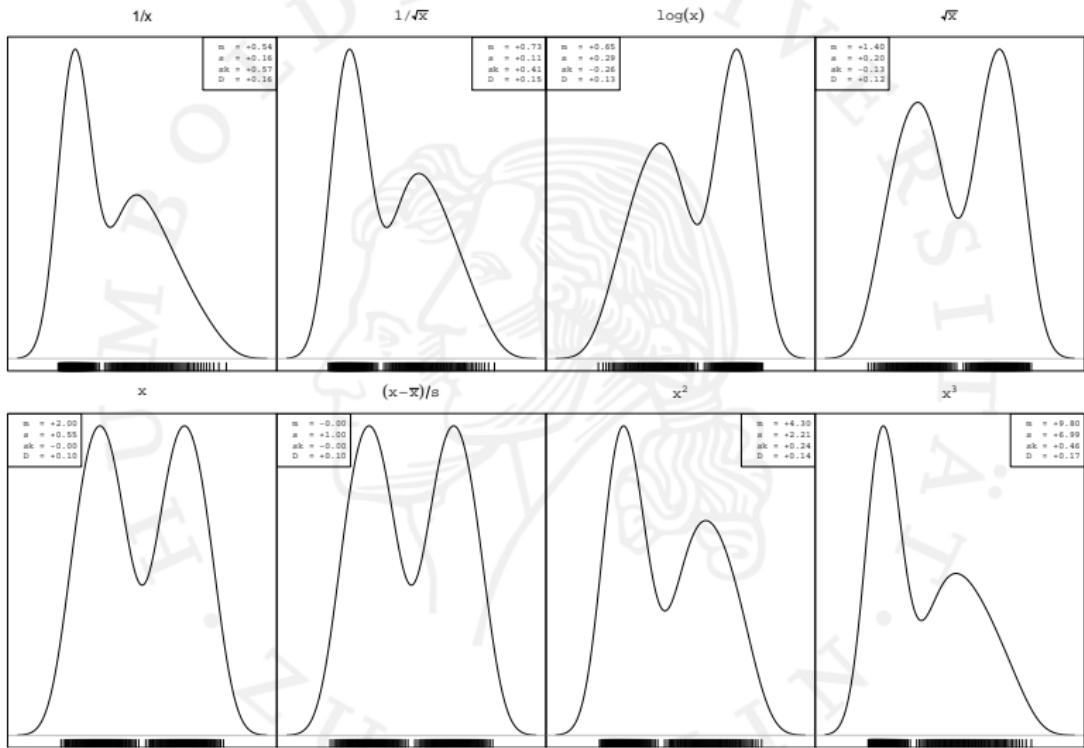
Ladder of transformations

p	... 3 ... 2 ... 1 ... 0.5 ... 0 ... -1 ... -2 ...
x^p	... x^3 ... x^2 ... x ... \sqrt{x} ... $\log(x)$... $\frac{1}{x}$... $\frac{1}{x^2}$...
symmetry	for left-skewed distributions no effect for right-skewed distributions
linearity	overproportional changes in Y if X grows no effect underproportional changes in Y if X grows









Box-Cox transformation

$$y_p^*(x) = \begin{cases} \frac{(x+c)^p - 1}{p} & p \neq 0 \\ \log(x + c) & p = 0 \end{cases}$$

- For a more symmetric distribution, choose p such that

$$y_{0.75} - y_{0.5} = y_{0.5} - y_{0.25}$$

- Maximum-Likelihood method to find p

► let $f(x)$ the density of X

► let $y_p^*(x)$ a transformation of x , then it holds for the density $f(x)$ of X

$$f(x) = g(y_p^*(x)) \left| \frac{dy_p^*(x)}{dx} \right|$$

► choose $g(y_p^*(x))$ as a normal density based on $y_p^*(x)$
 ► maximize the (profile) likelihood

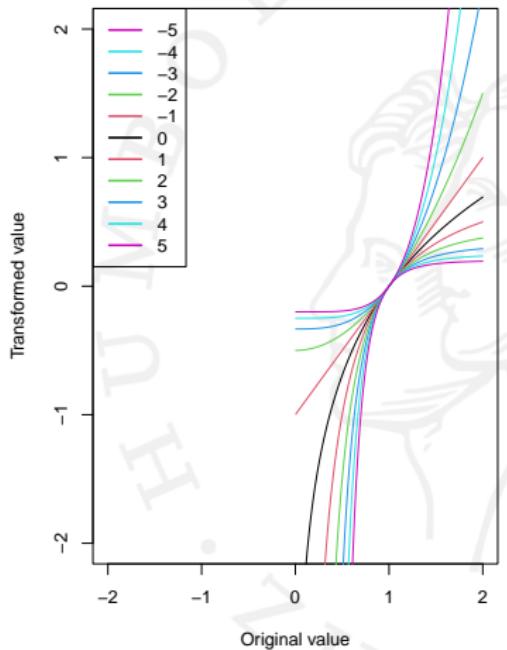
$$f(x) \approx \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x - \mu_p)^2}{2\sigma_p^2}\right) \left| \frac{dy_p^*(x)}{dx} \right|$$

Yeo-Johnson transformation

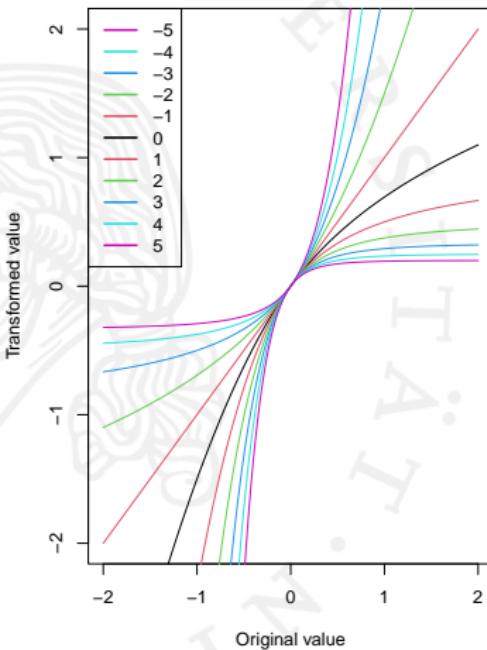
$$y(x) = \begin{cases} \frac{(x+1)^p}{p} & p \neq 0, x \geq 0 \\ \log(x + 1) & p = 0, x \geq 0 \\ -\frac{(1-x)^{2-p}}{2-p} & p \neq 2, x < 0 \\ -\log(1 - x) & p = 2, x < 0 \end{cases}$$

- $x > 0$ then $y(x)$ ist the Box-Cox transformation
- $x < 0$ then $y(x)$ ist the Box-Cox transformation of $1 - x$ and power $2 - p$
- Similar properties as Box-Cox transformation

Box-Cox



Yeo-Johnson



Further transformations

- Arcsin transformation

$$y(x) = \sqrt{n + c_1} \arcsin \left(\frac{x + c_2}{n + c_3} \right)$$

- ▶ transform binomial data to more normal data
- Folded-root transformation, folded-log transformation or logit transformation

$$y_F(x) = \sqrt{x} - \sqrt{1-x}$$

$$y_F^*(x) = \log(x) - \log(1-x)$$

- ▶ both transformation are used for rates, percentages etc.
- ▶ tails of the distribution are enlarged in relation of the center of the distribution

- Retransformation to original range

$$z(x) = a \cdot y(x) + b$$

- ▶ such that $z(x_0) = x_0$ and $z'(x_0) = 1$ for a central point x_0 (e.g. median)



Listing 7.2: example_logit.R

```
1 data(Boston, package="MASS")
2 library("car")
3 par(mfrow=c(1,2))
4 # logit
5 indus <- Boston$indus/100 # percentages
6 hist(indus, main="indus")
7 lindus <- logit(indus)
8 hist(lindus, main="logit(indus)")
```



```
car::logit(p, adjust)
```

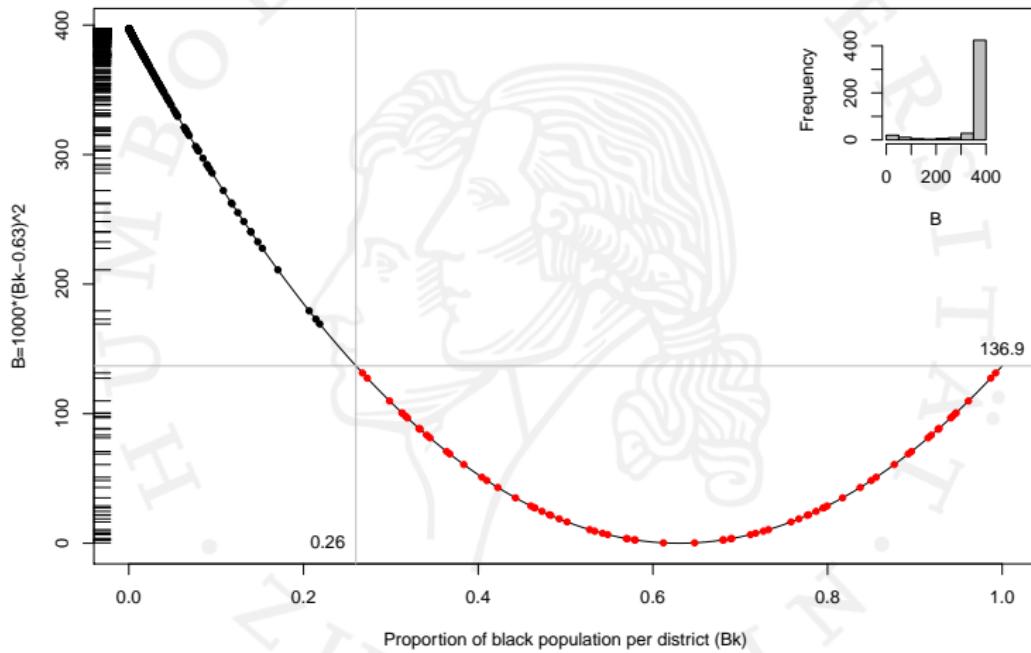
A large, faint watermark of a stylized horse's head and neck, facing left, is centered behind the text.

R Listing 7.3: example_boxcox.R

```
1 data(Boston, package="MASS")
2 library("car")
3 par(mfrow=c(1,2))
4 # Box-Cox
5 hist(Boston$crim, main="crim")
6 bccrim <- bcPower(Boston$crim, 0)
7 hist(bccrim, main="bcPower(crim,0)")
```

- R car::bcPower(U, lambda, jacobian.adjusted=FALSE, gamma=0)
- R car::yjPower(U, lambda, jacobian.adjusted=FALSE)
- R car::estimateTransform(X, Y, family="bcPower", start=NULL)

Interpretation of B



Robust statistics

November 3, 2022

Definition • Consequences • Stem-and-Leaf-Plot • Boxplot • Andrews curves • Standardized distance • Mahalanobis distance • Stahel-Donoho-Outlyingness • Grubbs test • Grubbs/Beck test • Dixon's r statistics • David-Hartley-Pearson test • Walsh's outlier test • Recap: Taylor series • Estimator classes • L-estimators • Breakdown points • M-estimator • Huber k • Hampel • Andrews wave • Tukeys biweight • Estimation • Robust estimators for dispersion

Definition

1. Observations far away from the mass of observations

- ▶ unprecise, but common

an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism

Hawkins, D. M. (1980). *Identification of Outliers*. Dordrecht: Springer Netherlands. isbn: 978-94-015-3996-8 978-94-015-3994-4. doi: 10.1007/978-94-015-3994-4. url: <http://link.springer.com/10.1007/978-94-015-3994-4> (visited on 09/12/2021).

2. Observations which appear to be unlikely given a distribution

- ▶ precise, but uncommon since data distributions must be specified
- ▶ but if we use later ML estimation then we have to specify data distributions

Consequences

- Outliers may influence the later analysis
- Outliers might be (only) extreme values
- Detection of outliers by
 - ▶ graphics,
 - ▶ characteristic parameters, and/or
 - ▶ tests

		Fallnummer	Wert
BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE	Größte Werte	1	136
		2	780
		3	569
		4	2249
		5	2692
	Kleinste Werte	1	1120
		2	1964
		3	1292
		4	1061
		5	1560

Stem-and-Leaf-Plot

- ALLBUS 2008 Data
- v386 (BEFR.: NETTOEINKOMMEN, OFFENE ABFRAGE)

Data	Stem	Leaf	Stem-and-Leaf
4300	=	$4 * 1000$	+3 * 100 0 6
2000	=	$2 * 1000$	+0 * 100 1 27
600	=	$0 * 1000$	+6 * 100 2 07
1200	=	$1 * 1000$	+2 * 100 3
1700	=	$1 * 1000$	+7 * 100 4 3
2700	=	$2 * 1000$	+7 * 100

Stem width: 1000
Each leaf: 1 case

- Stemwidths and leaves (in powers of ten) in SPSS:

Stem width	Leaves
10	0,1,2,3,4,5,6,7,8,9
5	0,1,2,3,4 5,6,7,8,9
2	0,1 2,3 4,5 6,7 8,9

- Additionally
 - ▶ Extremes (not outliers!)
 - ▶ Fractional leaves

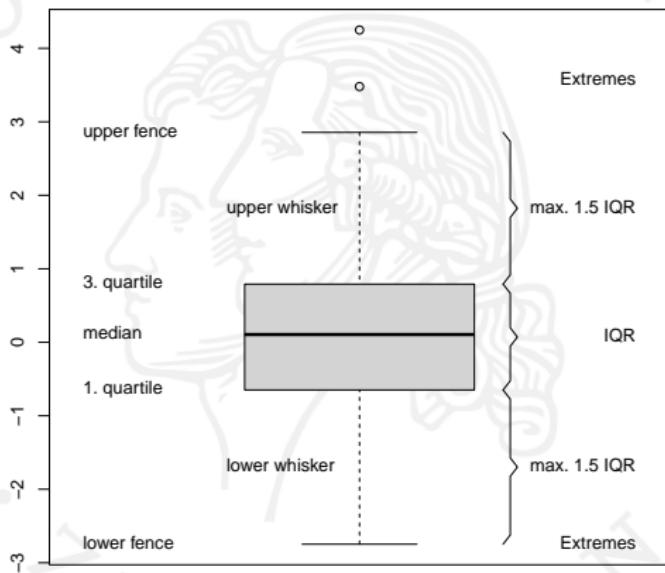
Tukey, John Wilder (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co. 688 pp. isbn: 978-0-201-07616-5.

⌚ Listing 8.1: example_stem.R

```
1 library("MASS") # for Boston Housing data
2 stem(Boston$medv)
```

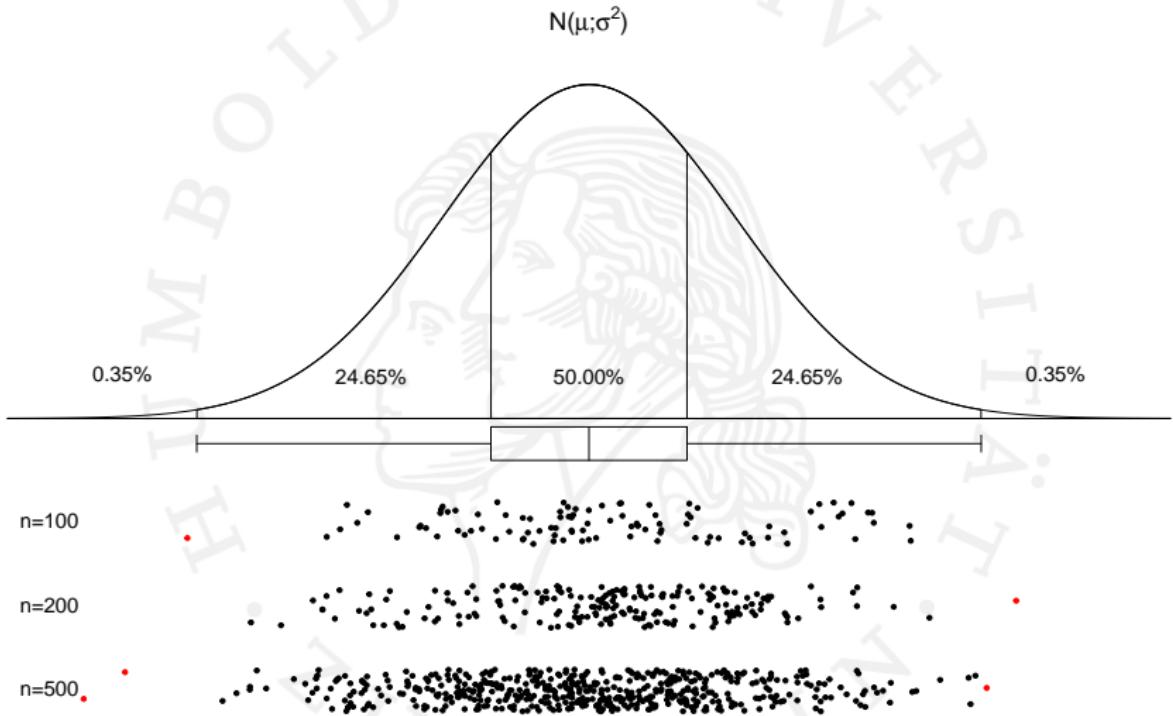
⌚ stem(x)

Boxplot



- Extremes are not necessarily outliers!
 - ▶ mild extremes (o):
 $x_i \in [x_{0.25} - 1.5IQR, x_{0.25} - 3IQR]$ or
 $x_i \in [x_{0.75} + 1.5IQR, x_{0.75} + 3IQR]$
 - ▶ strong extremes (*):
 $x_i \leq x_{0.25} - 3IQR$ or $x_i \geq x_{0.75} + 3IQR$
- For a normal distribution $N(\mu; \sigma^2)$ holds:
 - ▶ $x_{0.25} = \mu - 0.6745\sigma, x_{0.75} = \mu + 0.6745\sigma$
 - ▶ $IQR = 1.35\sigma$
 - ▶ $F(x_{0.25} - 1.5IQR) = 0.0035, F(x_{0.25} - 3.0IQR) = 0.00000117$
- k extreme values will appear if
 - ▶ Mild extremes: $n \geq 143.33k$
 - ▶ Strong extremes: $n \geq 426996k$

Tukey, John Wilder (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co. 688 pp. isbn: 978-0-201-07616-5.



R Listing 8.2: example_boxplot_outliers.R

```
1 n <- 10000
2 x <- rnorm(n)
3 boxplot(x, horizontal=T)
4 rug(x)
```

R boxplot(formula, data, range=1.5, varwidth=F, horizontal=F)

Andrews curves

- Visualization: outliers, clusters
 - ▶ Variables: three (or more) continuous variables
 - ▶ Observations: small
- Problem: order of variables!

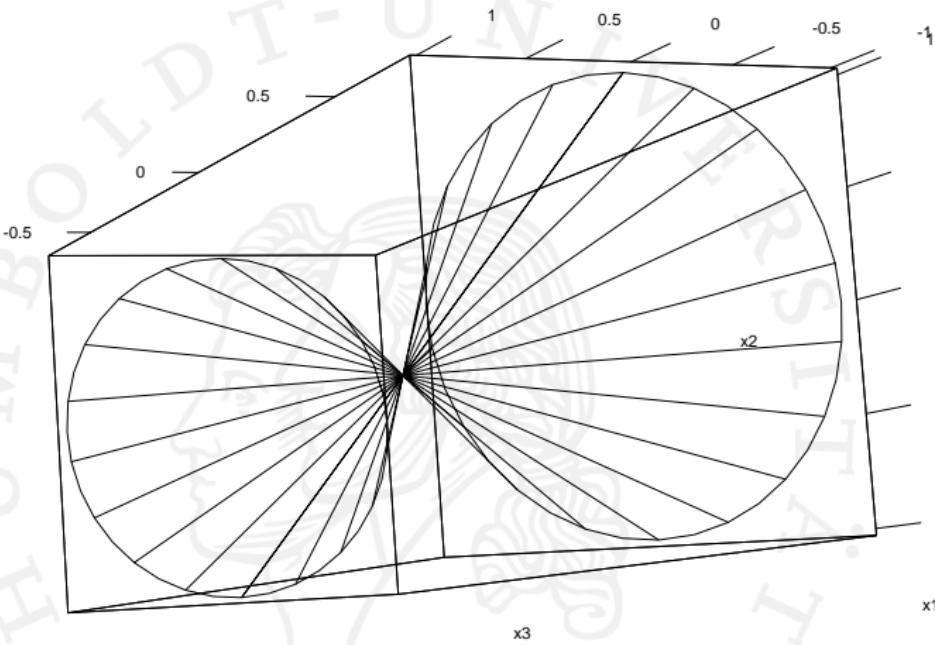
Andrews, D. F. (Mar. 1972). "Plots of High-Dimensional Data". In: *Biometrics* 28.1, p. 125.
issn: 0006341X. doi: 10.2307/2528964. url:
<http://www.jstor.org/stable/2528964?origin=crossref> (visited on 06/23/2016).

- Each observation $x_i = (x_{i,1}, \dots, x_{i,p})$ will be represented by a line

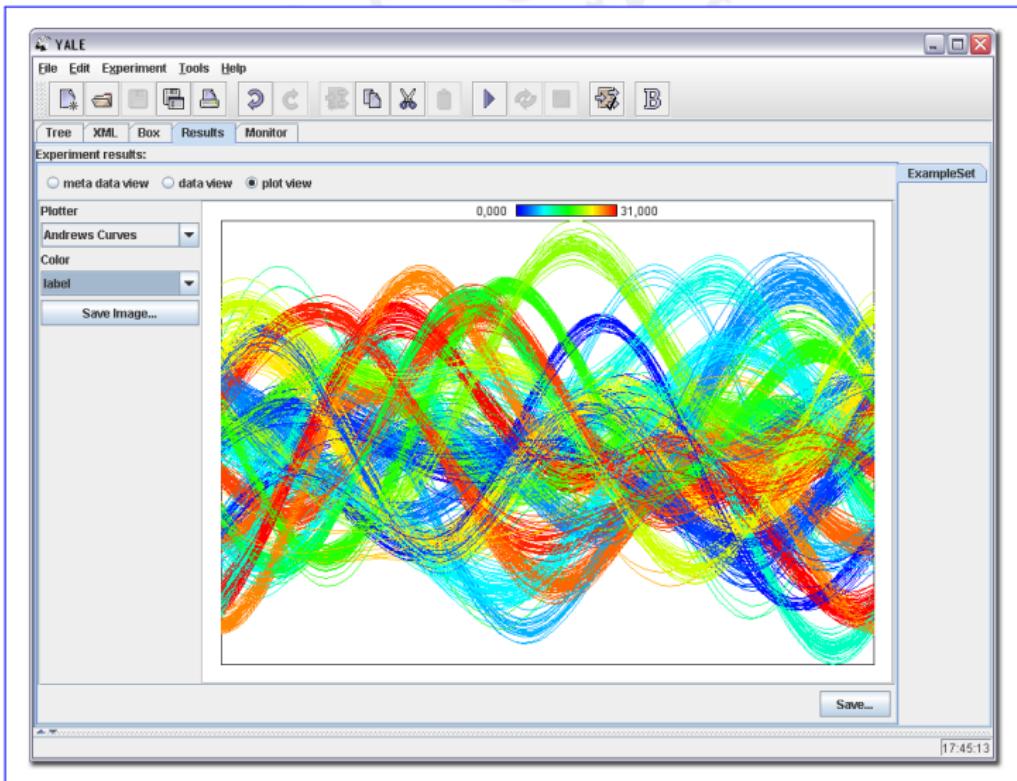
$$f_i(t) = \frac{x_{i,1}}{\sqrt{2}} + x_{i,2} \cos(t) + x_{i,3} \sin(t) + x_{i,4} \cos(2t) + x_{i,5} \sin(2t) + \dots \quad t \in [-\pi; \pi]$$

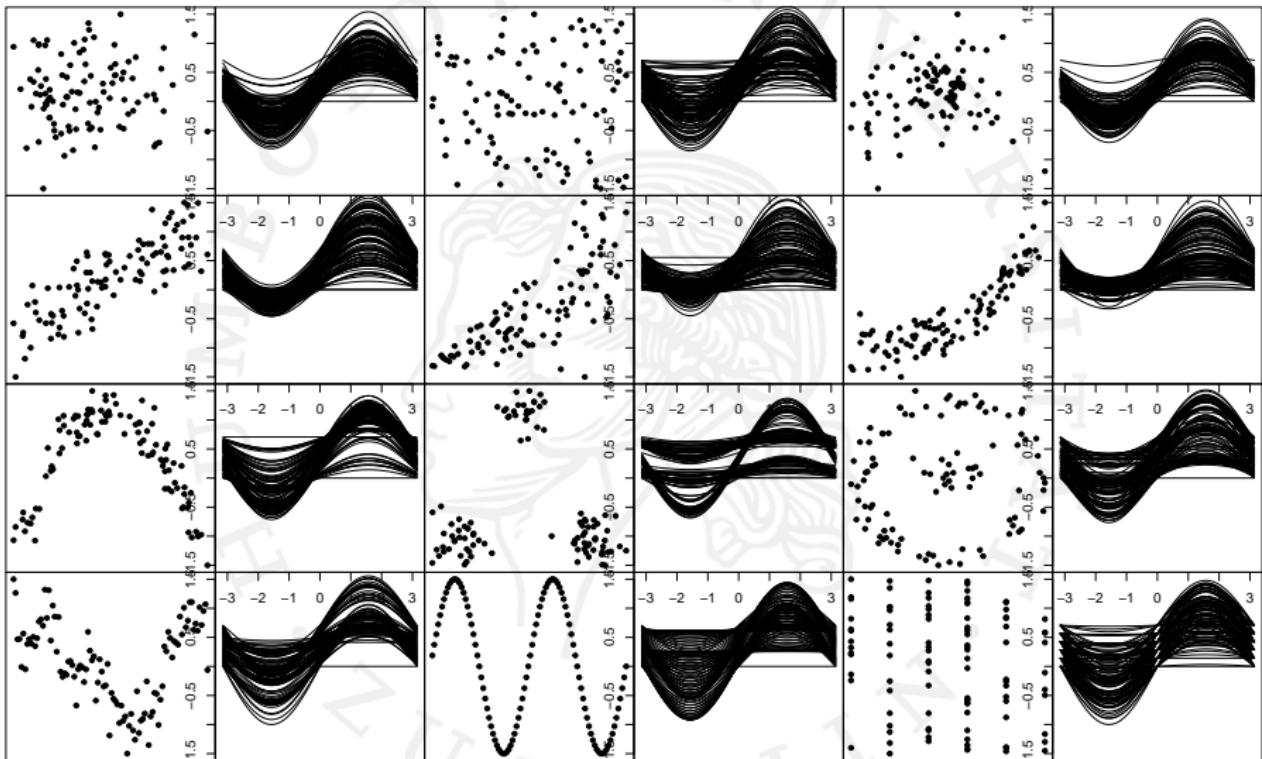
- It holds

$$\pi \underbrace{\sum_{k=1}^p (x_{i,k} - x_{j,k})^2}_{= \text{squared euclidian distance}} = \pi \|x_i - x_j\|^2 = \underbrace{\int_{-\pi}^{\pi} (f_i(t) - f_j(t))^2 dt}_{\approx \text{area between curves}}$$



projection vectors ($p = 3$): $(1/\sqrt{2}, \cos(t), \sin(t))$





 Listing 8.3: example_andrews.R

```

1 library("MASS")  # for Boston Housing data
2 library("andrews")
3 andrews(Boston)

```

 andrews::andrews(df, type=1)

type=2 $f_i(t) = x_{i,1} \cos(t) + x_{i,2} \sin(t) + x_{i,3} \cos(2t) + x_{i,4} \sin(2t) + \dots$

type=3 $f_i(t) = x_{i,1} \cos(t) + x_{i,2} \cos(\sqrt{2}t) + x_{i,3} \cos(\sqrt{3}t) + \dots$

type=4
$$\begin{aligned} f_i(t) = \frac{1}{\sqrt{2}} & [x_{i,1} + x_{i,2}(\sin(t) + \cos(t)) + x_{i,3}(\sin(t) - \cos(t)) \\ & + x_{i,4}(\sin(2t) + \cos(2t)) + x_{i,5}(\sin(2t) - \cos(2t)) + \dots] \end{aligned}$$

Standardized distance

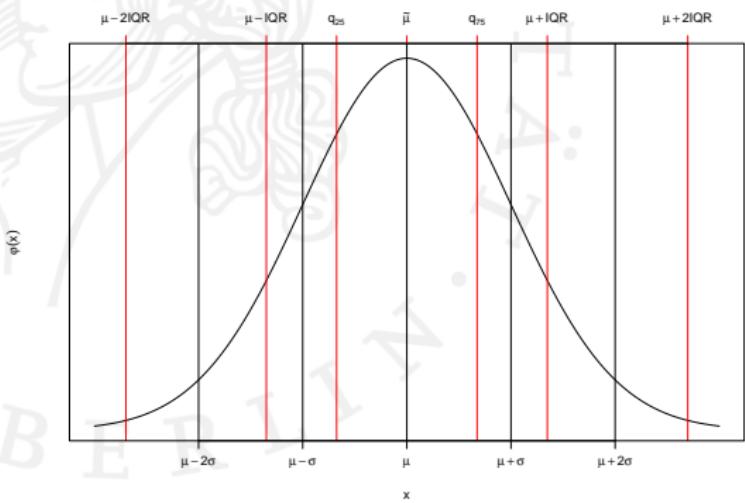
- Measure the distance of each data point to the “center” of the data

$$d(x_i) = \begin{cases} \frac{x_i - \bar{x}}{s_x} & \text{non-robust} \\ \frac{x_i - \tilde{x}}{1.34898 \text{ IQR}} & \text{robust} \end{cases}$$

$$\sigma = 1.34898 \text{ IQR}$$

$$N(\mu; \sigma^2)$$

Borders	Mass
$\mu \pm 0.5\text{IQR}$	50.00%
$\mu \pm \sigma$	68.25%
$\mu \pm \text{IQR}$	82.27%
$\mu \pm 2\sigma$	95.45%
$\mu \pm 2\text{IQR}$	99.30%
$\mu \pm 3\sigma$	99.73%
$\mu \pm 3\text{IQR}$	99.99%
$\mu \pm 4\sigma$	99.99%



Mahalanobis distance

- Compute for each data point

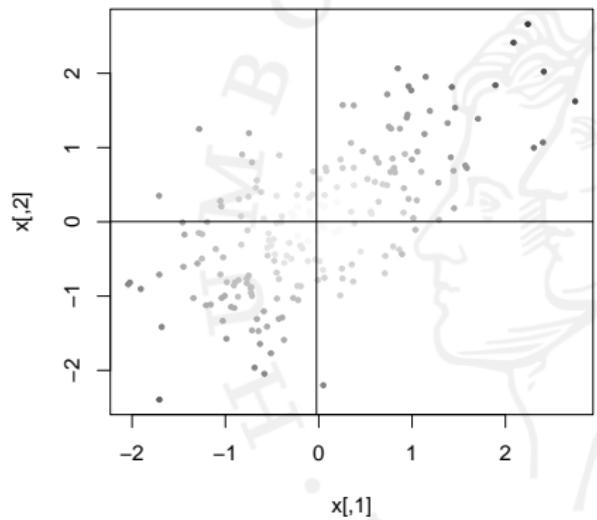
$$d_M(x_i) = \sqrt{(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})}$$

with \bar{x} the multivariate mean and S the covariance matrix of the data

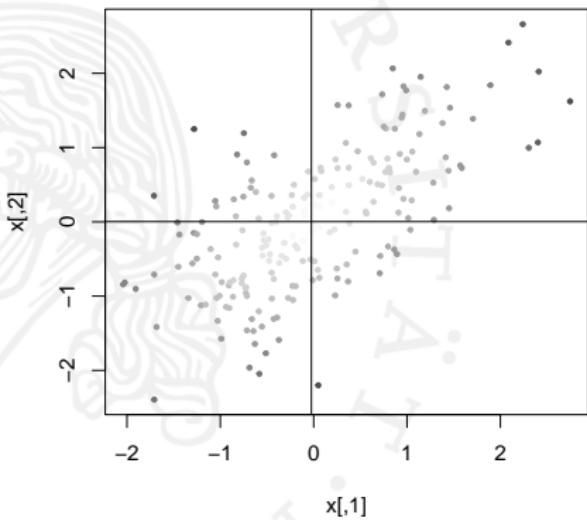
- measure the distance between the center of the observations taking the covariance structure into account
- if the data are uncorrelated then the Mahalanobis distance reduces to the euclidean distance of standardized variables
- disadvantage: it assumes spherical distributed data

Mahalanobis, P. C. (Apr. 1936). "On the generalised distance in statistics". In: *Proceedings National Institute of Science, India*. Vol. 2. 1, pp. 49–55. url:
<http://ir.isical.ac.in/dspace/handle/1/1268>.

Euclidean distance



Mahalanobis distance



 Listing 8.4: example_mahalanobis.R

```
1 library("MASS")    # for Boston Housing data
2 x <- Boston[,-c(1,4,9)]
3 dm <- sqrt(mahalanobis(x, colMeans(x), cov(x)))
4 sort(dm)
5 hist(dm)
6 rug(dm)
```

 mahalanobis(x, center, cov)

Stahel-Donoho-Outlyingness

- Compute for each (standardized) observation $x_i = (x_{i,1}, \dots, x_{i,p})$ and the projection $\alpha = (\alpha_1, \dots, \alpha_p)$

$$x_{i,\alpha} = \alpha^T x_i = \sum_{k=1}^p \alpha_k x_{i,k} \quad \sum \alpha_k^2 = 1$$

- Find the outlyingness $\text{out}(x_i)$ as

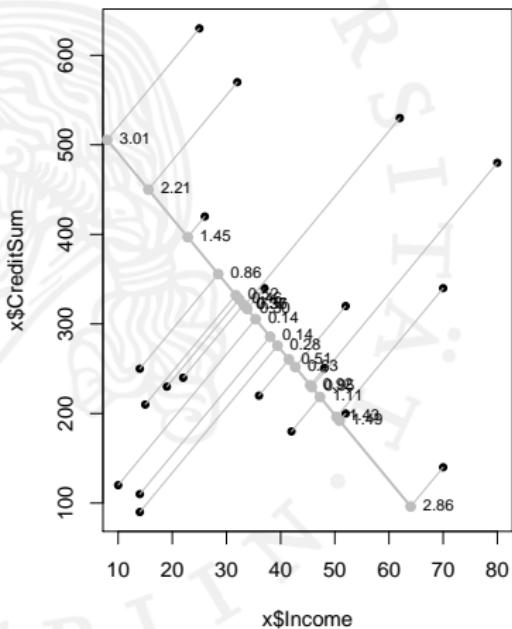
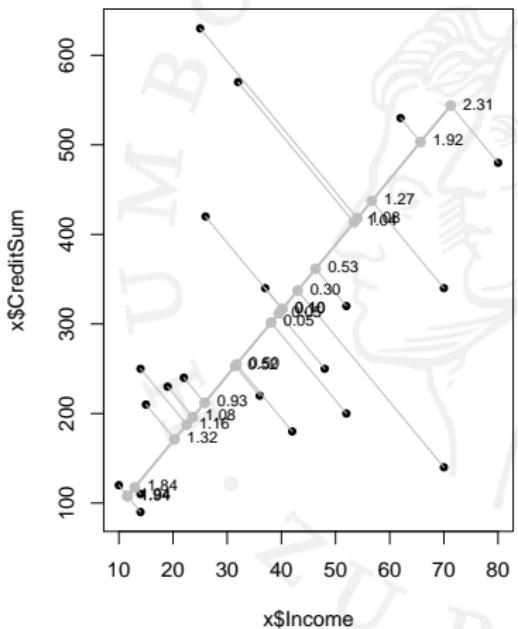
$$\text{out}(x_i) = \sup_{\alpha} \underbrace{\frac{|x_{i,\alpha} - \text{median}(x_{1,\alpha}, \dots, x_{n,\alpha})|}{\text{md}(x_{1,\alpha}, \dots, x_{n,\alpha})}}_{\text{robustified standardization}}$$

$$\text{md}(p_1, \dots, p_n) = \frac{1}{n} \sum_{i=1}^n |p_i - \text{median}(p_1, \dots, p_n)|$$

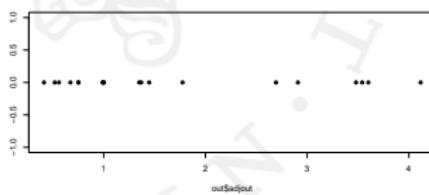
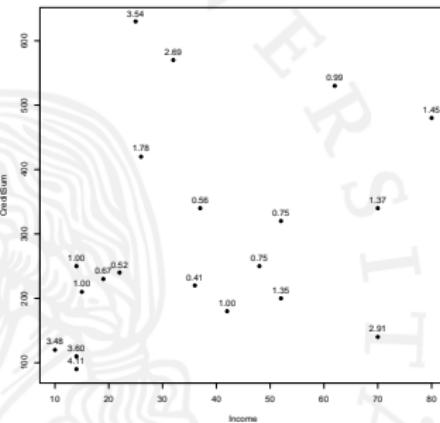
Stahel, Werner A. (1981). "Robuste Schätzungen, infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen". In: doi: 10.3929/ethz-a-000231580. url: <http://dx.doi.org/10.3929/ethz-a-000231580> (visited on 06/23/2016).

Donoho, D. (1982). "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University.

- In practice: Compute a lot of α 's and take the maximum



- R computes an “adjusted” outlyingness
- takes into account the skewness etc.
- the value does not matter, but the order



⌚ Listing 8.5: example_stahel_donoho.R

```
1 library("MASS")    # for Boston Housing data
2 library("robustbase")
3 x   <- Boston[,-c(1,4,9)]
4 set.seed(0)
5 out <- adjOutlyingness(x, ndir=2500)
6 sort(out$adjout)
7 hist(out$adjout)
8 rug(out$adjout)
```

⌚ robustbase::adjOutlyingness(x, ndir=250)

Grubbs test

Assumption: X is normal distributed

Hypotheses: $H_0(1) : x_{(1)}$ is not an outlier vs.

$H_1(1) : x_{(1)}$ is an outlier

$H_0(n) : x_{(n)}$ is not an outlier vs.

$H_1(n) : x_{(n)}$ is an outlier

Test statistics: with \bar{x} mean and s standard deviation

$$t_g(1) = \frac{\bar{x} - x_{(1)}}{s}, \quad t_g(n) = \frac{x_{(n)} - \bar{x}}{s}$$

Reject H_0 : if $t_g(1) > t_{g;n;\alpha}$ or $t_g(n) > t_{g;n;\alpha}$

$t_{g;n;\alpha}$'s are tabulated or

$$t_{g;n;\alpha} \approx \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/n,n-2}^2}{n-2+t_{\alpha/n,n-2}^2}}$$

Grubbs, Frank E. (Mar. 1950). "Sample Criteria for Testing Outlying Observations". In: *The Annals of Mathematical Statistics* 21.1, pp. 27–58. issn: 0003-4851. doi: 10.1214/aoms/1177729885. url: <http://projecteuclid.org/euclid.aoms/1177729885> (visited on 06/23/2016).

R Listing 8.6: example_grubbs.R

```
1 library("MASS") # for Boston Housing data
2 library("outliers")
3 grubbs.test(Boston$medv)
4 grubbs.test(Boston$medv, opposite=T)
```

R outliers::grubbs.test(x, opposite=FALSE)

Grubbs/Beck test

- Assumption: X is normal distributed
- Hypothesis:
 - ▶ $H_0(1) : x_{(1)}, x_{(2)}$ are not outliers vs. $H_1(1) : x_{(1)} \text{ or } x_{(2)}$ is an outlier
 - ▶ $H_0(n) : x_{(n)}, x_{(n-1)}$ are not outliers vs. $H_1(n) : x_{(n)} \text{ or } x_{(n-1)}$ is an outlier
- Test statistics:
 - ▶ $t_{gb}(1) = \frac{SQA(1, 2)}{SQA}$
 - ▶ $t_{gb}(n) = \frac{SQA(n, n - 1)}{SQA}$
$$SQA(k, l) = \sum_{i \neq k, l} (x_{(i)} - \bar{x}_{k, l})^2$$

$$\bar{x}_{k, l} = \frac{1}{n-2} \sum_{i \neq k, l} x_{(i)}$$

$$SQA = \sum_i (x_{(i)} - \bar{x})^2$$
- Critical values $t_{gb; n; \alpha}$ are tabulated
- Null hypothesis is rejected if $t_{gb}(1) > t_{gb; n; \alpha}$ or $t_{gb}(n) > t_{gb; n; \alpha}$

Grubbs, Frank E. and Beck, Glenn (Nov. 1972). "Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations". In: *Technometrics* 14.4, pp. 847–854. issn: 0040-1706, 1537-2723. doi: 10.1080/00401706.1972.10488981. url: <http://www.tandfonline.com/doi/abs/10.1080/00401706.1972.10488981> (visited on 06/24/2016).

R Listing 8.7: example_grubbs_beck.R

```

1  grubbs.beck.test <- function (x, left=T, B=1000, k=2) {
2    gbs <- function(sx, ind) { var(sx[ind])/var(sx) }
3    #
4    n <- length(x)
5    if (left) ind<-(k+1):n else ind<-1:(n-k)
6    vb <- replicate(B, gbs(sort(rnorm(n)), ind))
7    v  <- gbs(sort(scale(x)), ind=ind)
8    list(statistic=(n-3)/(n-1)*v, p.value=sum(vb<=v)/B)
9  }
10 #
11 library("MASS") # for Boston Housing data
12 library("outliers")
13 x <- Boston$medv[sample(506, 20)]
14 grubbs.test(x, type=20)
15 #
16 grubbs.beck.test(Boston$medv, left=F)

```

R outliers::grubbs.test(x, type=20, opposite=FALSE)

⚠ 3<=length(x)<=30

Dixon's r statistics

- Assumptions:
 - ▶ X is normal distributed
 - ▶ we believe there are g large outliers (with large values)
 - ▶ we believe there are k small outliers (with small values)
- Hypothesis:
 - ▶ $H_o(1) : x_{(1)} \text{ is not an outlier}$ vs. $H_1(1) : x_{(1)} \text{ is an outlier}$
 - ▶ $H_o(n) : x_{(n)} \text{ is not an outlier}$ vs. $H_1(n) : x_{(n)} \text{ is an outlier}$
- Test statistics:

$$\triangleright r_{kg}(1) = \frac{x_{(1+k)} - x_{(1)}}{x_{(n-g)} - x_{(1)}} \leq 1$$

$$\triangleright r_{gk}(n) = \frac{x_{(n)} - x_{(n-g)}}{x_{(n)} - x_{(1+k)}} \leq 1$$

$$\triangleright r_{kg}(1) = 1 \text{ or } r_{gk}(n) = 1 \iff \underbrace{x_{(1+k)} = \dots = x_{(n-g)}}_{\text{non-outlying obs.}}$$

- Critical values $r_{n;\alpha}$ are tabulated
- Null hypothesis is rejected if $r_{gk}(1) > r_{n;\alpha}$ or $r_{gk}(n) > r_{n;\alpha}$

Dixon, W. J. (Dec. 1950). "Analysis of Extreme Values". In: *Ann. Math. Statist.* 21.4, pp. 488–506. doi: 10.1214/aoms/1177729747. url: <http://dx.doi.org/10.1214/aoms/1177729747>.

— (Mar. 1951). "Ratios Involving Extreme Values". In: *Ann. Math. Statist.* 22.1, pp. 68–78. doi: 10.1214/aoms/1177729693. url: <http://dx.doi.org/10.1214/aoms/1177729693>.

Rorabacher, David B. (Jan. 1991). "Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level". In: *Analytical Chemistry* 63.2, pp. 139–146. issn: 0003-2700, 1520-6882. doi: 10.1021/ac00002a010. url: <http://pubs.acs.org/doi/abs/10.1021/ac00002a010> (visited on 08/21/2017).

 Listing 8.8: example_dixonr.R

```
1 library("outliers")
2 x <- c(rnorm(29),4)
3 dixon.test(x)
```

☞ outliers::dixon.test(x, type, opposite=F)

⚠ length(x)<=30

type	10	11	12	20	21	other
<i>k</i>	1	1	1	2	2	2
<i>g</i>	0	1	2	0	1	2

David-Hartley-Pearson test

- Assumption: X is normal distributed
- Hypothesis: $H_0 : x_{(1)}$ and $x_{(n)}$ are not outliers vs. $H_1 : x_{(1)}$ or $x_{(n)}$ is an outlier or both are outliers
- Test statistics $t_{dhp} = \frac{x_{(n)} - x_{(1)}}{\sqrt{\frac{SQA}{n-1}}} = \frac{\text{range}}{\text{standard deviation}}$
- Critical values $t_{dhp; n; \alpha}$ are tabulated
- Null hypothesis is rejected if $t_{dhp} > t_{dhp; n; \alpha}$

David, H. A., Hartley, H. O., and Pearson, E. S. (Dec. 1954). "The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation". In: *Biometrika* 41.3, p. 482. issn: 00063444. doi: 10.2307/2332728. url: <http://www.jstor.org/stable/2332728?origin=crossref> (visited on 08/21/2017).

Walsh's outlier test

Assumptions: n large enough, r known in advance

Hypotheses: H_0^{\min} : r th smallest observation is not an outlier vs.

H_1^{\min} : r th smallest observation is an outlier

H_0^{\max} : r th largest observation is not an outlier vs.

H_1^{\max} : r th largest observation is an outlier

Test statistics: $L_{\min} = X_{(r)} - (1 + a)X_{(r+1)} + aX_{(r+c)}$

$L_{\max} = X_{(n-r+1)} - (1 + a)X_{(n-r)} + aX_{(n+1-r-c)}$

$$c = \lfloor \sqrt{2n} \rfloor, b = \sqrt{1/\alpha}, a = \frac{1+b\sqrt{\frac{c-b^2}{c-1}}}{c-b^2-1}$$

Reject H_0 : $|l_{\min}| < 0$ or $|l_{\max}| > 0$

Remarks: $a > 0$ follows $\frac{1}{\lfloor \sqrt{2n} \rfloor - 1} > \alpha$

$\alpha = 10\% \Rightarrow n > 60, \alpha = 5\% \Rightarrow n > 220$

test is conservative since Chebyshev's inequality is used

Theorem (Chebyshev inequality)

For any random variable with existing expectation and variance, it holds

$$P\left(\frac{|X - E(X)|}{\sqrt{Var(X)}} \geq k\right) \leq \frac{1}{k^2}$$

Under the assumption that $E(L) \approx k\sqrt{Var(L)}$ and some Taylor series properties it follows

$$P(L < 0) = P\left(\frac{L - E(L)}{\sqrt{Var(L)}} < -k\right) \leq P\left(\frac{|L - E(L)|}{\sqrt{Var(L)}} > k\right) \leq \frac{1}{k^2}$$

Walsh, John E. (Sept. 1959). "Large sample nonparametric rejection of outlying observations". In: *Annals of the Institute of Statistical Mathematics* 10.3, pp. 223–232. issn: 0020-3157, 1572-9052. doi: 10.1007/BF02883943. url: <http://link.springer.com/10.1007/BF02883943> (visited on 08/21/2017).

Recap: Taylor series

- Let f an infinitely differentiable function. The Taylor series of f at x_0 is

$$T_k(x) = f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \dots + \frac{(x - x_0)^k}{k!} f^{(k)}(x_0)$$

- It can be shown under some regularity conditions holds

$$\|f(x) - T_k(x)\| \leq \frac{f^{(k+1)}(\xi)}{k!} (x - x_0)^{k+1}$$

with $\xi \in [x, x_0]$.

- Walsh needed some assumptions about the Taylor series of the true density which are fulfilled for the “standard” densities

Estimator classes

- L-estimator (linear estimator)
 - ▶ with $0 \leq w_i \leq 1$ and $\sum_i w_i = 1$

$$T_n = \sum_{i=1}^n w_i X_{(i)}$$

- ▶ with symmetric weights $w_i = w_{n-i+1}$ T_n is unbiased for symmetric distributions
- M-estimator (maximum likelihood-type estimator)
 - ▶ is solution of a minimization problem (negative log-likelihood)

$$\sum_{i=1}^n \theta(x_i, T_n) \rightarrow \text{minimal}$$

$$\sum_{i=1}^n \frac{d\theta}{dt}(x_i, \hat{t}_n) = 0$$

- not all L- or M-estimators are robust

L-estimators

- Mean \bar{x}

$$w_i = \frac{1}{n}$$

- Median $x_{0,5}$

$$w_i = \begin{cases} \frac{1}{2} & n \text{ even, } i = n/2, n/2 + 1 \\ 1 & n \text{ odd, } i = (n+1)/2 \\ 0 & \text{otherwise} \end{cases}$$

- α trimmed mean $\bar{x}_{tr;\alpha}$ ($0 \leq \alpha \leq 0.5$)

$$w_i = \begin{cases} \frac{1}{n-2v} & \text{if } v < i < n - v + 1 \\ 0 & \text{otherwise} \end{cases}$$

with v the largest integer number with $v = \lfloor n\alpha \rfloor \leq n\alpha$

- ▶ $\bar{x}_{tr;0.25}$ is called midmean

- Exact α trimmed mean $\bar{x}_{etr;\alpha}$ ($0 \leq \alpha \leq 0.5$)

$$w_i = \begin{cases} \frac{1}{n(1-2\alpha)} & \text{if } v < i < n - v + 1 \\ \frac{1+v-n\alpha}{n(1-2\alpha)} & \text{if } i = v + 1, n - v \\ 0 & \text{otherwise} \end{cases}$$

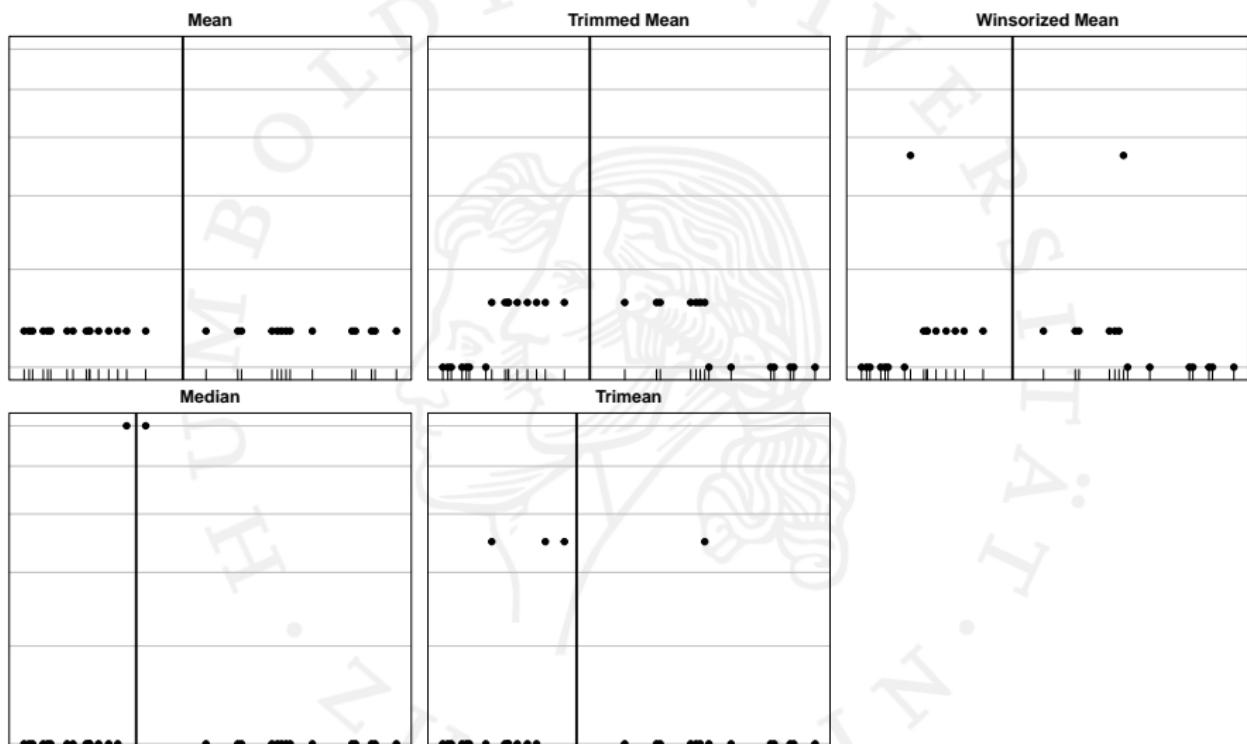
- α winsorized mean $\bar{x}_{w;\alpha}$ ($0 \leq \alpha \leq 0.5$)

$$w_i = \begin{cases} \frac{1}{n} & \text{if } v < i < n - v + 1 \\ \frac{v+1}{n} & \text{if } i = v + 1, n - v \\ 0 & \text{otherwise} \end{cases}$$

- α -Gastwirth-Cohen-mean $\bar{x}_{gc;\alpha;\lambda}$ ($0 \leq \alpha, \lambda \leq 0.5$)

$$w_i = \begin{cases} \lambda & \text{if } i = v + 1, n - v \\ \frac{1}{2} - \lambda & n \text{ even, } i = n/2, n/2 + 1 \\ 1 - 2\lambda & n \text{ odd, } i = (n + 1)/2 \\ 0 & \text{otherwise} \end{cases}$$

► trimean: $\bar{x}_{gc;0.25;0.25} = \frac{1}{4}x_{0.25} + \frac{1}{2}x_{0.5} + \frac{1}{4}x_{0.75}$



- Mean

⌚ `mean(x, na.rm=F)`
⌚ `AVERAGE(number1, number2, ...)`

- Median

⌚ `median(x, na.rm=F)`
⌚ `MEDIAN(number1, number2, ...)`

- Weighted mean

⌚ `weighted.mean(x, w, na.rm=F)`

- Trimmed mean

⌚ `mean(x, trim, na.rm=F)`
⌚ `TRIMMEAN(array, percent)`

- Winsorized mean

⌚ `psych::winsor.mean(x, trim=0.2, na.rm=T)`

Breakdown points

- a measure of “robustness” of an estimator
- the breakdown point ϵ_n^* is defined as the minimal percentage of observations from the left or right which is necessary to achieve an arbitrary result
 - ▶ mean: $\epsilon_n^* = \frac{1}{n}$
 - ▶ α trimmed mean: $\epsilon_n^* = \frac{v+1}{n} = \frac{\lfloor n\alpha \rfloor + 1}{n}$
 - ▶ median: $\epsilon_n^* = \begin{cases} \frac{n-2}{2n} & n \text{ even} \\ \frac{n-1}{2n} & n \text{ odd} \end{cases}$
- the asymptotical breakdown point ϵ^* is defined as

$$\epsilon^* = \lim_{n \rightarrow \infty} \epsilon_n^*$$

- ▶ mean: $\epsilon^* = 0$
- ▶ α trimmed mean: $\epsilon^* = \frac{v+1}{n} = \alpha$
- ▶ median: $\epsilon^* = \frac{1}{2}$

M-estimator

- to find an estimator \hat{t}_n maximize the likelihood function

$$L(x_1, \dots, x_n, t_n) = \prod_{i=1}^n f_i(x_i, t_n)$$

- this is equivalent to minimize the negative loglikelihood

$$-I(x_1, \dots, x_n, t_n) = -\log \left(\prod_{i=1}^n f_i(x_i, t_n) \right) = \sum_{i=1}^n -\log(f_i(x_i, t_n))$$

- Replace $-\log(f_i(x_i, t_n))$ by another function $\theta(x_i, t_n)$ and find \hat{t}_n by minimizing the sum of $\theta(x_i, t_n)$

- For non-scale invariant estimators minimize

$$\sum_{i=1}^n \theta(x_i - t_n)$$

- For example

$$\theta(x_i - t_n) = \frac{(x_i - t_n)^2}{2}$$

$$\sum_{i=1}^n \theta(x_i - t_n) = \frac{1}{2} \sum_{i=1}^n (x_i - t_n)^2 \rightarrow \text{minimal}$$

$$\hat{t}_n = \bar{x}$$

- ▶ \bar{x} is non-robust

- For scale invariant estimators minimize

$$\sum_{i=1}^n \theta\left(\frac{x_i - t_n}{cs_n}\right) = \sum_{i=1}^n \theta(z_i)$$

- ▶ with c a constant
- ▶ s_n a (robust) estimator for the dispersion
- ▶ for s_n usually $md(x_1, \dots, x_n)$ is chosen
- ▶ s_n (dispersion) is independent of t_n (location)
- It holds

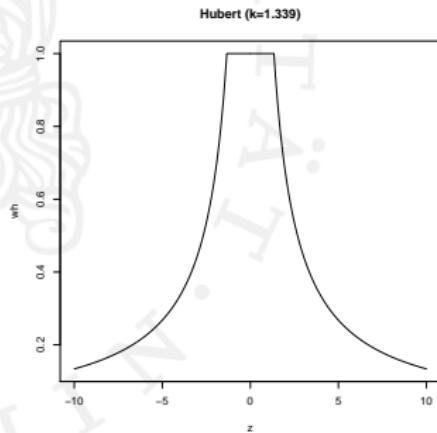
$$\sum_{i=1}^n \theta(z_i) \rightarrow \text{minimal}$$

$$0 = \sum_{i=1}^n \frac{d}{dz} \theta(z_i) = \sum_{i=1}^n \xi(z_i) = \sum_{i=1}^n w(z_i) z_i$$

Huber k

$$\begin{array}{c|cc} & \theta_{Hu}(z) & w_{Hu}(z) \\ \hline |z| \leq k : & \frac{z^2}{2} & 1 \\ |z| > k : & k|z| - \frac{k^2}{2} & \frac{k}{|z|} \end{array}$$

- SPSS: $k = 1.339$
- often: $k = 1.282$ (90% quantile of standard normal distribution)



```
R MASS::huber(y, k=1.5)
R robustbase::huberM(x, k=1.5, weights=NULL, mu=..., s=...)

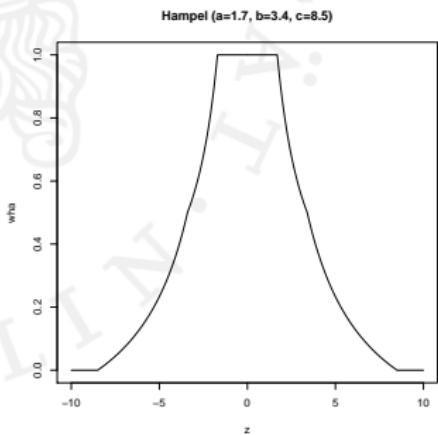
mu = if (is.null(weights))
      median(x)
    else
      wgt.himedian(x, weights)

s  = if (is.null(weights))
      mad(x, center=mu)
    else
      wgt.himedian(abs(x - mu), weights)
```

Hampel

	$\theta_{Ha}(z)$	$w_{Ha}(z)$
$ z \leq a :$	$\frac{z^2}{2}$	1
$a < z \leq b :$	$a z - \frac{a^2}{2}$	$\frac{a}{ z }$
$b < z \leq c :$	$ab - \frac{a^2}{2} + (c-b)\frac{a}{2} \left(1 - \left(\frac{c- z }{c-b}\right)^2\right)$	$\frac{a}{ z } \frac{c- z }{c-b}$
$ z > c :$	$ab - \frac{a^2}{2} + (c-b)\frac{a}{2}$	0

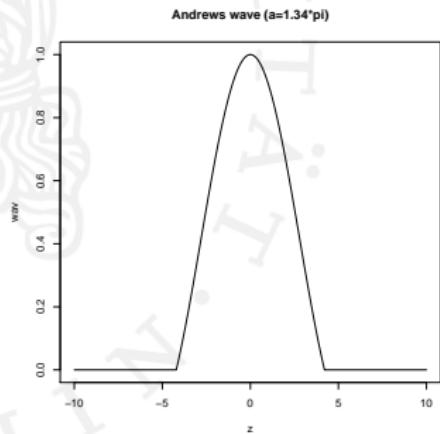
- SPSS: $a = 1.7$, $b = 3.4$, $c = 8.5$



Andrews wave

$$\frac{\theta_{Aw}(z)}{|z| \leq a : \frac{a^2}{\pi^2} \left(1 - \cos\left(\frac{\pi z}{a}\right)\right)} \quad \frac{w_{Aw}(z)}{|z| > a : \frac{2a^2}{\pi^2} \sin\left(\frac{\pi z}{a}\right) \quad 0}$$

- SPSS: $a = 1.34\pi$

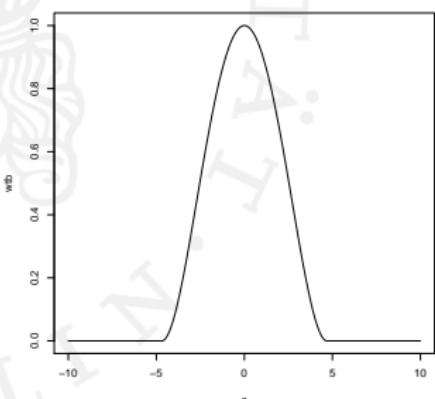


Tukeys biweight

$$\frac{\theta_{Tb}(z)}{|z| \leq a : \quad \frac{a^2}{6} \left(1 - \left(1 - \frac{z^2}{a^2} \right)^3 \right)} \quad \frac{w_{Tb}(z)}{\left(1 - \frac{z^2}{a^2} \right)^2}$$

$$|z| > a : \quad \frac{a^2}{6} \quad 0$$

- SPSS: $a = 4.685$

Tukey's biweight ($a=4.685$)

Estimation

- Direct computation via Newton-Raphson-Verfahren

$$x^{(r+1)} = x^{(r)} - \frac{f(x^{(r)})}{f'(x^{(r)})}$$

$$t_n^{(r+1)} = t_n^{(r)} + cs_n \frac{\sum_{i=1}^n \frac{d\theta}{dz}(z_i^{(r)})}{\sum_{i=1}^n \frac{d^2\theta}{dz^2}(z_i^{(r)})}$$

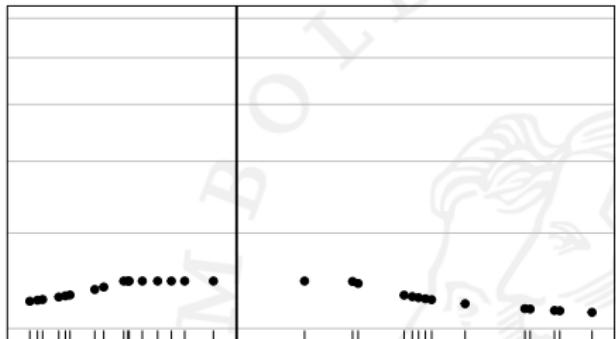
- Iteratively reweighted least-squares (SPSS)

$$z_i^{(r)} = \frac{x_i - t_n^{(r)}}{cs_n}$$

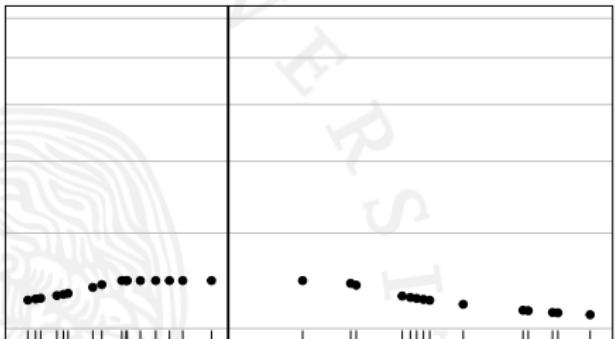
$$t_n^{(r+1)} = \frac{\sum_{i=1}^n w(z_i^{(r)}) x_i}{\sum_{i=1}^n w(z_i^{(r)})}$$

- ▶ $t_n^{(0)} = x_{0.5}$ and stop if changes in $t_n^{(\bullet)}$ are small or $r > 30$

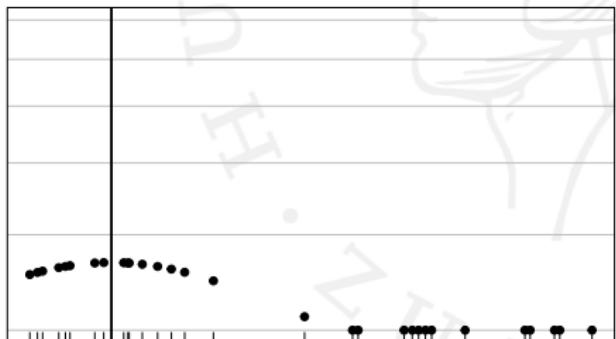
Huber ($k=0.5$)



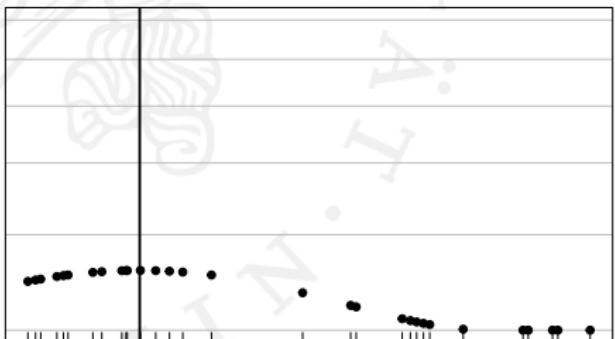
Hampel ($a=0.5, b=1, c=5$)



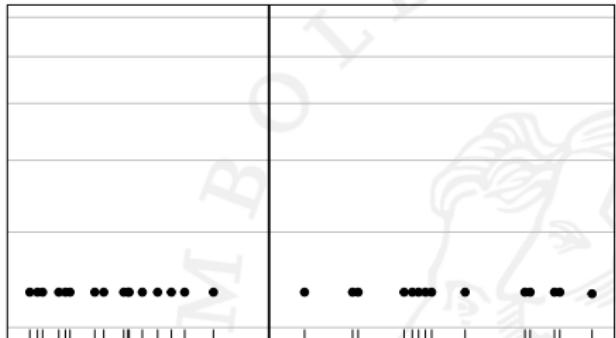
Andrews Wave ($a=1$)



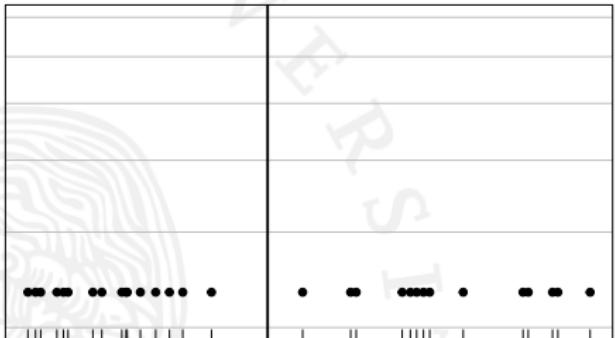
Tukeys Biweight ($a=1.5$)



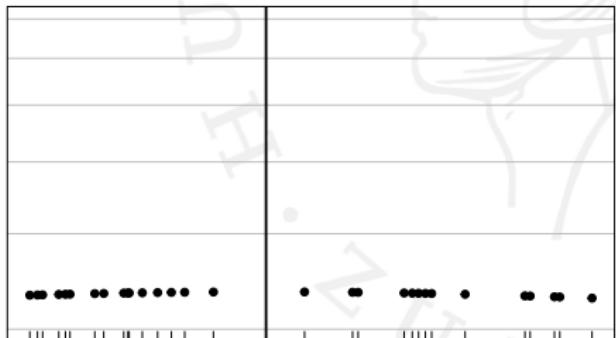
Huber k



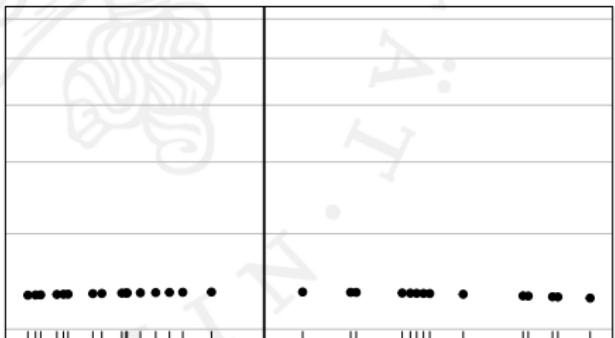
Hampel



Andrews Wave



Tukeys Biweight



Robust estimators for dispersion

- Mean absolute deviation from the mean

$$d(\bar{x}) = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

- Mean absolute deviation from the median

$$d(x_{0.5}) = \frac{1}{n} \sum_{i=1}^n |x_i - x_{0.5}|$$

- Median absolute deviation

$$MAD = \text{median}(|x_i - x_{0.5}|)$$

- ▶ $X_i \sim N(\mu; \sigma^2) \implies E(MAD) = 0.6745\sigma$

- Interquartile range

$$IQR = x_{0.75} - x_{0.25}$$

- α trimmed variance

$$s_{tr;\alpha}^2 = \frac{1}{n - 2v} \sum_{i=v+1}^{n-v+1} (x_{(i)} - \bar{x}_{tr;\alpha})^2$$

- α trimmed absolute deviation

$$d_{tr;\alpha} = \frac{1}{n - 2v} \sum_{i=v+1}^{n-v+1} |x_{(i)} - \bar{x}_{tr;\alpha}|$$

- Mean absolute deviation from the mean

```
R mm <- mean(abs(x-mean(x)))
```

- Mean absolute deviation from the median

```
R m5 <- mean(abs(x-median(x)))
```

- Median absolute deviation

```
R mad(x, center=median(x), na.rm = FALSE)
```

- Interquartile range

```
R IQR(x, na.rm=F, type=7)
```

- α winsorized variance

```
R winsor.var(x, trim=0.2, na.rm=T)
```

```
R winsor.sd(x, trim=0.2, na.rm=T)
```

Missing values

November 3, 2022

Poverty endangering of children in germany • Importance of missing value handling • Coding of missing values • Types of missing values • Consequences • Exploratory missingness analysis • Impact of missingness • Imputation methods • A simulated example • Multiple imputation • EM-Algorithm • Examples • Data augmentation • NORM • Chained equation • Combine estimates • Hints for imputation

Poverty endangering of children in germany

Example 9.15

- 2009 the DIW estimated the poverty risk of german children at 16.3% (using SOEP data for 2005)
- three weeks before the Bundestagswahl (27 Sep 2009) the figure was published by OECD (OECD average: 12.3%)
- heated debate about poverty risk of children in germany
- january 2010 increase of Kindergeld by 20 EUR ($\geq +10\%$)
- 2011 DIW published only a rate of 8.3% for 2005 without any further notice
- problem: in case of non-response the income of a family was set to zero
- DIW changed the methodology in 2010

Rademaker, Maike (2011). "Fehlerhafte Statistik: Kinderarmut nur halb so hoch wie gedacht".

In: *Financial Times Deutschland*. url: www.ftd.de/politik/deutschland/:fehlerhafte-statistik-kinderarmut-nur-halb-so-hoch-wie-gedacht/60048191.html.

Importance of missing value handling

- Q: When do we need to worry about procedures to impute missing values?
- A: If we want to ensure that the inference to the population is still possible.
- Missing values can have various reasons
 - ▶ shame, (political) correctness
 - ▶ ignorance, non-applicable
 - ▶ exhaustion
 - ▶ misstyping
 - ▶ ...

Coding of missing values

- IEEE Standard for Floating-Point Arithmetic (IEEE 754)

$$(-1)^s \times c \times b^q$$

- sign $s = 0$ or 1 , base/radix $b = 2$ or 10 ,
- a non-negative integer significand c and an integer exponent q
- $-12.345 = (-1)^1 \times 12345 \times 10^{-3}$
- Format requires special coding for
 - infinity: $+\infty$ and $-\infty$.
 - not a number (NaN): a quiet and a signaling one
- Floating points numbers
 - single precision: 1 bit sign, 8 bit exponent, 23 bit significand
 - double precision: 1 bit sign, 11 bit exponent, 52 bit significand
 - quiet NaN: $s = 0$, exponent bits are all set to 1
 - signaling NaN: $s = 1$, exponent bits are all set to 1
 - NaN: unused bits can carry payload, e.g. for source of NaN

 Listing 9.1: example_na.R

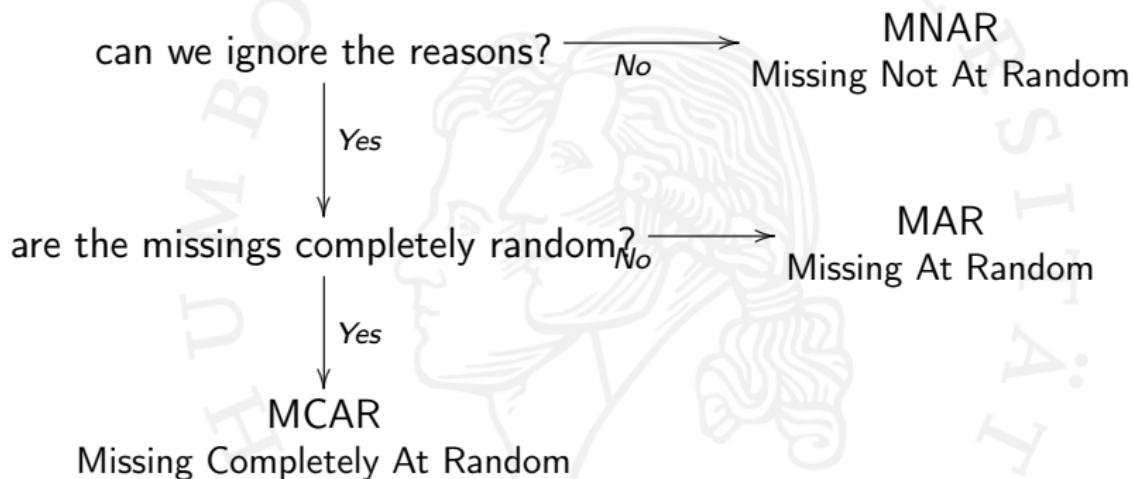
```
1 # NA - datum not available
2 class(NA)
3 y <- c(1, 2, 3, NA)
4 is.na(y)
5 is.nan(y)
6 mean(y)
7 mean(y, na.rm=T)
8 # NaN - invalid operation
9 class(0/0)
10 y <- c(1, 2, 3, 0/0)
11 is.na(y)
12 is.nan(y)
13 mean(y)
14 mean(y, na.rm=T)
```

☞ is.na(x)

⚠ does not support a payload, e.g. for source of NA

☞ is.nan(x)

Types of missing values



- the formal definition depends on the reason why a missing value appears
- the reason determines how missing values can be imputed

- Formal definition:

X variable with no missing values,

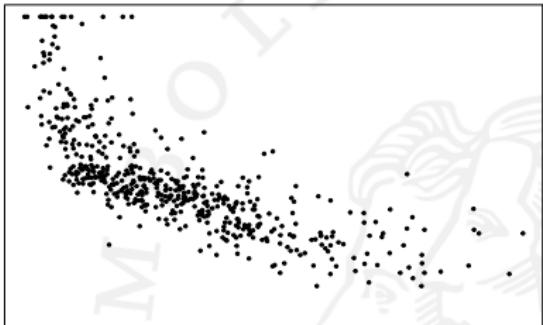
Y variable with missing values

- ▶ Missing Completely At Random: the missingness does not depend on X nor on Y
- ▶ Missing At Random: the missingness depends on X , but not on Y
- ▶ Missing Not At Random: the missingness depends on Y

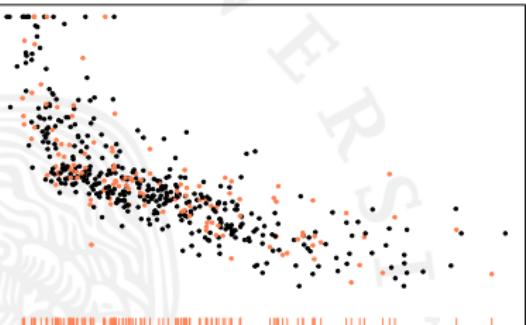
- Examples:

- ▶ MCAR: recording error
- ▶ MAR: double sampling, sampling for non-response followup, planned missingness
- ▶ MNAR: sample surveys with non-respondents (not at home, unwilling to answer), respondent out of range (in survival analysis)

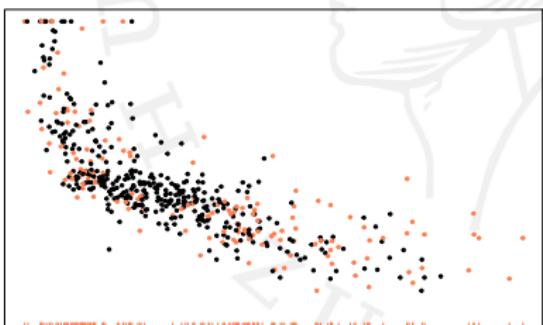
Full data (x: LSTAT, y: MEDV)



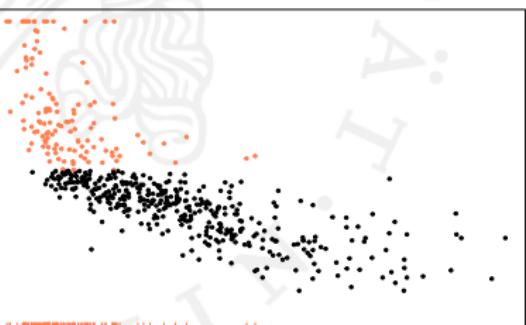
MCAR



MAR



MNAR



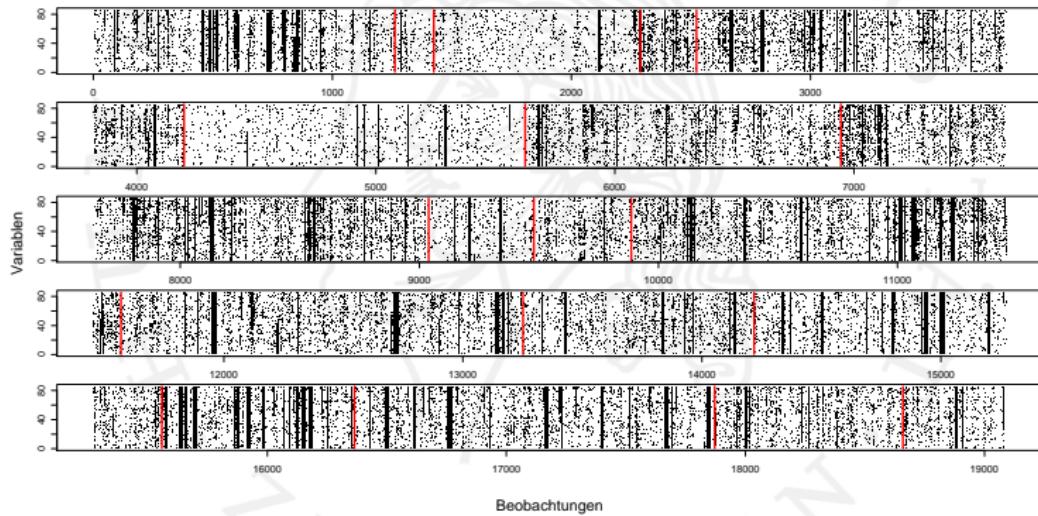
Consequences

- MCAR
 - ▶ can be handled with appropriate methods
 - ▶ consistent estimators on the observed data are still consistent
 - ▶ may imply larger variances on (parameter) estimates
- MAR
 - ▶ there is no possibility to test whether missings are MAR or MNAR
 - ▶ in most cases we can expect a departure from the MAR assumption
 - ▶ in many realistic cases an erroneous specification has only minor impact on estimates
- MNAR
 - ▶ needs to be modelled explicitly
 - ▶ obtain follow up answer from non-respondents

Exploratory missingness analysis

- Which variables have a lot of missings?
- Which observations have a lot of missings?
 - ▶ possibly exclude variables or observations from further analysis
- Bivariate missingness
 - ▶ create for each pair of variables a binary variable (0=value not missing, 1=value missing)
 - ▶ is the missingness in pairs of variables are dependent?
- Missingness patterns
 - ▶ create frequency table of missing patterns
 - ▶ do we observe a specific patterns?
- Missingness map
 - ▶ use a scatterplot to plot the missing values
 - ▶ a point/bar is colored differently if missing or not
 - ▶ reorder columns or rows, e.g. after number of missings

Missingness map for $n \approx 19000$ and $p = 89$



The background of the slide features a large, faint watermark of the HU Berlin logo, which consists of a stylized profile of a person's head and the text "HANNOVER UNIVERSITÄT BERLIN".

R Listing 9.2: example_miss_misssmap.R

```
1 library("VIM")
2 mm <- aggr(sleep) # plot(mm)
3 print(mm)
4 summary(mm)
```

R Listing 9.3: example_miss_explore.R

```
1 library("VIM")
2 par(mfrow=c(2,2))
3 barMiss(sleep$Danger)
4 histMiss(sleep$NonD)
5 scattMiss(cbind(sleep$Dream, sleep$Sleep))
6 marginplot(cbind(sleep$Dream, sleep$Sleep))
```

R VIM::aggr(x, plot=TRUE)

R VIM::...Miss(x)

R VIM::marginplot(x)

Impact of missingness

- create for each variable with missings a binary variable (0=value not missing, 1=value missing)
- on other variables apply subgroup analysis based on the binary variable
- if we have differences in the distribution or the parameter → MAR or MNAR
- if we have no differences in the distribution or the parameter → MCAR
- Little's test compares two models and tests for the MCAR case
- there is no test available to distinguish between MAR and MNAR case

⌚ BaylorEdPsych::LittleMCAR(x)

Little, Roderick J. A. (Dec. 1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values". In: *Journal of the American Statistical Association* 83.404, p. 1198. issn: 01621459. doi: 10.2307/2290157. url: <http://www.jstor.org/stable/2290157?origin=crossref> (visited on 10/19/2017).

Imputation methods

- Older methods
 - ▶ case deletion/listwise deletion
 - ▶ available case analysis
 - ▶ reweighting
- Single imputation
 - ▶ k nearest neighbour
 - ▶ unconditional mean (item average)
 - ▶ unconditional distribution/hot deck
 - ▶ conditional mean
 - ▶ conditional distribution/predictive distribution
 - ▶ Maximum-Likelihood
- Multiple imputation

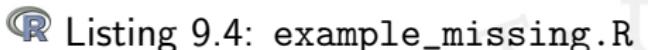
⌚ VIM::kNN(data, k=5)

⌚ VIM::hotdeck(data)

- case deletion
 - ▶ may reduce the dataset considerably
 - ▶ valid under MCAR in simple cases, but inefficient
 - ▶ under MAR biased estimates
 - ⚠ in SPSS often the default action
- available case analysis
 - ▶ if for an analysis only a subset of variables is required then use all complete observations in this subset
 - ▶ example: correlation matrix

Rubin, Donald B., ed. (June 9, 1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. isbn: 978-0-470-31669-6 978-0-471-08705-2. url:
<http://doi.wiley.com/10.1002/9780470316696> (visited on 08/22/2016).

Schafer, Joseph L. (2000). *Analysis of incomplete multivariate data*. 1. ed., 1. CRC Press reprint. Monographs on statistics and applied probability 72. OCLC: 249266966. Boca Raton: Chapman & Hall/CRC. 430 pp. isbn: 978-0-412-04061-0.

 Listing 9.4: example_missing.R

```
1 data("allbus2018", package="mmstat4")
2 vars <- startsWith(names(allbus2018), "pt")
3 trust <- allbus2018[,vars]
4 # count number of missing values
5 r <- is.na(trust)
6 # no. per column
7 apply(r, 2, sum)
8 # no. per row
9 apply(r, 1, sum)
10 # number of complete cases
11 cc <- complete.cases(trust)
12 sum(cc)
13 # filter functions
14 head(na.omit(trust))
15 try(na.fail(trust))
```

④ complete.cases(x, ...)

④ na.omit(x)

④ na.fail(x)



Listing 9.5: example_impute1.R

```
1  data("allbus2012", package="mmstat4")
2  body      <- as.data.frame(allbus2012)
3  names(body) <- c("age", "height", "weight")
4  # number of NAs
5  nobody <- is.na(body)
6  apply(nobody, 2, sum)
7  # full data
8  mean(body$weight)
9  cor(body)
10 # case deletion
11 mean(body$weight, na.rm=T)
12 cor(body, use="complete.obs")
13 sum(complete.cases(body))
14 # available case analysis
15 cor(body, use="pairwise.complete.obs")
16 crossprod(!nobody)
```

A simulated example

- X systolic blood pressure measured in january
- Y systolic blood pressure measured in february (follow-up)
- data simulated from

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(\begin{pmatrix} 125 \\ 125 \end{pmatrix}, \begin{pmatrix} 625 & 375 \\ 375 & 625 \end{pmatrix} \right), \rho = 0.6$$

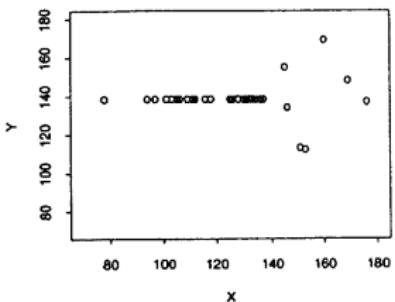
- MCAR on Y : select randomly from the Y observations
- MAR on Y : second reading if $X > 140$
- MNAR on Y : second recording if $Y > 140$
 - ▶ other possibility: second recording only when first and second reading differ
- Parameters to be analyzed: $\mu_Y, \sigma_Y, \rho, \beta_{Y|X}, \beta_{X|Y}$

Schafer, Joseph L. and Graham, John W. (2002). "Missing data: Our view of the state of the art.". In: *Psychological Methods* 7.2, pp. 147–177. issn: 1939-1463, 1082-989X. doi: 10.1037/1082-989X.7.2.147. url: <http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.7.2.147> (visited on 08/22/2016).

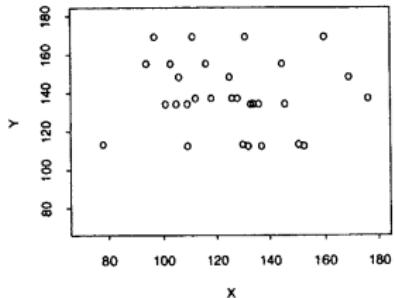
*Performance of Listwise Deletion for Parameter Estimates and Confidence Intervals Over 1,000 Samples
(N = 50 Participants)*

Parameter	MCAR	MAR	MNAR
Average parameter estimate (with RMSE in parentheses)			
μ_Y = 125.0	125.0 (6.95)	143.3 (19.3)	155.5 (30.7)
σ_Y = 25.0	24.6 (5.26)	20.9 (5.84)	12.2 (13.2)
ρ = .60	.59 (.19)	.33 (.37)	.34 (.36)
β_{YX} = .60	.61 (.27)	.60 (.51)	.21 (.43)
β_{XY} = .60	.60 (.25)	.20 (.44)	.60 (.52)
Coverage (with average interval width in parentheses)			
μ_Y	94.3 (30.0)	18.8 (25.0)	0.0 (14.7)
σ_Y	94.3 (23.3)	90.7 (19.4)	17.4 (11.4)
ρ	95.4 (0.76)	82.5 (0.93)	82.7 (0.94)
β_{YX}	94.6 (1.10)	95.9 (2.20)	40.0 (0.73)
β_{XY}	95.3 (1.08)	37.7 (0.71)	96.6 (2.23)

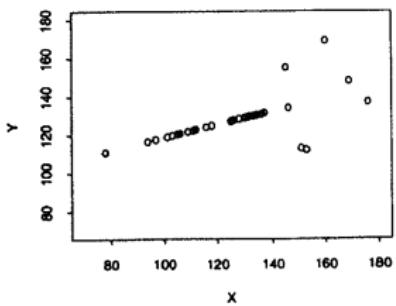
(a) Mean Substitution



(b) Hot Deck



(c) Conditional Mean



(d) Predictive Distribution

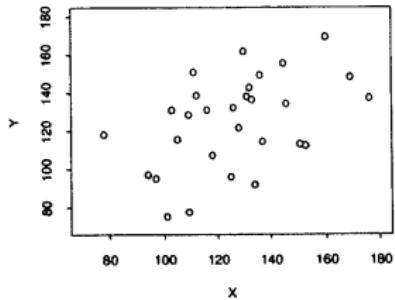


Table 3

*Performance of Single-Imputation Methods for Parameter Estimates and Confidence Intervals Over 1,000 Samples
(N = 50 Participants)*

Parameter	MCAR				MAR				MNAR			
	MS	HD	CM	PD	MS	HD	CM	PD	MS	HD	CM	PD
Average parameter estimate (with RMSE in parentheses)												
$\mu_Y = 125.0$	125.1 (7.18)	125.2 (7.89)	125.2 (6.26)	125.1 (6.57)	143.5 (19.4)	143.5 (19.5)	124.9 (18.1)	124.8 (18.3)	155.5 (30.7)	155.5 (30.73)	151.6 (26.9)	151.6 (26.9)
$\sigma_Y = 25.0$	12.3 (13.0)	23.4 (5.40)	18.2 (8.57)	24.7 (5.37)	10.6 (14.6)	20.0 (6.68)	20.4 (10.7)	27.0 (8.77)	6.20 (18.9)	11.7 (13.7)	8.42 (16.9)	12.9 (12.7)
$p = .60$.30 (.32)	.16 (.46)	.79 (.27)	.59 (.20)	.08 (.52)	.04 (.57)	.64 (.48)	.50 (.40)	.15 (.47)	.08 (.53)	.55 (.40)	.38 (.37)
$\beta_{YX} = .60$.16 (.45)	.16 (.47)	.61 (.25)	.60 (.27)	.04 (.56)	.04 (.57)	.61 (.57)	.62 (.57)	.04 (.56)	.04 (.56)	.21 (.43)	.21 (.43)
$\beta_{XY} = .60$.61 (.26)	.17 (.46)	1.12 (.64)	.60 (.24)	.20 (.44)	.06 (.56)	.78 (.75)	.45 (.40)	.61 (.55)	.19 (.53)	1.63 (1.72)	.76 (.68)
Coverage (with average interval width in parentheses)												
μ_Y	39.2 (7.0)	60.0 (13.3)	58.5 (10.4)	71.0 (14.1)	0.2 (6.0)	2.4 (11.4)	25.7 (11.6)	32.3 (15.3)	0.0 (3.5)	0.0 (6.7)	0.0 (4.8)	0.0 (7.3)
σ_Y	0.7 (5.1)	63.7 (9.6)	31.3 (7.5)	65.4 (10.2)	0.1 (4.4)	45.3 (8.2)	30.0 (8.4)	49.4 (11.1)	0.0 (2.5)	1.7 (4.8)	0.7 (3.5)	4.4 (5.3)
p	25.5 (0.50)	5.5 (0.53)	21.7 (0.19)	65.0 (0.35)	0.0 (0.55)	0.0 (0.55)	19.6 (0.21)	40.7 (0.34)	2.2 (0.54)	0.5 (0.54)	37.6 (0.31)	50.0 (0.43)
β_{YX}	1.2 (0.27)	16.5 (0.54)	38.6 (0.22)	63.5 (0.44)	0.0 (0.25)	0.8 (0.47)	17.2 (0.22)	33.5 (0.45)	0.0 (0.14)	0.1 (0.27)	3.1 (0.13)	7.4 (0.26)
β_{XY}	98.1 (1.18)	23.9 (0.63)	8.9 (0.50)	71.1 (0.47)	91.2 (1.43)	14.5 (0.75)	18.6 (0.56)	60.0 (0.46)	97.4 (2.50)	71.3 (1.30)	19.1 (1.46)	56.2 (1.05)

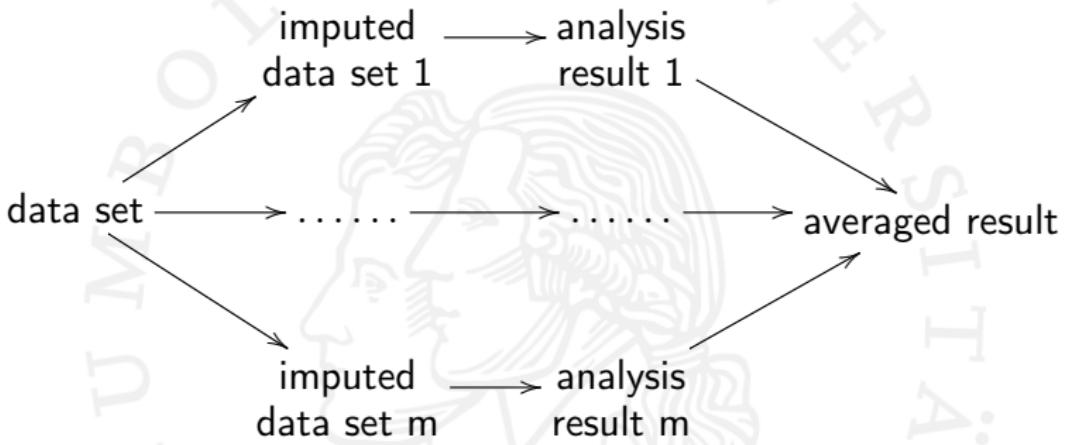
- table explanations
 - ▶ RMSE = Root Mean Squared Error
 - ▶ if parameter standard error estimate is one half of absolute size of parameter then serious bias in estimate, marked in bold
 - ▶ coverage = percentage of confidence intervals covering the parameters (should be 95%), less than 90% are in bold
- MNAR
 - ▶ none of the methods work
 - ▶ the unbiased estimate of $\beta_{X|Y}$ is an artefact (normal distribution)
- MAR/MCAR
 - ▶ Predictive distribution works
 - ▶ Mean substitution works only for estimating μ_Y
- partially severe undercoverage
 - ▶ also due to less information

 Listing 9.6: example_schafer1.R

```
1 library("MASS")
2 sig <- matrix(c(625,375,375,625), ncol=2)
3 n   <- 50
4 x   <- mvrnorm(n, mu=c(125,125), Sigma=sig)
5 plot(x)
6 #na <- runif(n)<0.6                                # MCAR
7 na <- (x[,1]<140)                                    # MAR
8 #na <- (x[,2]<140)                                    # MNAR
9 #x <- x[!na,]                                         # case deletion
10 #x[na,2] <- sample(x[!na,2], sum(na), r=T)          # hot deck
11 #x[na,2] <- mean(x[!na,2])                          # mean substitution
12 lm <- lm(V2~V1, data=as.data.frame(x[!na,]))
13 b <- coefficients(lm)
14 x[na,2] <- b[1]+b[2]*x[na,1]                         # cond. mean
15 r <- residuals(lm)
16 x[na,2] <- x[na,2]+sample(r, sum(na), r=T)          # pred. dist
17 points(x, col="red", pch=19, cex=0.75)
```

Multiple imputation

- Maximum-Likelihood
 - ▶ requires additional information
 - ▶ delivers better results (confidence intervals)
- EM-Algorithm
 - ▶ consider missing values as random variables
 - ▶ estimate unknown parameters via ML-Method
 - ▶ repeat the last two steps until estimate stabilizes
- Multiple imputation
 - ▶ fill missing values by data augmentation
 - ▶ does not help under MNAR
 - ▶ but provides at least information about the variability



- average results $\hat{\theta}_{MI} = \frac{1}{m} \sum_i \hat{\theta}_i$
- $\widehat{Var}(\hat{\theta}_i)$ can be computed easily
- m odd, not too large ($m = 3, 5, 7, 11$)

EM-Algorithm

- Instead of full loglikelihood

$$l(\theta) = \sum_{\text{non-miss}} l_i(\theta) + \sum_{\text{miss}} l_i(\theta)$$

- Maximize $Q(\theta, \theta_t)$, the expectation over the missing observations

$$\begin{aligned} Q(\theta, \theta_t) &= \int_{\text{miss}} l(\theta) \prod_{\text{miss}} f_i(\theta_t) d\text{miss} \\ &= \sum_{\text{non-miss}} l_i(\theta) + \sum_{\text{miss}} \int_{\text{miss}} l_i(\theta) \prod_{\text{miss}} f_i(\theta_t) d\text{miss} \\ &= \sum_{\text{non-miss}} l_i(\theta) + \underbrace{\sum_{\text{miss}} \int_{\text{miss}} l_i(\theta) f_i(\theta_t) d\text{miss}}_{H(\theta, \theta_t)}; \end{aligned}$$

Example 9.16

$$\begin{aligned}
 X_i &\sim N(\mu; 1) \rightarrow f_i(\mu) = \exp(-0.5(x_i - \mu)^2) \\
 L(\mu) &= \exp(-0.5(x_1 - \mu)^2) \exp(-0.5(x_2 - \mu)^2) \\
 l(\mu) &= -0.5(x_1 - \mu)^2 - 0.5(x_2 - \mu)^2 \\
 &\quad 2 \text{ obs., } x_2 \text{ missing} \\
 Q(\mu, \mu_t) &= \int_{-\infty}^{\infty} (-0.5(x_1 - \mu)^2) - 0.5(x_2 - \mu)^2 * \\
 &\quad \exp(-0.5(x_2 - \mu_t)^2) dx_2 \\
 &= -0.5(x_1 - \mu)^2 \\
 &\quad -0.5 \underbrace{\int_{-\infty}^{\infty} (x_2 - \mu)^2 \exp(-0.5(x_2 - \mu_t)^2) dx_2}_{H(\mu, \mu_t)} \\
 \mu_{t+1} &= \max_{\mu} Q(\mu, \mu_t)
 \end{aligned}$$

- Initialize
 - ▶ find a preliminary estimate θ_0 , e.g. from non-missing observations
- E-Step: $E(I(X)) = \int I(x)f(x)dx$
 - ▶ compute $H(\theta, \theta_t)$
- M-Step:
 - ▶ compute θ_{t+1} as $\max_\theta Q(\theta, \theta_t)$ (M-step)
- Repeat until estimate stabilizes (converges)
- Maximizing $I(\theta)$ is the “same” as maximizing $Q(\theta, \theta_t)$
 - ▶ it can be shown that the Kullback-Leibler-Divergence does not increase in the E-step
 - ▶ holds, e.g. if data are MAR and parameter of interest ψ are separate from the parameter(s) η which govern missingness ($\theta = (\psi, \eta)$)

Examples

- $Y \sim N(\mu, \psi)$ with $\psi = \sigma^2$
- Full Maximum likelihood estimator

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{\psi} = \frac{1}{n} \sum_{i=1}^n y_i^2 - \hat{\mu}^2$$

- EM-Iterations

$$\hat{\mu}_{(t+1)} = \frac{1}{n} \left(\sum_{\text{non-miss}} y_i + n_{\text{miss}} \hat{\mu}_{(t)} \right)$$

$$\hat{\psi}_{(t+1)} = \frac{1}{n} \left(\sum_{\text{non-miss}} y_i^2 + n_{\text{miss}} (\hat{\psi}_{(t)} + \hat{\mu}_{(t)}^2) \right) - \hat{\mu}_{(t+1)}^2$$

- $Y \sim N(\mu, \Sigma)$ bivariate with second variable has missing values
- Full loglikelihood

$$-\frac{n}{2} \log |2\pi\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)$$

- Parameters to estimate: μ_i, σ_{ij} with $i, j = 1, 2$
- decompose $f_i(y_1, y_2) = f_i(y_1, \mu_1, \sigma_{11})f_i(y_2, y_1, \beta_0, \beta_1, \sigma_r)$ for missing observations
- maximization and integrations leads to a linear regression between y_1 and y_2

- the following relationship is used

$$\beta_1 = \sigma_{12}/\sigma_{11} \text{ (slope)}$$

$$\beta_0 = \mu_2 - \beta_1 \mu_1 \text{ (intercept)}$$

$$\sigma_r = \sigma_{22} - \sigma_{12}^2/\sigma_{11} \text{ (std.dev. residuals)}$$

- and can be estimated in closed form
 - $\hat{\mu}_1$ and $\hat{\sigma}_{11}$ with standard ML estimator
 - $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_r$ only from non-missing observations
- e.g. μ_2 is estimated by

$$\hat{\mu}_2 = \frac{1}{n} \left(\sum_{\text{non-miss}} y_{i2} + \sum_{\text{miss}} \hat{y}_{i2} \right)$$

- Two binary variables with missings on both variables

	parameter		observed		
	$Y_2 = 0$	$Y_2 = 1$	$Y_2 = 0$	$Y_2 = 1$	
$Y_1 = 0$	θ_{11}	θ_{12}	n_{11}	n_{12}	n_{1+}
$Y_1 = 1$	θ_{21}	θ_{22}	n_{21}	n_{22}	n_{2+}
			n_{+1}	n_{+2}	n

- Full loglikelihood

$$I(\theta) = n_{11} \log(\theta_{11}) + n_{12} \log(\theta_{12}) + n_{21} \log(\theta_{22}) + n_{22} \log(\theta_{22})$$

- Full Maximum likelihood estimator

$$\hat{\theta}_{ij} = \frac{n_{ij}}{n}$$

- n_{ij}^A non-missing observation
- n_{i+}^B missing observations in Y_2
- n_{+j}^C missing observations in Y_1
- EM-Iterations

$$\theta_{ij}^{(t+1)} = \frac{1}{n} \left(n_{ij}^A + n_{i+}^B \frac{\theta_{ij}^{(t)}}{\theta_{i+}^{(t)}} + n_{+j}^C \frac{\theta_{ij}^{(t)}}{\theta_{+j}^{(t)}} \right)$$

Data augmentation

- Idea of data augmentation
 - ▶ EM algorithm delivers an estimate $\hat{\theta}$
 - ▶ $\hat{\Theta}$ is random $\implies \hat{\Theta}$ has a distribution
 - ▶ draw m “plausible” $\hat{\theta}$ ’s from the distribution
- Markov-Chain-Monte-Carlo (MCMC) method
 - ▶ get a start estimate $\hat{\theta}_i^{(0)}$
 - ▶ impute from $\hat{\theta}_i^{(k)}$ and get a complete dataset
 - ▶ recompute an estimate $\hat{\theta}_i^{(k+1)}$ and repeat
 - ▶ $\hat{\theta}_i^{(k)}$ iterates through the distribution of $\hat{\theta}$
 - ▶ choose m “plausible” $\hat{\theta}$ (imputed datasets)

A. P. Dempster N. M. Laird, D. B. Rubin (1977). “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38. issn: 00359246. url: <http://www.jstor.org/stable/2984875>.

Tanner, Martin A. and Wong, Wing Hung (June 1987). “The Calculation of Posterior Distributions by Data Augmentation”. In: *Journal of the American Statistical Association* 82.398, p. 528. issn: 01621459. doi: 10.2307/2289457. url: <http://www.jstor.org/stable/2289457?origin=crossref> (visited on 08/22/2016).

NORM

- Assumes data are multivariate normal distributed

$$X \sim N(\mu, \Sigma)$$

- Estimate μ and Σ , e.g. with EM algorithm
- Impute values by random draws from a multivariate normal distribution with estimated μ and Σ
- Reestimate μ and Σ until estimates converge
- Performs well in practice, even for ordinal data
- package *Amelia* (after [Amelia Earhart](#))
 - ▶ Schafer developed a windows program NORM



Listing 9.7: example_amelia.R

```
1 # run example_mar.R before
2 library("Amelia")
3 library("mitools")
4 # run NORM
5 aobj <- amelia(xmar, noms=c("chas", "rad"))
6 # compute linear regressions
7 models <- lapply(aobj$imputations, function(x) {
8   lm(medv~lstat, data=x)
9 })
10 # look at one model
11 summary(models[[1]])
12 # extract
13 beta <- MIextract(models, fun="coef")
14 vcov <- MIextract(models, fun="vcov")
15 summary(MIcombine(beta, vcov))
16 summary(MIcombine(models))
```



```
Amelia:::amelia(x, m=5, parallel=c("no", "multicore", "snow"), noms,
                 ords)
```

Chained equation

- multiple imputations by chained equation
 - ▶ applies iteratively regression methods for imputing each variable

numeric	various linear regression models, mean/class-mean imputation, quadratic term regression model
factor, 2 levels	logistic regression
factor, ≥ 2 levels	polytomous logistic regression
ordered,	linear discriminant analysis
any	proportional odds model
	predictive mean matching, sample

- theoretical properties are not fully explored, but performs well in practice

 Listing 9.8: example_mice.R

```
1 # run example_mar.R before
2 source(list.files(pattern='example_mar.R', recursive=TRUE)[1])
3 #
4 library("mice")
5 # run NORM
6 xmar$chas <- factor(xmar$chas)
7 xmar$rad <- factor(xmar$rad)
8 mobj <- mice(xmar)
9 # compute linear regressions
10 models = list()
11 for (i in 1:5) models[[i]] <- lm(medv~lstat,
12                                     data=complete(mobj, i))
13 # look at one model
14 summary(models[[1]])
```

```
④ mice::mice(data, m=5, method=vector("character",length=ncol(data)),
              predictorMatrix=(1-diag(1,ncol(data))))
④ mice::complete(x, action=1)
```

Combine estimates

- multiple imputation estimate can easily combined by averaging
- allows also for variance estimation

$$\hat{\theta}_{MI} = \frac{1}{m} \sum_i \hat{\theta}_i$$

$$\widehat{Var}(\hat{\theta}_{MI}) = \underbrace{\frac{1}{m} \sum_i \widehat{Var}(\hat{\theta}_i)}_{= \text{within}} + \left(1 - \frac{1}{m}\right) \underbrace{\frac{1}{m-1} \sum_i (\hat{\theta}_i - \hat{\theta}_{MI})^2}_{= \text{between}}$$

 Listing 9.9: example_mi.R

```
1 # run example_mar.R/example_mice.R before
2 source(list.files(pattern='example_mar.R', recursive=TRUE)[1])
3 source(list.files(pattern='example_mice.R', recursive=TRUE)[1])
4 #
5 library("mitools")
6 # extract
7 beta <- MIextract(models, fun="coef")
8 vcov <- MIextract(models, fun="vcov")
9 summary(MIcombine(beta, vcov))
10 summary(MIcombine(models))
```

- ⌚ mitools::imputationList(datasets, ...)
- ⌚ mitools::MIextract(results, expr, fun)
- ⌚ mitools::MIcombine(results, ...)

Hints for imputation

- Don't round off imputations for dummy variables. May introduce bias in estimation process.
- Don't transform skewed variables. Changes relationship between variables, might impute outliers.
- Use more imputations. Rule-of-thumb: number of imputations = percent of missing values.
- Create multiplicative terms before imputing. Creating the multiplicative terms after imputation may bias the regression parameters of the multiplicative term.
- Alternatives to multiple imputation aren't usually better. Other techniques (e.g. listwise deletion) impose more stringent assumptions than MAR.

Source: www.theanalysisfactor.com/multiple-imputation-5-recent-findings-that-change-how-to-use-it

Subgroup analysis

November 3, 2022

- Subgroup analysis • Stacked bar plot • Grouped bar plot • Mosaic plot • Conditional densities • Population pyramid • Error bar diagramm • Two sample Gauss test • Two sample t-test • p sample median test • Mann-Whitney U test • ANOVA • Kruskal-Wallis H test • Paired two sample t -test • Wilcoxon signed-rank test • Friedman test • Permutation or exact tests • Omnibus test • Post-Hoc test • Bonferroni correction • Least Significant Difference test • Bonferroni corrected LSD test • Tukey Honest Significant Differences test • Student-Newman-Keuls & Duncan test • Scheffé test • Homogeneous subgroups • Spread level plot • F test • Levene test

Subgroup analysis

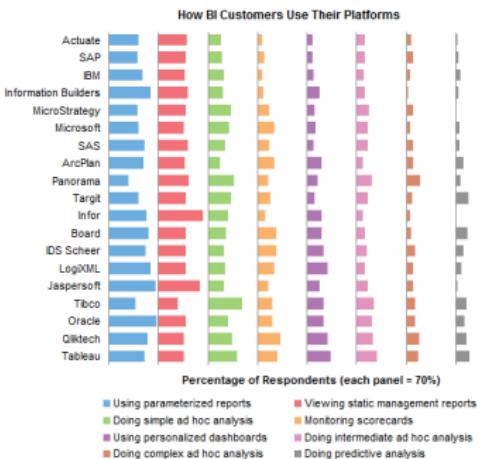
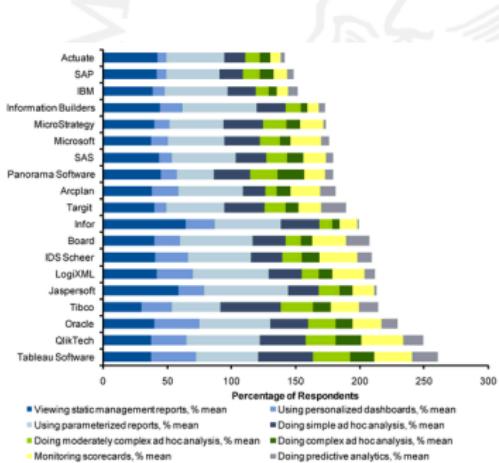
- Excel via ... IF functions
 - ▶ counting
 - ☒ COUNTIF(range, criteria)
 - ☒ COUNTIFS(range1, criterial1, range2, criteria2, ...)
 - ▶ mean
 - ☒ AVERAGEIF(range, criteria , average_range)
 - ☒ AVERAGEIFS(average_range, range1, criterial1, range2, criteria2 ...)
- SPSS
 - ▶ direct support, e.g. boxplot
 - ▶ apply analysis only to one subset of observations
 - ☒ Daten → Fälle auswählen
 - ▶ apply analysis to subsets of observations
 - ☒ Daten → Datei aufteilen
- R: either direct support or use
 - ☒ tapply(X, INDEX, FUN=NULL, ..., simplify=T)

R Listing 10.1: subgroup_example.R

```
1 data(Boston, package="MASS")
2 boxplot(medv~rad, data=Boston, notch=T, varwidth=T)
3 #
4 tapply(Boston$medv, factor(Boston$rad), mean)
```

Stacked bar plot

- Visualization: distributions
 - ▶ two categorical variables
 - ▶ unlimited number of observations
- Problems:
 - ▶ with too much categories per bar a comparison is difficult



Source: <http://peltiertech.com/WordPress/stacked-bar-chart-alternatives/>

⌚ Listing 10.2: example_barchart_stacked_graphics.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$chas, Boston$rad)
3 barplot(tab)
```

⌚ Listing 10.3: example_barchart_stacked_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 tab2 <- table(Boston$rad, Boston$chas)
4 barchart(tab2)
```

⌚ Listing 10.4: example_barchart_stacked_ggplot.R

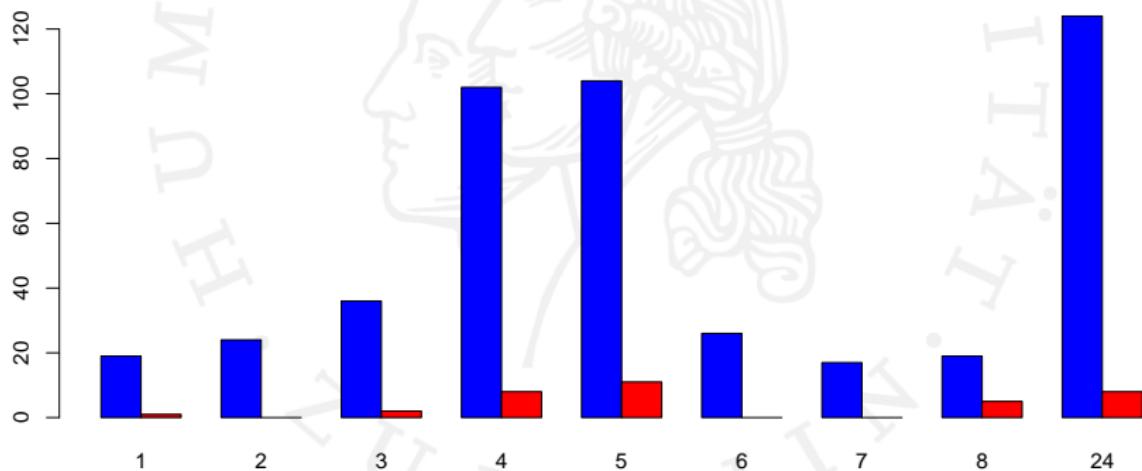
```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 d <- ggplot(Boston, aes(x=factor(rad), fill=factor(chas)))
4 d + geom_bar(position="stack")
```

⌚ barplot(tab2)

⌚ lattice::barchart(tab2)

Grouped bar plot

- Visualization: distributions / conditional distributions
 - ▶ two categorical variables
 - ▶ unlimited number of observations





Listing 10.5: example_barchart_grouped_graphics.R

```
1 library("MASS") # for Boston Housing data
2 tab <- table(Boston$chas, Boston$rad)
3 barplot(tab, beside=T)
```



Listing 10.6: example_barchart_grouped_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 tab2 <- table(Boston$rad, Boston$chas)
4 ldat <- data.frame(rad=rep(rownames(tab2), 2),
5                      count=as.vector(tab2),
6                      chas=c(rep(colnames(tab2)[1],9),
7                             rep(colnames(tab2)[2],9)))
8 barchart(count~rad, group=chas, data=ldat, ylim=c(0, 5+max(ldat$count)))
```

⌚ Listing 10.7: example_barchart_grouped_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 d <- ggplot(Boston, aes(x=factor(rad),fill=factor(chas)))
4 d + geom_bar(position=position_dodge())
```

⌚ barplot(tab2, beside=T)

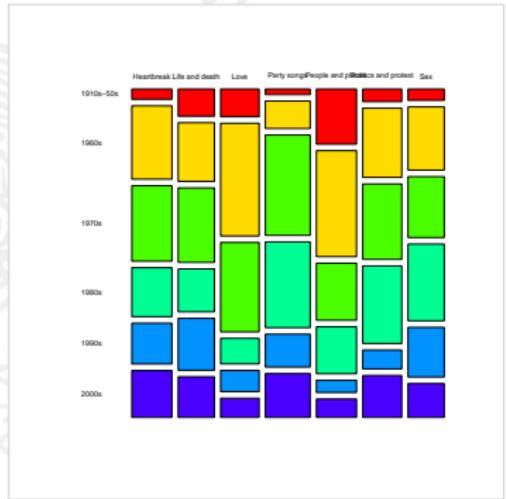
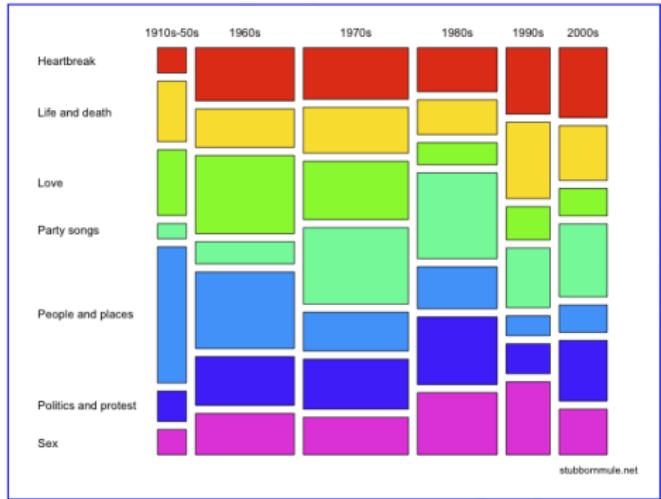
⌚ lattice::barchart(formula, data, group)

Mosaic plot

- Hartigan, Kleiner, (1981), Minard (1844)
- Visualization: association
 - ▶ Variables: two (or more) categorical variables
 - ▶ Observations: unlimited
- Plot for two variables:
 - ▶ width of a block: proportional to song frequency in a decade
 - ▶ height of a block: proportional to conditional song frequency theme given by decade
 - ▶ area of block: proportional to the song frequency by theme and decade
- Problems: order of the variables

Minard, C.J. (1844). *Tableaux figuratifs de la circulation de quelques chemins de fer.*

Hartigan, J. A. and Kleiner, B. (1981). "Mosaics for Contingency Tables". In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. Ed. by William F. Eddy. New York, NY: Springer US, pp. 268–273. isbn: 978-0-387-90633-1 978-1-4613-9464-8. url: http://link.springer.com/10.1007/978-1-4613-9464-8_37 (visited on 08/26/2015).



Joint and marginal frequency distributions

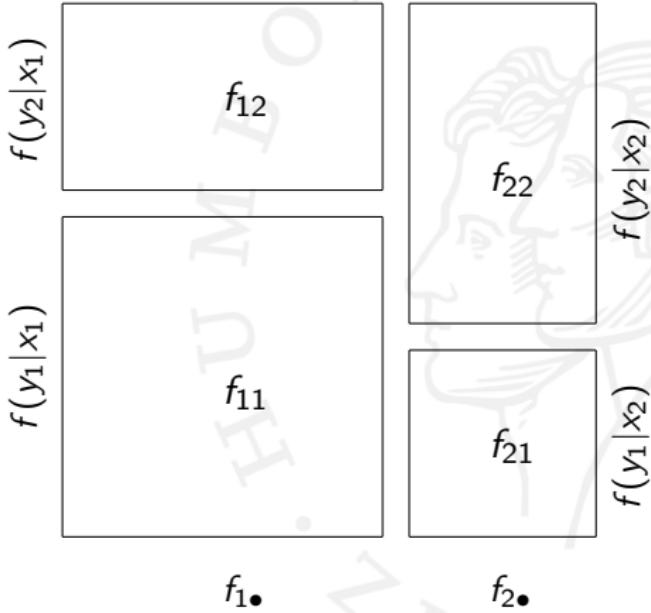
$X \setminus Y$	0	1	
0	f_{11}	f_{12}	$f_{1\bullet}$
1	f_{21}	f_{22}	$f_{2\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	1

Conditional frequency distributions (conditioned on X)

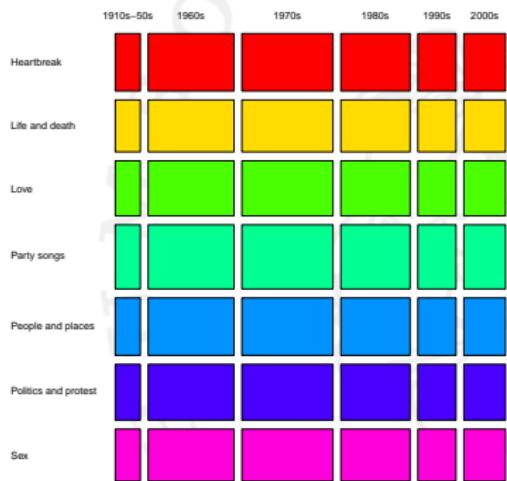
	0	1	
0	$f(y_1 x_1) = f_{11}/f_{1\bullet}$	$f(y_2 x_1) = f_{12}/f_{1\bullet}$	1
1	$f(y_2 x_1) = f_{21}/f_{2\bullet}$	$f(y_2 x_2) = f_{22}/f_{2\bullet}$	1

Conditional frequency distributions (conditioned on Y)

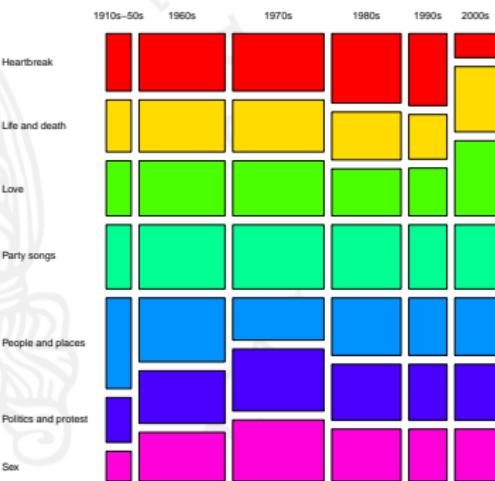
	0	1	
0	$f(x_1 y_1) = f_{11}/f_{\bullet 1}$	$f(x_1 y_2) = f_{12}/f_{\bullet 2}$	
1	$f(x_2 y_1) = f_{21}/f_{\bullet 1}$	$f(x_2 y_2) = f_{22}/f_{\bullet 2}$	
	1	1	



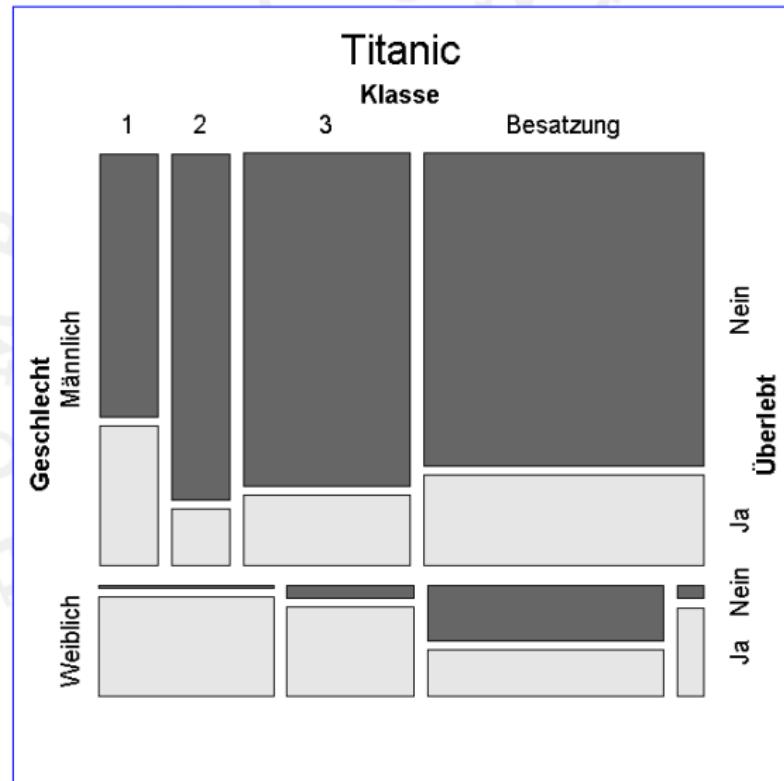
- area = height \times width
- $f_{ij} = f(y_j|x_i) \times f_{i\bullet}$
- if X and Y are independent it holds $f(y_i|x_1) = f(y_i|x_2)$
- if X and Y are independent the horizontal gaps should be at same height
- two conditional frequencies are possible (conditioned on X and Y)
- order of categorical variables plays a role



Full independence



Partial independence



⌚ Listing 10.8: example_mosaic_plot.R

```
1 plot(Titanic)
```

⌚ Listing 10.9: example_mosaic_graphics.R

```
1 mosaicplot(Titanic)
```

⌚ `plot(table)`

⌚ `mosaicplot(table, type=c("pearson", "deviance", "FT"))`

② Listing 10.10: example_mosaic_vcd.R

```
1 library("vcd")
2 mosaic(Titanic)
```

② Listing 10.11: example_mosaic_cotabplot.R

```
1 library("vcd")
2 cotabplot(Titanic)
```

② vcd::mosaic(table, direction)

② vcd::cotabplot(table)

Conditional densities

- Visualization: conditional distribution
 - ▶ Variables: at least one continuous and one categorical variable
 - ▶ Observations: unlimited
- Compute a smoothed estimate for each group such that holds

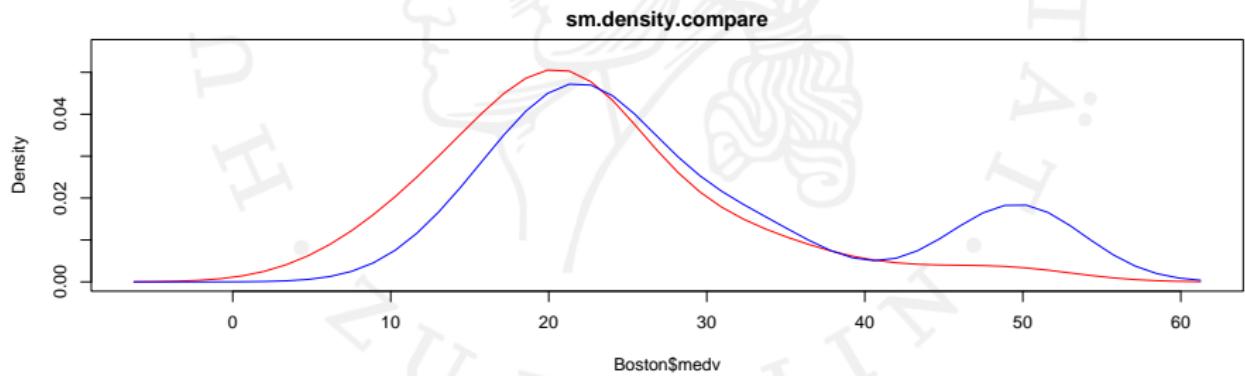
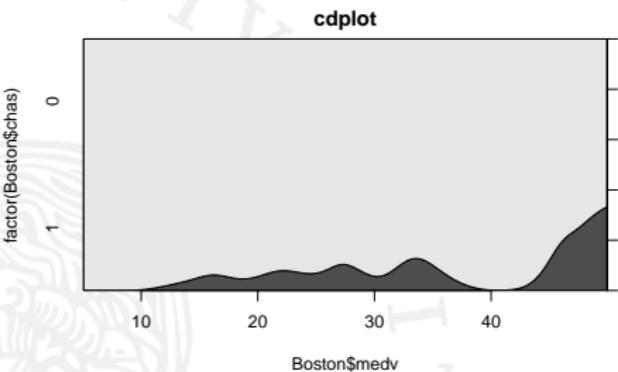
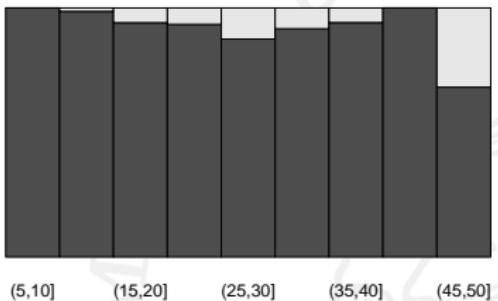
$$\sum_{k=1}^K f_{h,k}(x) = 1$$

- Compute a kernel density estimate for each group such that holds

$$\int_{-\infty}^{+\infty} f_{h,k}(x) dx = 1 \quad (k = 1, \dots, K)$$

Bowman, A. W. and Azzalini, Adelchi (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Oxford statistical science series 18. Oxford : New York: Clarendon Press ; Oxford University Press. 193 pp. isbn: 978-0-19-852396-3.

Hofmann, Heike and Theus, Martin (2005). "Interactive graphics for visualizing conditional distributions".



④ Listing 10.12: example_cdplot.R

```
1 library("MASS") # for Boston Housing data
2 cdplot(Boston$medv, factor(Boston$chas))
```

④ Listing 10.13: example_sm_density_compare.R

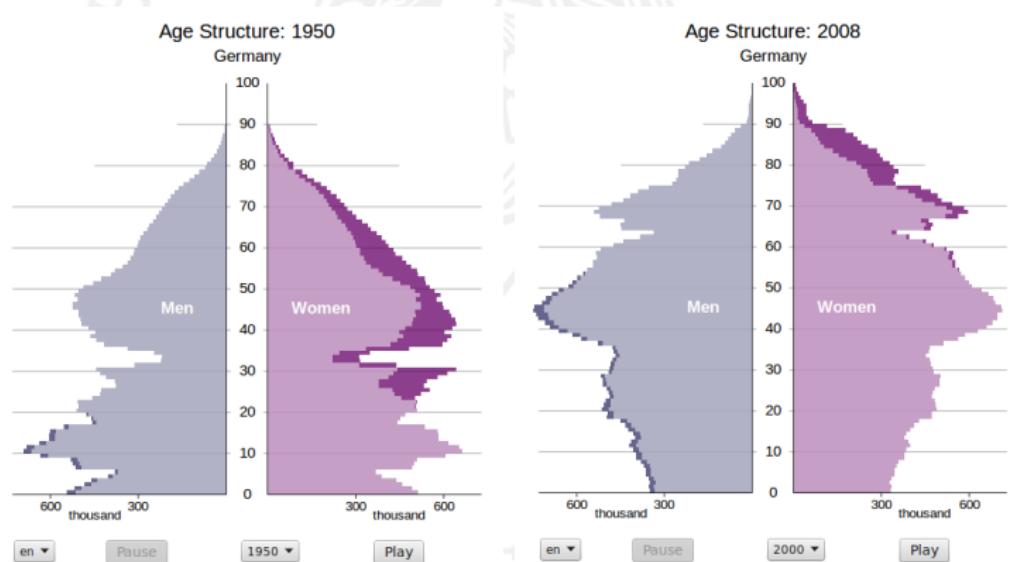
```
1 library("MASS") # for Boston Housing data
2 library("sm")
3 sm.density.compare(Boston$medv, factor(Boston$chas),
4         col=c("red", "blue"), lty=c("solid", "solid"), lwd=2)
5 legend("topright", legend=c("CHAS==0", "CHAS==1"),
6         col=c("red", "blue"), lwd=2)
```

④ `cdplot(x, y, plot = TRUE, bw = "nrd0", n = 512)`

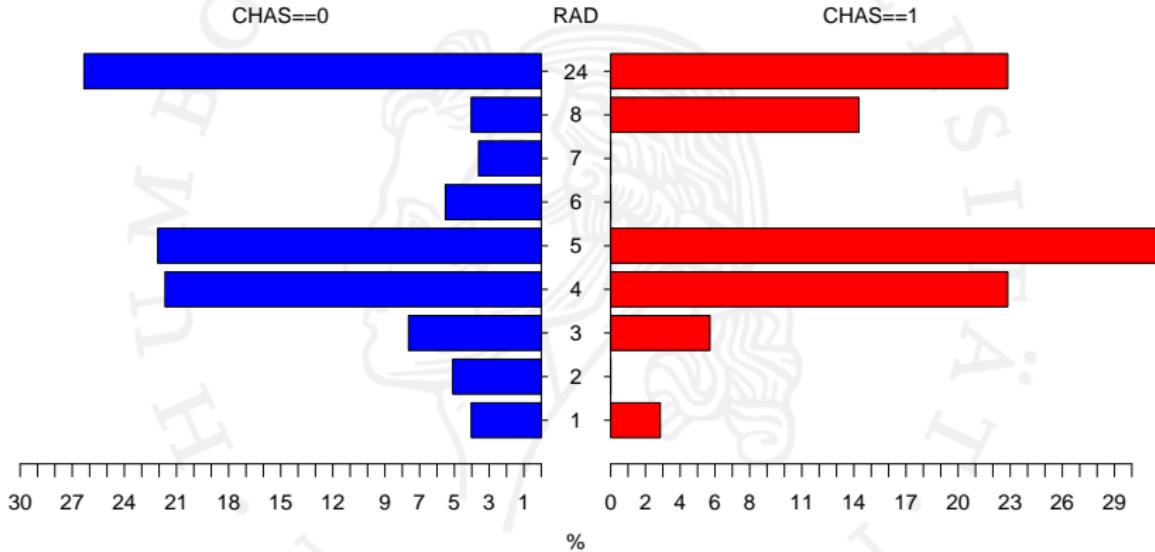
④ `sm::sm.density.compare(x, group, h)`

Population pyramid

- Visualization: distributions
 - ▶ one categorical or continuous variable or one binary variable
 - ▶ unlimited number of observations
- software often implements a population pyramid for age and sex



Source : <https://www.destatis.de/bevoelkerungspyramide/>





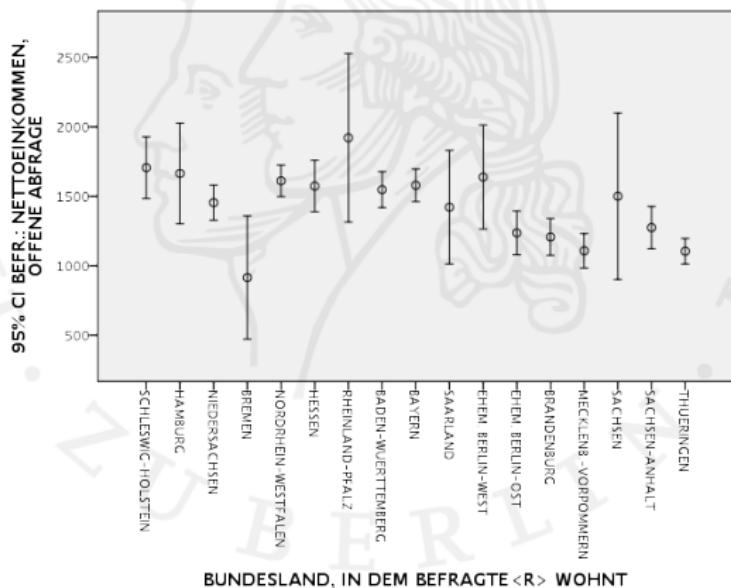
Listing 10.14: example_pyramid.R

```
1 library("MASS") # for Boston Housing data
2 library("plotrix")
3 tab <- table(Boston$rad, Boston$chas)
4 tab <- 100*sweep(tab, 2, colSums(tab), "/")
5 pyramid.plot(tab[,1], tab[,2],
6               top.labels=c("CHAS==0", "RAD", "CHAS==1"),
7               labels=rownames(tab), gap=2,
8               lxcoll="blue", rxcol="red")
```

④ `plotrix::pyramid.plot(xx, xy, labels=NA,`
 `top.labels=c("Male", "Age", "Female"))`

Error bar diagramm

- Show mean values for each group including the $1 - \alpha$ confidence interval
- Non over-lapping confidence intervals hint to significant different means



⌚ Listing 10.15: example_plotmeans.R

```
1 library("MASS")
2 library("gplots")
3 plotmeans(medv~rad, data=Boston, connect=F)
```

⌚ gplots::plotmeans(formula, data, p=0.95, connect=T, bars=T)

Two sample Gauss test

Assumptions: $X_{k,i}$ metric and independent ($k = 1, 2$), σ_k known
and either $X_{k,i} \sim N(\mu_k; \sigma_k)$ or
 $X_{k,i} \sim (\mu_k; \sigma_k)$, $n_i > 30$ (CLT)

Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistics: $V = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_D} \approx N(0; 1)$

$$\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Reject H_0 :
 $|v| > z_{1-\alpha/2}$
 $v < -z_{1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \geq 0$
 $v > +z_{1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \leq 0$

Remark: $D = \bar{X}_1 - \bar{X}_2 \approx N(\mu_D; \sigma_D)$

Effect sizes: $d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{(n_1\sigma_1^2 + n_2\sigma_2^2)/(n_1 + n_2)}}$

$d = 0.2$ small, $d = 0.5$ medium, $d = 0.8$ large effect

$$r = \sqrt{\frac{d^2}{d^2 + \frac{(n_1+n_2)^2}{n_1 n_2}}}$$

$r = 0.1$ small, $r = 0.3$ medium, $r = 0.5$ large effect

Two sample t-test

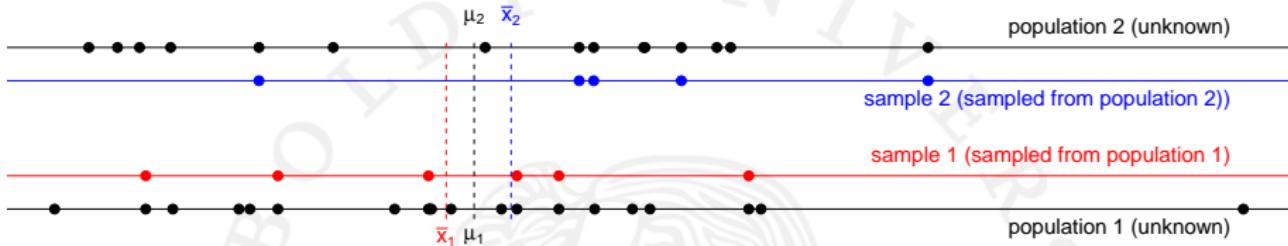
Assumptions: $X_{k,i}$ metric and independent ($k = 1, 2$)
 $\sigma_1 = \sigma_2$ unknown and either $X_{k,i} \sim N(\mu_k; \sigma_k)$ or
 $X_{k,i} \sim (\mu_k; \sigma_k)$, $n_i > 30$ (CLT)

Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistics: $V = \frac{(\bar{X}_1 - \bar{X}_2)}{S_D} \approx t_{n_1+n_2-2} \stackrel{n_1+n_2-2>30}{\approx} N(0; 1)$
 $S_D^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

Reject H_0 : $|v| > t_{n_1+n_2-2; 1-\alpha/2}$
 $v < -t_{n_1+n_2-2; 1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \geq 0$
 $v > +t_{n_1+n_2-2; 1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \leq 0$

Remark The test is also called Welch test



Effect sizes: $d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)}}$

$d = 0.2$ small, $d = 0.5$ medium, $d = 0.8$ large effect

$$r = \sqrt{\frac{d^2}{d^2 + \frac{(n_1+n_2)^2}{n_1 n_2}}}$$

$r = 0.1$ small, $r = 0.3$ medium, $r = 0.5$ large effect

Remark: Effect sizes for the two sample t test and Welch test are the same

Assumptions: $X_{k,i}$ metric and independent ($k = 1, 2$)
 $\sigma_1 \neq \sigma_2$ unknown and either $X_{k,i} \sim N(\mu_k; \sigma_k)$ or
 $X_{k,i} \sim (\mu_k; \sigma_k)$, $n_i > 30$ (CLT)

Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistics: $V = \frac{(\bar{X}_1 - \bar{X}_2)}{S_D} \approx t_f \stackrel{n > 30}{\approx} N(0; 1)$
 $S_D^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$, $f = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$

Reject H_0 : $|v| > t_{f;1-\alpha/2}$
 $v < -t_{f;1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \geq 0$
 $v > +t_{f;1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \leq 0$

Remark: The test is also called Welch test

Welch, B. L. (Jan. 1947). "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved". In: *Biometrika* 34.1, p. 28. issn: 00063444. doi: 10.2307/2332510. url: <http://www.jstor.org/stable/2332510?origin=crossref> (visited on 12/10/2015).

⌚ Listing 10.16: twottest_example.R

```
1 data(Boston, package="MASS")
2 t.test(Boston$medv[Boston$chas==0] ,
3        Boston$medv[Boston$chas==1])
4 t.test(medv~chas, Boston)
```

```
⌚ t.test(x, y, var.equal=T, alternative=c("two.sided", "less",
   "greater"))

⌚ t.test(x, y, alternative=c("two.sided", "less", "greater"),
   var.equal=F)
```

p sample median test

Assumption(s): $X_{k,i}$ is continuous metric ($k = 1, \dots, p$), independent

Hypotheses: $H_0 : \tilde{\mu}_1 = \dots = \tilde{\mu}_p$ vs.

H_1 : it exist at least one pair (i, j) with $\tilde{\mu}_i \neq \tilde{\mu}_j$

Test statistics: $V = \sum_{k=1}^p \frac{(n_{k-} - n_k/2)^2}{n_k/2} + \frac{(n_{k+} - n_k/2)^2}{n_k/2} \approx \chi^2_{p-1}$

\tilde{x} median about all observations x_{ki}

$n_{j-} = \#\{x_{ki} < \tilde{x}\}, n_{j+} = \#\{x_{ki} > \tilde{x}\}$

Reject H_0 : $v > \chi^2_{p-1; 1-\alpha}$

- basically a χ^2 independence test:

group	1	\dots	p
below total median	n_{1-}	\dots	n_{p-}
above total median	n_{1+}	\dots	n_{p+}
total	n_1	\dots	n_p

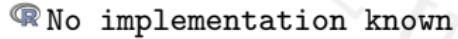
- the test has a bad power

Dixon, W. J. and Mood, A. M. (Dec. 1946). "The Statistical Sign Test". In: *Journal of the American Statistical Association* 41.236, p. 557. issn: 01621459. doi: 10.2307/2280577. url: <http://www.jstor.org/stable/2280577?origin=crossref> (visited on 08/14/2015).

Freidlin, Boris and Gastwirth, Joseph L. (Aug. 2000). "Should the Median Test be Retired from General Use?" In: *The American Statistician* 54.3, pp. 161–164. issn: 0003-1305, 1537-2731. doi: 10.1080/00031305.2000.10474539. url: <http://www.tandfonline.com/doi/abs/10.1080/00031305.2000.10474539> (visited on 08/14/2015).

 Listing 10.17: example_p_median_test.R

```
1 p.median.test <- function (x, f) {
2     med <- median(x)
3     lst <- tapply(x, f, function(v, med) { c(length(v), sum(v<med), sum(v>=med), sum(v>med)) })
4     tab <- do.call('rbind', lst)
5     n0 <- tab[,1]-tab[,2]-tab[,3]
6     as.table(tab[,2:3]+n0/2)
7 }
8 #
9 data(Boston, package="MASS")
10 tab <- p.median.test(Boston$medv, Boston$rad)
11 tab
12 chisq.test(tab)
```

 No implementation known

Mann-Whitney U test

Assumption(s): $X_{k,i}$ independent, $k = 1, 2$, at least ordinal

It holds $F_1(x + a) = F_2(x)$

Hypotheses: $H_0 : a = 0$ vs. $H_1 : a \neq 0$

Test statistics: $V = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \approx N(m_u; s_u)$

R_1 rank sum of sample 1

$$m_u = \frac{n_1 n_2}{2}, s_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

- $n_1 > 3, n_2 > 3$ and $n > 25$ the normal approximations can be used
- otherwise critical values are tabulated
- in case of ties the computation of s_u becomes more complicated

⚠ rejection of H_0 implies that $\mu_1 \neq \mu_2$ and $\tilde{\mu}_1 \neq \tilde{\mu}_2$

Wilcoxon, Frank (Dec. 1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6, p. 80. issn: 00994987. doi: 10.2307/3001968. url: <http://www.jstor.org/stable/10.2307/3001968?origin=crossref> (visited on 06/09/2016).

Mann, H. B. and Whitney, D. R. (Mar. 1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1, pp. 50–60. issn: 0003-4851. doi: 10.1214/aoms/1177730491. url: <http://projecteuclid.org/euclid.aoms/1177730491> (visited on 06/09/2016).

R Listing 10.18: example_mann_whitney.R

```
1 data(Boston, package="MASS")
2 wilcox.test(Boston$medv[Boston$chas==0],
3             Boston$medv[Boston$chas==1])
4 wilcox.test(medv~chas, Boston)
```

R wilcox.test(x, y, alternative=c("two.sided", "less", "greater"),
exact=NULL, correct=TRUE)

ANOVA

Assumption(s): $X_{k,i} \sim N(\mu_k, \sigma)$ with $k = 1, \dots, p$

Hypotheses: $H_0 : \mu_1 = \dots = \mu_p$ vs. $H_1 : \text{at least one pair } \mu_i \neq \mu_j$

Test statistics: $V = \frac{(n - p)SS_{\text{between}}}{(p - 1)SS_{\text{within}}} \sim F_{p-1; n-p}$

$$SS_{\text{between}} = \sum_{k=1}^p n_k \bar{X}_k^2 - n \bar{X}^2$$

$$SS_{\text{within}} = \sum_{k=1}^p \sum_{i=1}^{n_k} X_{k,i}^2 - \sum_{k=1}^p n_k \bar{X}_k^2$$

Remark: ANOVA = ANalysis Of VAriance

⚠ If $n_1 = \dots = n_p$ then the ANOVA is robust against departure from the assumptions!

$$\begin{aligned}(n - 1)S^2 &= \sum_{k=1}^p \sum_{i=1}^{n_k} (X_{k,i} - \bar{X})^2 \\&= \sum_{k=1}^p \sum_{i=1}^{n_k} (X_{k,i} - \bar{X}_k + \bar{X}_k - \bar{X})^2 \\&= \underbrace{\sum_{k=1}^p n_k \bar{X}_k^2 - n \bar{X}^2}_{=SS_{between}} + \underbrace{\sum_{k=1}^p \sum_{i=1}^{n_k} X_{k,i}^2 - \sum_{k=1}^p n_k \bar{X}_k^2}_{=SS_{within}} \\&= \sum_{k=1}^p n_k \bar{X}_k^2 - n \bar{X}^2 + \sum_{k=1}^p (n_k - 1) S_k^2.\end{aligned}$$

data variation = variation between groups + variation within group

1. Assumption: $\sigma_k = \sigma$
2. General: $Var(X) = E(X^2) - E^2(X) \Rightarrow E(X^2) = \sigma_X^2 + \mu_X^2$

$$E((n-1)S^2) = (n-1)\sigma^2$$

$$E(SS_{within}) = \sum_{k=1}^p (n_k - 1)\sigma_k^2 \stackrel{1.}{=} (n-p)\sigma^2$$

$$E(SS_{between}) = \sum_{k=1}^p n_k E(\bar{X}_k^2) - nE(\bar{X}^2)$$

$$\stackrel{2.}{=} \sum_{k=1}^p n_k \left(\frac{\sigma_k^2}{n_k} + \mu_k^2 \right) - n \left(\frac{\sigma^2}{n} + \mu^2 \right)$$

$$\stackrel{1.}{=} (p-1)\sigma^2 + \underbrace{\sum_{k=1}^p n_k (\mu_k^2 - \mu^2)}_{=0 \text{ under } H_0, >0 \text{ under } H_1}$$

- Violating the independence assumption
 - ▶ ANOVA results are difficult to interpret
 - ▶ Mean difference can result from dependence
 - ▶ Try to achieve $n_1 = \dots = n_p$
- Distribution form is not normal
 - ▶ If data are not unimodal then is the mean to right parameter to analyze?
 - ▶ If the data in the groups are all skewed to the same direction then ANOVA can be applied (maybe a transformation is necessary before)
 - ▶ Use the group medians instead of the group means
 - ▶ Apply a non-parametric test
- Variances are heterogeneous
 - ▶ If $\frac{\text{largest variance}}{\text{smallest variance}} < 4$ then ANOVA can be applied
 - ▶ Otherwise transform data or apply a non-parametric test

Regression model

$$Y_j = \beta_0 + \sum_{i=1}^I \beta_i w_{ji} + E_j$$

ANOVA model

$$X_{ki} = \mu_k + E_{ki}$$

ANOVA as regression model

$$X_{ki} = \underbrace{\mu_I}_{=\beta_0} + \sum_{\substack{k=1 \\ k \neq I}}^P (\underbrace{\mu_k - \mu_I}_{=\beta_k}) \underbrace{I(\text{obs. } i \text{ in group } k)}_{=w_{ki}} + E_{ki}$$

- F-Test (=ANOVA) in linear regression checks whether
 - ▶ $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$
 - ▶ $H_0 : \text{all } \beta_k = 0$ vs. $H_1 : \text{at least one } \beta_k \neq 0$
- F-Test in linear regression with continuous variables can be considered as ANOVA with only one element in each group



Listing 10.19: example_anova.R

```
1 data(Boston, package="MASS")
2 # check number of observations
3 tapply(Boston$medv, Boston$rad, length)
4 # check normality
5 tapply(Boston$medv, Boston$rad, shapiro.test)
6 # check variances
7 tapply(Boston$medv, Boston$rad, var)
8 # check skewness
9 library("DescTools")
10 tapply(Boston$medv, Boston$rad, Skew)
11 # ANOVA :
12 summary(aov(medv~rad, data=Boston))
```



aov(y~f)

Kruskal-Wallis H test

Assumption(s): $X_{k,i}$ independent, $k = 1, \dots, p$ at least ordinal, $n_j > 4$

Hypotheses: H_0 : the mean ranks in each group are the same

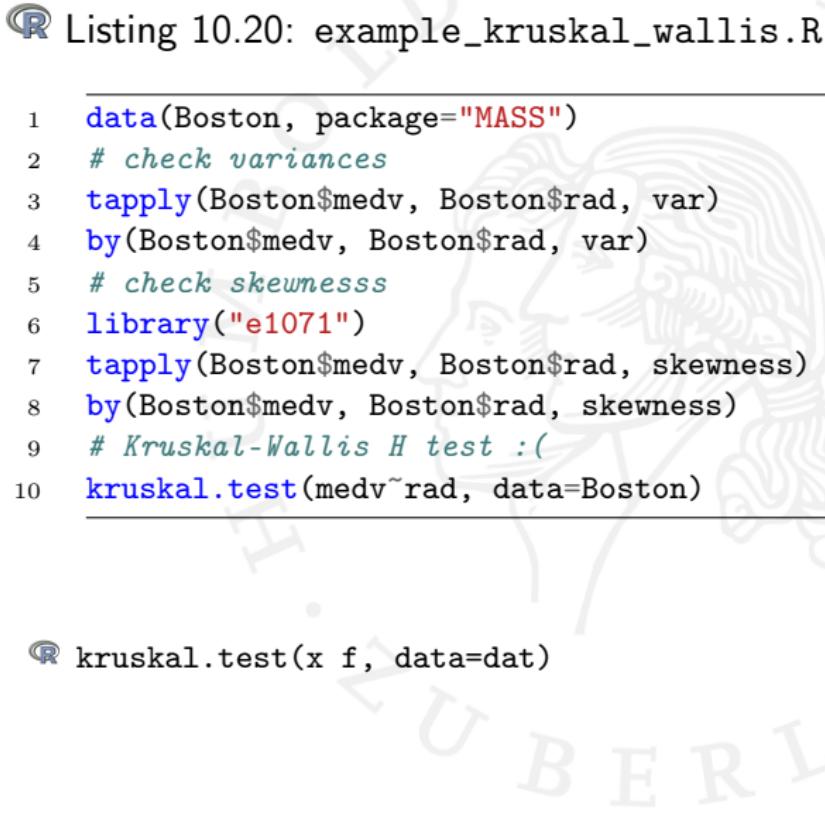
H_1 : the mean ranks in each group are not the same

Test statistics: $V = \frac{12}{n(n+1)} \sum_{j=1}^p n_j \left(R_j - \frac{n+1}{2} \right)^2 \approx \chi_{p-1}^2$

R_j rank sum of of group j

- other approximation conditions are possible
- only if the groups distributions have the “same” shape then the rejection of H_0 implies that the means μ_i and the medians $\tilde{\mu}_i$ can not be all equal

Kruskal, William H. and Wallis, W. Allen (Dec. 1952). "Use of Ranks in One-Criterion Variance Analysis". In: *Journal of the American Statistical Association* 47.260, pp. 583–621. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1952.10483441. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441> (visited on 06/09/2016).

R Listing 10.20: example_kruskal_wallis.R

```
1 data(Boston, package="MASS")
2 # check variances
3 tapply(Boston$medv, Boston$rad, var)
4 by(Boston$medv, Boston$rad, var)
5 # check skewnesss
6 library("e1071")
7 tapply(Boston$medv, Boston$rad, skewness)
8 by(Boston$medv, Boston$rad, skewness)
9 # Kruskal-Wallis H test :
10 kruskal.test(medv~rad, data=Boston)
```

```
R kruskal.test(x f, data=dat)
```

Paired two sample t -test

Assumptions: $X_{k,i}$ metric and dependent ($k = 1, 2$)

$D_i = X_{1,i} - X_{2,i} \approx N(\mu_1 - \mu_2; \sigma_D)$ or $n > 30$ (CLT)

D_i independent

Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistics: $V = \frac{\bar{D}}{S_D} \approx t_{n-1}$

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

Reject H_0 : $|v| > t_{n-1; 1-\alpha/2}$

$v < -t_{n-1; 1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \geq 0$

$v > +t_{n-1; 1-\alpha}$ if $H_0 : \mu_1 - \mu_2 \leq 0$

⌚ Listing 10.21: example_paired_ttest.R

```
1 # 100 m running times before new training
2 b <- c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
3 # 100 m running times after new training
4 a <- c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
5 #
6 t.test(b, a, paired=TRUE)
```

⌚ `t.test(x, y, paired=T, alternative=c("two.sided", "less", "greater"))`

Wilcoxon signed-rank test

Assumption(s): $X_{k,i}$ dependent, $k = 1, 2$ at least metric,

$D_i = X_{1,i} - X_{2,i}$ independent, symmetric and continuous, $n > 20$

Hypotheses: $H_0 : \tilde{\mu}_D = 0$ vs. $H_1 : \tilde{\mu}_D \neq 0$

Test statistics: $V = R^+ \approx N\left(\frac{n(n+1)}{4}; \sigma = \frac{n(n+1)(2n+1)}{24}\right)$

R_i rank of $|D_i|$,

R^+ rank sum of positive D_i ,

R^- rank sum of negative D_i

- if the approximation conditions not fulfilled then the critical values of $\min(R^+, R^-)$ are tabulated
- is robust against violations of the assumptions of D_i

④ Listing 10.22: example_wilcoxon.R

```
1 # 100 m running times before new training
2 b <- c(12.98, 13.54, 12.85, 15.67, 17.24, 19.23, 12.63, 15.35, 14.48,
3 # 100 m running times after new training
4 a <- c(12.72, 13.63, 12.05, 15.23, 16.88, 20.02, 12.07, 15.92, 16.09,
5 #
6 wilcox.test(b, a, paired=TRUE)
```

④ `wilcox.test(x, y, alternative=c("two.sided", "less", "greater"),
paired=TRUE, exact=NULL, correct=TRUE)`

Friedman test

Assumption(s): $X_{k,i}$ at least ordinal, $X_{\bullet,i}$ dependent
 $k = 1, \dots, p$, $i = 1, \dots, n$

Hypotheses: $H_0 : \tilde{\mu}_1 = \dots = \tilde{\mu}_p$
 $H_1 : \text{at least for one pair } (r, s) \text{ holds } \tilde{\mu}_r \neq \tilde{\mu}_s$

Test statistics: $S_R = \sum_{k=1}^p \left(R_k - \frac{n(p+1)}{2} \right)^2$
 $R_k = \frac{1}{n} \sum_{i=1}^n R_{ki}$

with R_{ki} the rank of the observation x_{ki} in x_{1i}, \dots, x_{pi}
 for small n and p use tabulated values

for $np \leq 40$: $\frac{12n(p-1)(S_R-1)}{n^2(p^3-p)+24} \approx \chi_{p-1}^2$

for $40 \leq np \leq 60$: $U = \frac{12S_R}{np(p+1)} \approx \chi_{p-1}^2$

for $np > 60$: $\frac{(n-1)U}{n(p-1)-U} \approx F_{p-1;(n-1)(p-1)}$

- Idea
 - ▶ assign all values from one observation a rank from 1 to k
 - ▶ under the null the mean rank R_k for each observation should be constant
- in case of ties the test statistics must be modified by

$$U_c = \frac{U}{1 - \frac{1}{np(p^2-1)} \sum_j (t_j^3 - t_j)}$$

- Friedman is a non-parametric ANOVA for dependent data

Friedman, Milton (Dec. 1937). "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". In: *Journal of the American Statistical Association* 32.200, p. 675. issn: 01621459. doi: 10.2307/2279372. url: <http://www.jstor.org/stable/2279372?origin=crossref> (visited on 06/09/2016).

R Listing 10.23: example_friedman.R

```
1 dat <- data.frame(t1 = c(1, 2, 5, 3, 2, 1, 1, 3, 2 ,1),
2
3
4 friedman.test(as.matrix(dat))
```

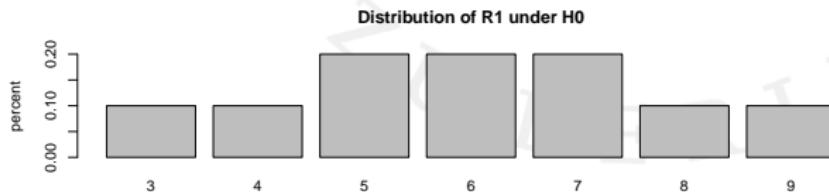
R friedman.test(mat)

Permutation or exact tests

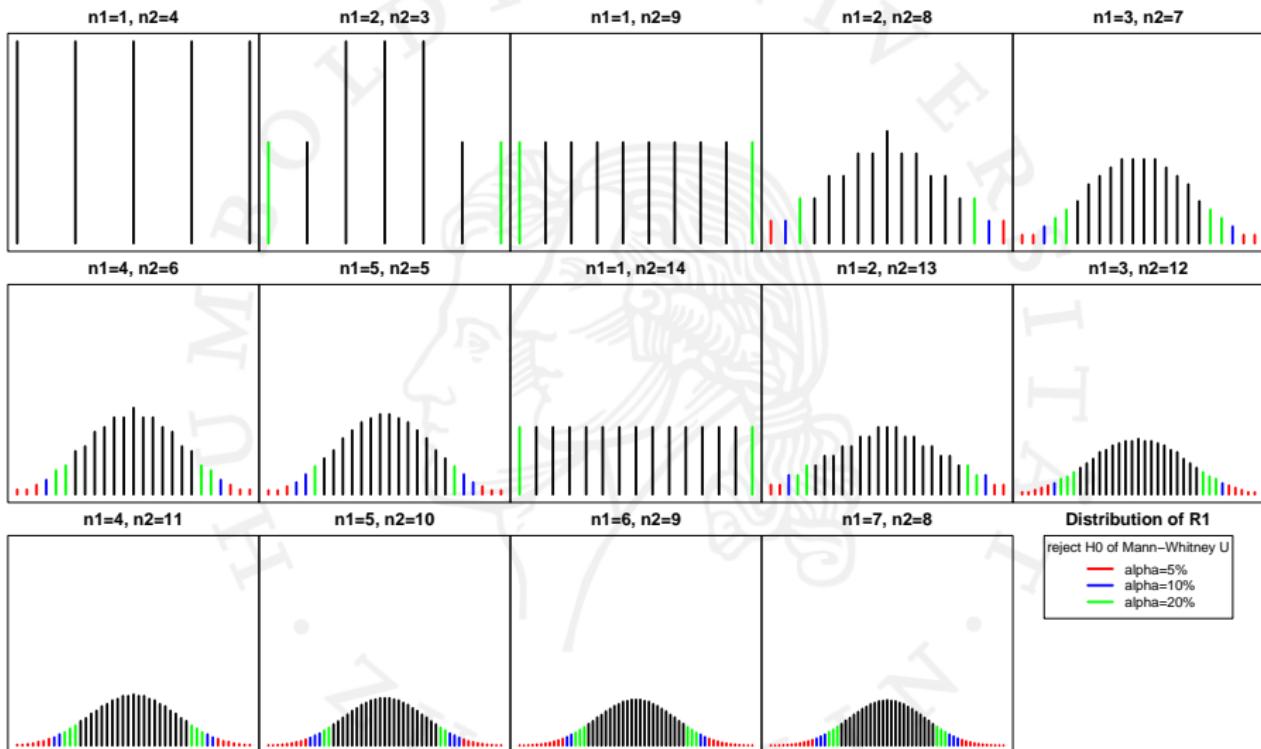
- Permutation or exact test
 - ▶ A finite number of observations is assigned a label
 - ▶ Under H_0 the labels can be rearranged
 - ▶ The distribution of the test statistics under H_0 can be computed by all “permutations” of the labels
- Permutation tests:
 - ▶ Mann-Whitney U test
 - ▶ Kruskal-Wallis H test
 - ▶ Friedman test

Example 10.17

- Mann-Whitney U test for $n_1 = 2$ and $n_2 = 3$
- for $\alpha = 20\%$ reject H_0 if $r_1 = 3$ or $r_1 = 9$



Labels	R_1
11222	3
12122	4
12212	5
12221	6
21122	5
21212	6
21221	7
22112	7
22121	8
22211	9



Omnibus test

- Omnibus test
 - ▶ Tests whether a difference between $p > 2$ populations exist
 - H_0 : there is no difference between the populations
 - H_1 : between at least two populations exist a difference
 - ▶ Does not say *which* populations are different
- Examples
 - ▶ ANOVA
 - ▶ Kruskal-Wallis H test
 - ▶ Levene test
- Problem
 - ▶ the *global* null hypothesis (e.g. $\mu_1 = \dots = \mu_p$) is composed by *local* null hypotheses (e.g. $\mu_i = \mu_j$)
 - ▶ a rejection of one local null hypothesis should result in a rejection of the global null hypothesis

Post-Hoc test

- Tests for the local hypotheses
 - ▶ Simple solution for ANOVA: use e.g. an error bar plot to view the differences
- for the ANOVA are at least two types of Post hoc tests available
 - ▶ equal variances in all groups
 - ▶ unequal variances in at least two groups
- Tests which assume homogenous variances
 - ▶ tests based on pairwise t test
 - ★ Least Significant Difference (LSD)
 - ★ Bonferroni
 - ▶ tests based on variation width
 - ★ Tukey
 - ★ Student-Newman-Keuls
 - ★ Duncan
 - ★ Scheffé

Bonferroni correction

- For ANOVA we could do pairwise t-tests between group means
- Problem: how to choose the global significance level based on multiple tests with a “local” significance level?
- How to compute a local α to get a globally correct α_G :

$$\begin{aligned}1 - \alpha_G &= P(\text{"H}_0\text{:global"} | H_0) \\&= P\left(\bigcap_{k=1}^g (\text{"H}_{0,k"} | H_{0,k})\right) \\&= (1 - \alpha)^g \\ \alpha &= 1 - \sqrt[g]{1 - \alpha_G} \\ \alpha &\approx \alpha_G/g \text{ (more conservative approach)}\end{aligned}$$

Bonferroni, C. E. (1936). “Teoria Statistica Delle Classi e Calcolo Delle Probabilità”. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8, pp. 3–62.

Example 10.18

With

- $g = 4$ comparisons
- global significance level
 $\alpha_G = 5\%$ for the ANOVA

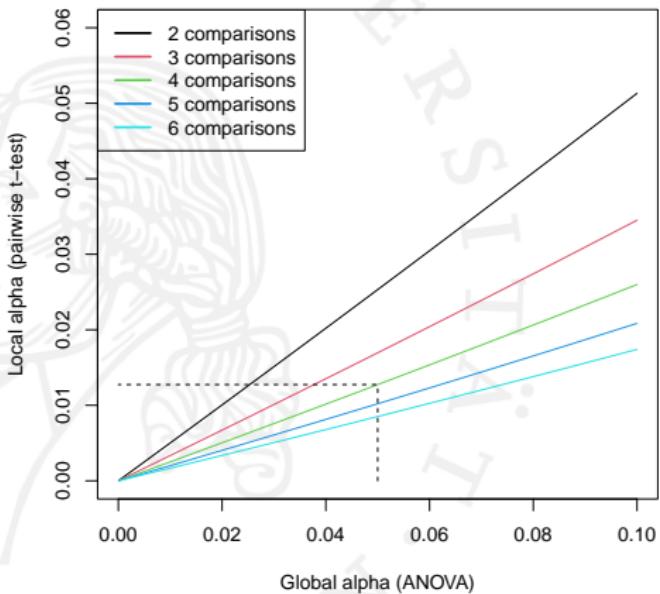
choose the local significance level for the pairwise t -tests

$$\alpha = 1 - \sqrt[4]{1 - 0.05} = 0.01274$$

or

$$\alpha = 0.05/4 = 0.0125$$

$$1 - \sqrt[g]{1 - \alpha_G} = \alpha_G/g$$



Least Significant Difference test

- pairwise comparison of group means
- based on two sample t-test for differences in mean

$$\frac{\bar{x}_i - \bar{x}_j}{s_d \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n-p; 1-\alpha/2}$$

$$s_d^2 = \frac{1}{n-p} \sum_{j=1}^p (n_j - 1) s_j^2$$

- the variance is estimated from *all* groups since we assume the variance in all groups equal
- α are *not* Bonferroni corrected
- test should be used only for assumed differences

R Listing 10.24: example_lsd.R

```
1 data(Boston, package="MASS")
2 # ANOVA
3 fit <- aov(medv~rad, data=Boston)
4 # LSD test
5 library("agricolae")
6 LSD.test(fit, "rad", console=T)
```

R `agricolae::LSD.test(y, trt, alpha=0.05,
 p.adj=c("none", "holm", "hommel", "hochberg",
 "bonferroni", "BH", "BY", "fdr"), group=TRUE,
 console=FALSE)`

⚠ y is the result of the R function `aov`

Bonferroni corrected LSD test

- as LSD test, but α is corrected using the Bonferroni inequality

$$P\left(\bigcup_{k=1}^g H_{0;k}\right) \leq \sum_{k=1}^g P(H_{0;k})$$

- choose $\alpha = \alpha_G/g$
- should only be applied if the number of groups is small

R Listing 10.25: example_bonferroni.R

```
1 data(Boston, package="MASS")
2 # ANOVA
3 fit <- aov(medv~rad, data=Boston)
4 # LSD test
5 library("agricolae")
6 LSD.test(fit, "rad", p.adj="bonferroni", console=T)
```

```
agrيلae::LSD.test(y, trt, p.adj="bonferroni", alpha=0.05,
group=TRUE, console=FALSE))
```

⚠ y is the result of the R function aov

Tukey Honest Significant Differences test

- $H_0 : \mu_{(i)} = \dots = \mu_{(i+k)}$ vs. $H_1 : \text{at least one mean is different}$

$$\frac{|\bar{x}_{(i)} - \bar{x}_{(i+k)}|}{s_d / \sqrt{n_i}} \sim Q_{p;n-p}$$

- the quantiles of $Q_{p;n-p}$ are tabulated
- only for balanced case $n_1 = \dots = n_p$
- the global significance niveau is ensured by $Q_{p;n-p}$
- the test procedure is used iteratively
 - ▶ start with $(\bar{x}_{(1)}, \bar{x}_{(p)})$
 - ▶ continue with $(\bar{x}_{(1)}, \bar{x}_{(p-1)})$ and $(\bar{x}_{(2)}, \bar{x}_{(p)})$
 - ▶ repeat

R Listing 10.26: example_tukey.R

```
1 data(Boston, package="MASS")
2 # ANOVA
3 Boston$rad      <- as.factor(Boston$rad)
4 fit <- aov(medv~rad, data=Boston)
5 # HSD test
6 TukeyHSD(fit, "rad")
7 #
8 library("agricolae")
9 HSD.test(fit, "rad", console=T)
```

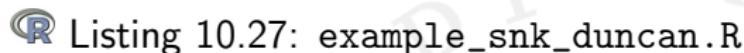
R TukeyHSD(aov, which, ordered=FALSE)

R agricolae::HSD(aov, trt, alpha=0.05, group=TRUE, console=FALSE))

⚠ aov is a result of the R function aov

Student-Newman-Keuls & Duncan test

- Student-Newman-Keuls-Test
 - ▶ $H_0 : \mu_{(i)} = \dots = \mu_{(i+k)}$ vs. $H_1 : \text{at least one mean is different}$
- Duncan-Test
 - ▶ as Student-Newman-Keuls test
 - ▶ the quantiles are corrected by $\alpha_k = 1 - (1 - \alpha)^k$
- the quantiles of $Q_{k;n-p}$ are tabulated
- only for balanced case $n_1 = \dots = n_p$
- the test procedure is used iteratively

 Listing 10.27: example_snk_duncan.R

```
1 data(Boston, package="MASS")
2 library("agricolae")
3 # ANOVA
4 Boston$rad      <- as.factor(Boston$rad)
5 fit <- aov(medv~rad, data=Boston)
6 # Student-Newman-Keuls test
7 SNK.test(fit, "rad", console=T)
8 # Duncan test
9 duncan.test(fit, "rad", console=T)
```

☞ `agricolae::SNK.test(aov, trt, alpha=0.05, group=TRUE,
 console=FALSE))`

☞ `agricolae::duncan.test(aov, trt, alpha=0.05, group=TRUE,
 console=FALSE))`

⚠ `aov` is a result of the R function `aov`

Scheffé test

- Definition linear contrast

$$\Lambda = \sum_{i=1}^p c_i \mu_i \text{ with } 0 = \sum_{i=1}^p c_i$$

- If all $\mu_i = \mu$ and (c_1, \dots, c_p) is a contrast then it follows $\Lambda = 0$

Example 10.19

Linear contrast: $\Lambda = \mu_i - \mu_j$ with $c_i = 1$, $c_j = -1$ and otherwise 0

$$H_0 : \mu_i = \mu_j \text{ vs. } H_1 : \mu_i \neq \mu_j \implies H_0 : \Lambda = 0 \text{ vs. } H_1 : \Lambda \neq 0$$

- Can be applied in the unbalanced case
- $H_0 : \Lambda = 0$ vs. $H_1 : \Lambda \neq 0$

$$\frac{\frac{1}{p-1} \left(\sum_{i=1}^p c_i \bar{x}_i \right)^2}{s_d^2 \sum_{i=1}^p \frac{c_i^2}{n_i}} \sim F_{p-1; n-p}$$

- Based on $1 = \frac{Var(\Lambda)}{Var(\Lambda)} = \frac{E(\Lambda^2) - E^2(\Lambda)}{Var(\Lambda)}$
- Checks all independent contrasts with two means
- Robust against violation of
 - ▶ homogeneous variances
 - ▶ normal distribution
- Conservative

④ Listing 10.28: example_scheffe.R

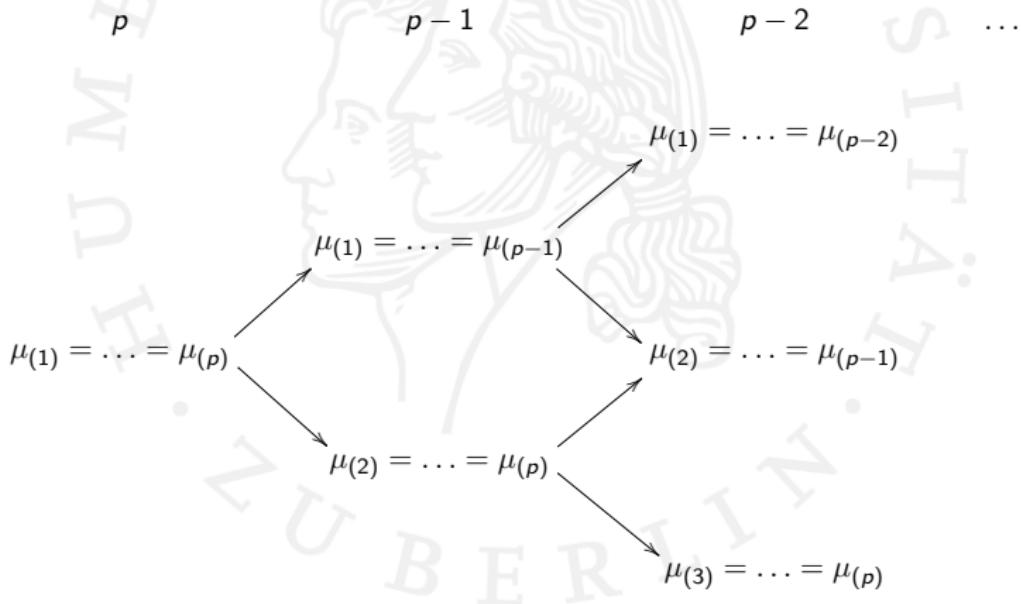
```
1 data(Boston, package="MASS")
2 # ANOVA
3 fit <- aov(medv~rad, data=Boston)
4 # Scheffé test
5 library("agricolae")
6 scheffe.test(fit, "rad", console=T)
```

④ `agricolae::scheffe.test(aov, trt, alpha=0.05, group=TRUE,
 console=FALSE)`

⚠ `aov` is a result of the R function `aov`

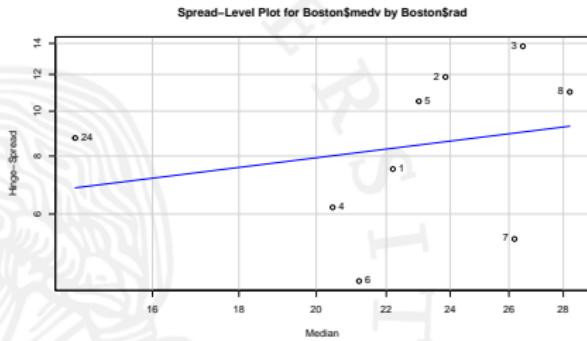
Homogeneous subgroups

- Efficient method to find homogeneous subgroups
 - ▶ sort groups by empirical means \bar{x}_i
 - ▶ test sequences of means



Spread level plot

- plot median vs. interquartile range or hinge-spread
- estimate regression line
- slope can be used to transform data
- $p = 1 - \text{slope}$
(round to nearest 0.5)



	LowerHinge	Median	UpperHinge	Hinge-Spread
24	11.15	14.40	19.90	8.75
4	17.50	20.45	23.70	6.20
6	18.80	21.20	23.10	4.30
1	20.35	22.20	27.85	7.50
5	19.50	23.00	30.00	10.50
2	21.40	23.85	33.25	11.85
7	24.30	26.20	29.60	5.30
3	21.10	26.50	34.90	13.80
8	23.65	28.25	34.65	11.00

Suggested power transformation: 0.5458092

⌚ Listing 10.29: example_spreadlevel.R

```
1 data(Boston, package="MASS")
2 library("car")
3 spreadLevelPlot(Boston$medv, by=Boston$rad)
```

⌚ car::spreadLevelPlot(x, by)

F test

Assumption(s): $X_{k,i} \sim N(\mu_k; \sigma_k)$ independent, $k = 1, 2$,

Hypotheses: $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$

Test statistics: $V = \frac{S_1^2}{S_2^2} \sim F_{n_1-1; n_2-1}$

- not robust against departure from the normal distribution (outliers)
- basis for derived tests
 - ▶ ANOVA
 - ▶ Goldfeld-Quandt-Test (heteroscedasticity)
 - ▶ Chow-Test (structural break)

⌚ Listing 10.30: example_ftest.R

```
1 data(Boston, package="MASS")
2 var.test(medv~chas, data=Boston)
```

⌚ var.test(x, y, ratio=1, alternative=c("two.sided", "less",
"greater"))

Levene test

Assumption(s): $X_{k,i} \sim (\mu_k; \sigma_k)$ independent & continuous,
 $k = 1, \dots, p$

Hypotheses: $H_0 : \sigma_1^2 = \dots = \sigma_p^2$ vs.
 $H_1 : \text{at least for one pair } \sigma_j^2 \neq \sigma_k^2$

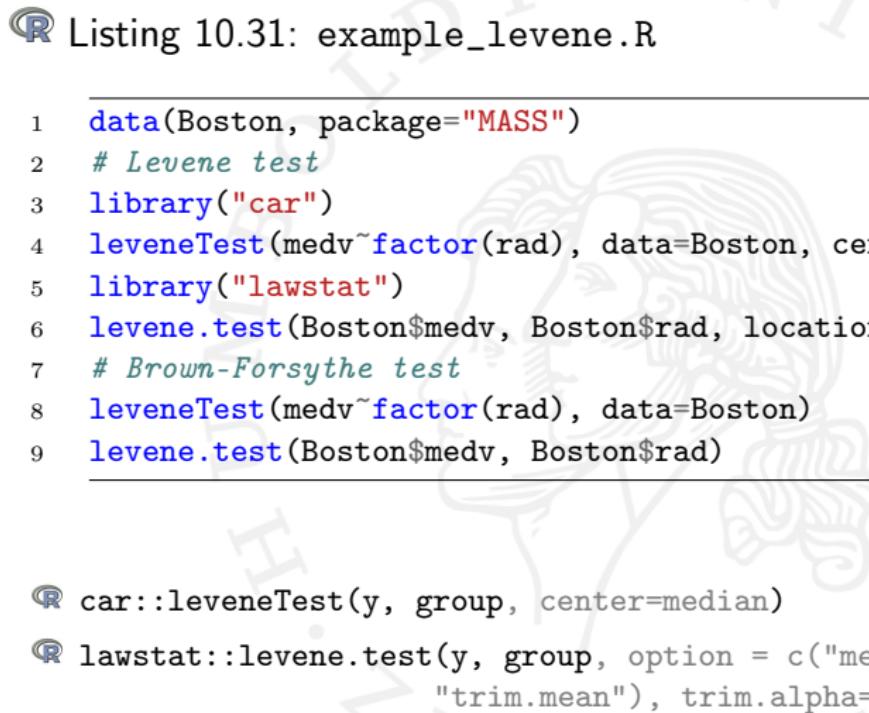
Test statistics:

$$V = \frac{n-p}{p-1} \frac{\sum_{j=1}^p n_j (\bar{Y}_j - \bar{Y})^2}{\sum_{j=1}^p \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2} \approx F_{p-1; n-p}$$

with $\bar{Y}_{ji} = |X_{ji} - \bar{X}_j|$

- Remark(s):
- The Levene test utilizes the ANOVA
 - The Brown-Forsythe test uses the group medians instead of the group means (more robust)

- Levene, H. (1960). "Robust tests for equality of variances". In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ed. by I. Olkin et al. Menlo Park, CA: Stanford University Press, pp. 278–292.
- Brown, Morton B. and Forsythe, Alan B. (June 1974). "Robust Tests for the Equality of Variances". In: *Journal of the American Statistical Association* 69.346, pp. 364–367. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1974.10482955. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482955> (visited on 06/22/2016).

A large, faint watermark-style graphic of a human head profile facing left, with internal brain structures visible.

R Listing 10.31: example_levene.R

```
1 data(Boston, package="MASS")
2 # Levene test
3 library("car")
4 leveneTest(medv~factor(rad), data=Boston, center=mean)
5 library("lawstat")
6 levene.test(Boston$medv, Boston$rad, location="mean")
7 # Brown-Forsythe test
8 leveneTest(medv~factor(rad), data=Boston)
9 levene.test(Boston$medv, Boston$rad)
```

- R car::leveneTest(y, group, center=median)
- R lawstat::levene.test(y, group, option = c("mean", "median",
 "trim.mean"), trim.alpha=0.25, bootstrap=FALSE,
 num.bootstrap=1000)

Correlation and association

November 3, 2022

- Scatterplot • Sunflowerplot • Matrix visualization • Bivariate statistics • Association • Covariance • Correlation • Steiger's Z test • Spearman's rank correlation • Ties • Test for Spearman's rank correlation • Kendall's τ
- Differences between rank correlations • Somer's D • Contingency tables
- Conditional frequencies • χ^2 based coefficients • χ^2 independence test • Fisher's exact test • Linear-by-Linear association test • PRE measures • Goodman and Kruskal's λ • Goodman and Kruskal's τ • Eta squared • Cohen's κ • Cohen's κ test • McNemar's test • Cochran-Mantel-Haenszel Test • Relative risk • Test on relative risk • Odds ratio • Test on odds ratio • Common odds ratio

Scatterplot

- Variables: two continuous variables
- Observations: unlimited
- Visualization: association
- Problems:
 - ▶ overplotting
 - ▶ not well suited for discrete variables
- rectangular coordinates: 1500 Leonardo da Vinci
- smoothed line in scatterplot: 1832 John Frederick W. Herschel
- described by Francis Galton

R Listing 11.1: scatterplot_example.R

```
1 library("spdep")
2 data("boston")
3 plot(boston.c$LON, boston.c$LAT, main="Boston school districts")
```

R plot(x, y=NULL, ...)

⚠ plot is a generic function like summary!

R title(...)

R lines(x, y=NULL, ...)

R points(x, y=NULL, ...)

R text(x, y=NULL, ...)

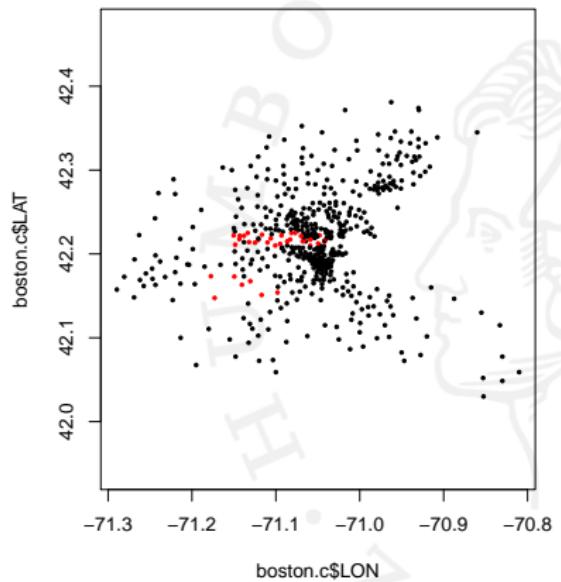
R axis(side, at=NULL, labels=T, ...)

Optional parameters for plot and friends:

- type: what to plot
- pch, cex, col: plot symbols, size and color
- lty, lwd: line type and width
- xlim, ylim: plotting range
- asp: aspect ratio
- main, main.cex, main.col, main.font: plot title, size, color and font
- sub, sub.cex, sub.col, sub.font: plot subtitle, size, color and font
- xlab, ylab, lab.cex, lab.col, lab.font: plot labels, size, color and font
- axes=T, lab.axis, lab.col, axis.font: axes plotting, size, color and font

 Listing 11.2: example_scatterplot2.R

```
1 library("spdep")
2 data("boston")
3 #
4 col <- ifelse(boston.c$CHAS==1, "red", "black")
5 plot(boston.c$LON, boston.c$LAT, main="Boston school districts", pch=19)
6 #
7 library("lattice")
8 xyplot(LAT~LON|CHAS, data=boston.c, pch=19)
9 #
10 library("ggplot2")
11 d <- ggplot(boston.c, aes(x=LON, y=LAT, colour=CHAS))
12 d + geom_point(shape=19)
```



R Listing 11.3: map_boston.R

```
1 library("spdep")
2 data("boston")
3 pdf("boston_pts.pdf", width=5, height=6)
4 plot(boston.c$LONG, boston.c$LAT, pch=19, asp=T, cex=0.4, col=boston.c$O
5 dev.off()
6 #
7 library("ggmap")
8 map <- get_map(location=c(-71.5, 42.1, -70.5, 42.6), source="osm")
9 pdf("boston_osm.pdf", width=5, height=6)
10 ggmap(map)
11 dev.off()
```

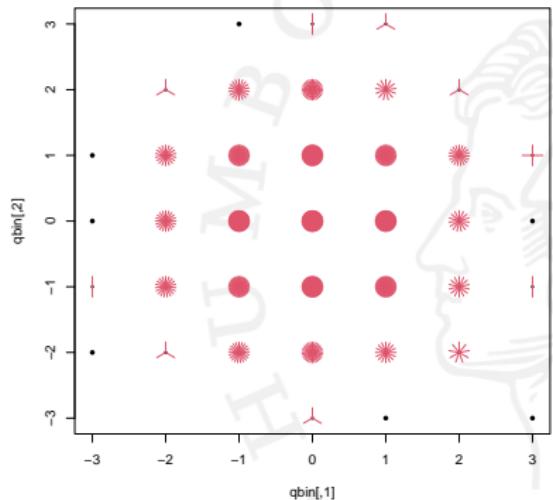
Sunflowerplot

- Cleveland, McGill (1984), Carr, Littlefield, Nicholson, Littlefield (1987)
- Visualization: association
 - ▶ Variables: two categorical variables
 - ▶ Observations: unlimited
- use symbol, size or color for the number of “hidden” values
- Problems: rounding for continuous variables
- later with hexagonal bins

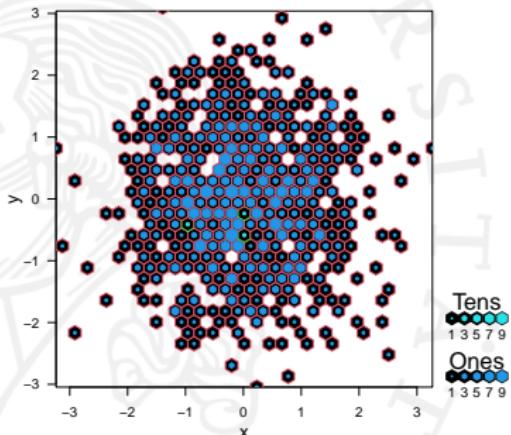
Cleveland, W.S. and McGill, R. (1984). "The Many Faces of a Scatterplot". In: *Journal of the American Statistical Association* 79.388, pp. 807–822.

Carr, D.B. et al. (1987). "Scatterplot matrix techniques for large N". In: *Journal of the American Statistical Association* 82.398, pp. 424–436.

Sunflower Plot of Rounded $N(0,1)$



Sunflower Plot of Rounded $N(0,1)$



④ Listing 11.4: example_sunflower.R

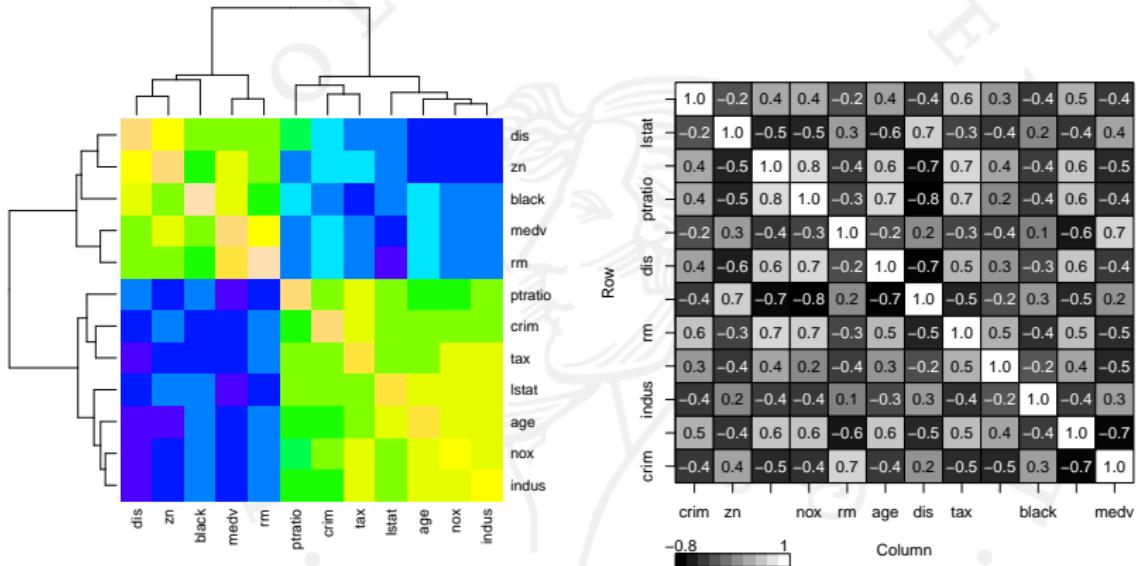
```
1 library("MASS") # for Boston Housing dat
2 Boston$lstat <- 2*round(Boston$lstat/2)
3 Boston$medv  <- 2*round(Boston$medv/2)
4 sunflowerplot(medv~lstat, data=Boston)
```

④ sunflowerplot(x, y)
④ hexbin::hexbin(x, y, xbins=30)

Matrix visualization

- Variables/observations: matrix coding
- Visualization: matrix structure
- Problem: using colors may lead to “misinterpretation”

	crim	zn	indus	nox	rm	age	dis	tax	ptratio	black	lstat	medv
crim	1.00	-0.20	0.41	0.42	-0.22	0.35	-0.38	0.58	0.29	-0.39	0.46	-0.39
zn	-0.20	1.00	-0.53	-0.52	0.31	-0.57	0.66	-0.31	-0.39	0.18	-0.41	0.36
indus	0.41	-0.53	1.00	0.76	-0.39	0.64	-0.71	0.72	0.38	-0.36	0.60	-0.48
nox	0.42	-0.52	0.76	1.00	-0.30	0.73	-0.77	0.67	0.19	-0.38	0.59	-0.43
rm	-0.22	0.31	-0.39	-0.30	1.00	-0.24	0.21	-0.29	-0.36	0.13	-0.61	0.70
age	0.35	-0.57	0.64	0.73	-0.24	1.00	-0.75	0.51	0.26	-0.27	0.60	-0.38
dis	-0.38	0.66	-0.71	-0.77	0.21	-0.75	1.00	-0.53	-0.23	0.29	-0.50	0.25
tax	0.58	-0.31	0.72	0.67	-0.29	0.51	-0.53	1.00	0.46	-0.44	0.54	-0.47
ptratio	0.29	-0.39	0.38	0.19	-0.36	0.26	-0.23	0.46	1.00	-0.18	0.37	-0.51
black	-0.39	0.18	-0.36	-0.38	0.13	-0.27	0.29	-0.44	-0.18	1.00	-0.37	0.33
lstat	0.46	-0.41	0.60	0.59	-0.61	0.60	-0.50	0.54	0.37	-0.37	1.00	-0.74
medv	-0.39	0.36	-0.48	-0.43	0.70	-0.38	0.25	-0.47	-0.51	0.33	-0.74	1.00



The logo of the University of Berlin (HU Berlin) is a watermark-like watermark in the background of the slide. It features a profile of the Roman goddess Minerva, wearing a helmet and holding a spear, with the text "HANNOVERIANA HU BERLIN" around it.

R Listing 11.5: example_heatmap1.R

```
1 library("MASS")    # for Boston Housing data
2 cor <- cor(Boston[,c(-4,-9)])
3 heatmap(cor, revC=T, col=topo.colors(12))
4 #
5 library("lattice")
6 levelplot(cor, at=(-10:10)/10)
```

```
R image(x, y, z, col=heat.colors(12), breaks)
R heatmap(x, scale=c("row", "column", "none"), revC=F, na.rm=T, zlim)
R lattice::levelplot(x, at)
R gplots::heatmap.2(x, scale=c("row", "column", "none"), revC=F,
                    na.rm=T, zlim)
R plotrix::color2D.matplot(x, show.legend=F, show.values=F, yref=T)
```



Listing 11.6: example_heatmap2.R

```
1 library("MASS")    # for Boston Housing data
2 cor <- cor(Boston[,c(-4,-9)])
3 #
4 library("ggplot2")
5 library("reshape2")
6 mcor <- melt(cor)
7 d <- ggplot(mcor, aes(x=Var1, y=Var2, z=value))
8 d <- d+geom_tile(aes(fill= value))
9 d + scale_fill_gradient(low = "white", high = "steelblue")
```

Bivariate statistics

- Two variables
 - ▶ dependence unknown: association or correlation
 - ▶ dependence known: regression or subgroup analysis
- ⚠ Fallacy: correlation implies causation
- ▶ reverse causation: The faster windmills are observed to rotate, the more wind is observed to be. Therefore wind is caused by the rotation of windmills.
 - ▶ third factor: As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning.
 - ▶ coincidence: At full moon returns for gold are no less than 5 times smaller than the returns around the new moon. Therefore, the moon phases causes gold price returns.

Lucey, Brian M. (June 2010). "Lunar seasonality in precious metal returns?" en. In: *Applied Economics Letters* 17.9, pp. 835–838. issn: 1350-4851, 1466-4291. doi: 10.1080/17446540802516188. url: <http://www.tandfonline.com/doi/abs/10.1080/17446540802516188> (visited on 08/09/2015).

Association

- Measure the association between two variables directly
 - ▶ coefficients are
 - ★ metric: Bravais-Pearson coefficient (, covariance)
 - ★ quasi-metric: tetra- and polychoric correlation
 - ★ ordinal: Kendalls τ coefficient (, Spearman)
 - ★ nominal: χ^2 , contingency coefficient(s)
 - ▶ for variables with different measurement levels we can use lower level coefficients
- Test for association with ordinal coefficients
 - ▶ do not use χ^2 independence test (information loss)
 - ▶ use instead some Z statistics $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta \neq \vartheta_0$

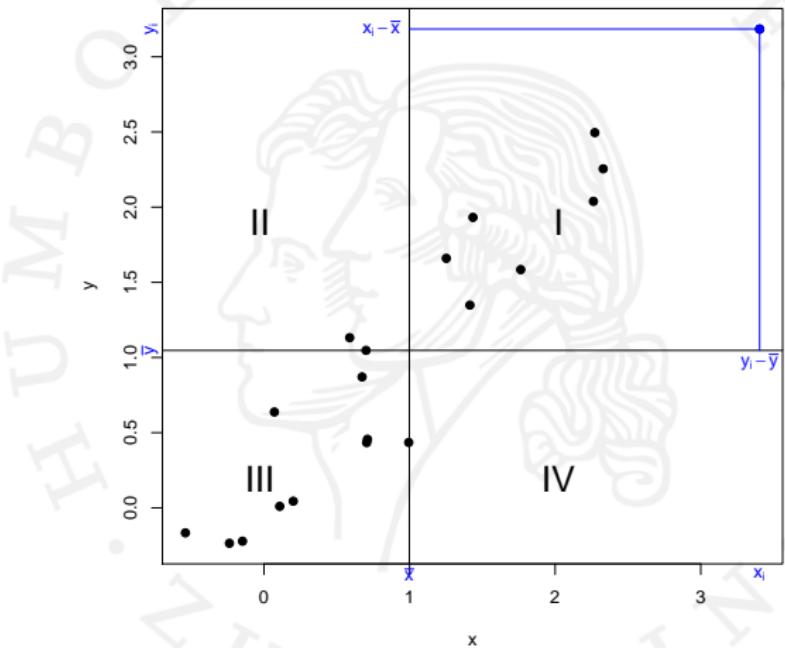
$$Z = \frac{\vartheta - \vartheta_0}{\hat{\sigma}(\hat{\theta})} \approx N(0; 1)$$

Covariance

- Covariance

$$s_{xy} = \begin{cases} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) & \text{descriptive} \\ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) & \text{estimate for } \sigma_{xy} \end{cases}$$

- $s_{xy} > 0$ positive relationship (most points are in I and III)
- $s_{xy} < 0$ negative relationship (most points are in II and IV)
- $-\infty < s_{xy} < +\infty$ difficult to judge strength of association
- measures the linear relation
- sensitive to outliers



⌚ Listing 11.7: example_covariance.R

```
1 library("MASS") # for Boston Housing data
2
3 cov(Boston)
```

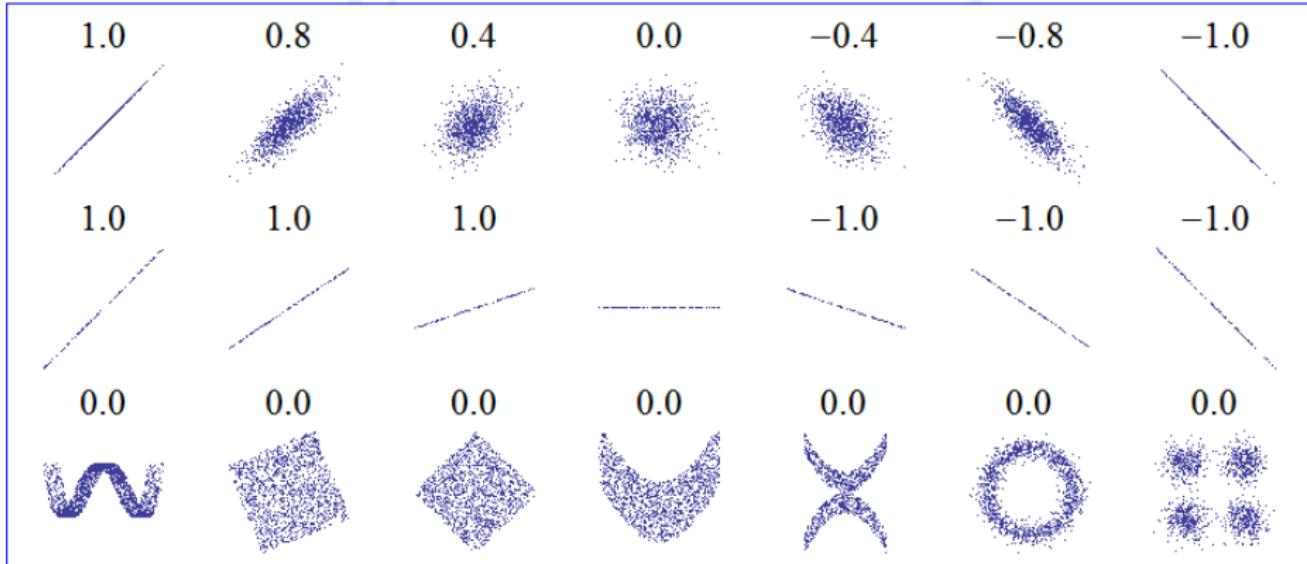
⌚ cov(x, y, use="everything")
⌚ var(x, y, na.rm=FALSE, use)
⌚ cov2cor(matrix)

Correlation

- Bravais-Pearson correlation or product moment correlation

$$\begin{aligned}
 r_{xy} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2)(\sum_{i=1}^n (y_i - \bar{y})^2)}} \\
 &= \frac{s_{xy}}{s_x s_y} \\
 &= s_{uv} \text{ with } u_i = \frac{x_i - \bar{x}}{s_x}, v_i = \frac{y_i - \bar{y}}{s_y}
 \end{aligned}$$

- invariant under linear transformation
- r_{xy} does NOT require normally distributed data
- r_{xy} is sensitive to outliers



Bravais, Auguste (1846). "Analyse Mathématique sur les Probabilités des Erreurs de Situation d'un Point". Mémoires par divers Savans à l'Académie des Sciences de l'Institut de France. In: *Sciences Mathématiques et Physiques* 9, pp. 255–332.

Galton, Francis (1888). "Co-Relations and Their Measurement, Chiefly from Anthropometric Data". In: *Proceedings of the Royal Society of London* 45, pp. 135–145. issn: 03701662. url: <http://www.jstor.org/stable/114860>.

Pearson, Karl (1895). "Note on Regression and Inheritance in the Case of Two Parents". In: *Proceedings of the Royal Society of London* 58, pp. 240–242. issn: 03701662. url: <http://www.jstor.org/stable/115794>.

Steiger's Z test

- $Z = f(R_{xy}) = 0.5 \log \left(\frac{1+R_{xy}}{1-R_{xy}} \right) = \text{atanh}(R_{xy})$
- $(X, Y) \sim N(\bullet; \bullet) \implies Z \approx N \left(f(\rho) + \frac{\rho}{n-2}; n-3 \right)$

Assumptions: (X, Y) are bivariate normal distributed

Hypotheses: $H_0 : \rho_{xy} = 0$ vs. $H_1 : \rho_{xy} \neq 0$

Test statistics: $T = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}} \sim t_{n-2}$

Reject H_0 : $|t| > t_{n-2; 1-\alpha/2}$

$t < -t_{n-2; 1-\alpha}$ if $H_0 : \rho_{xy} \leq 0$

$t > +t_{n-2; 1-\alpha}$ if $H_0 : \rho_{xy} \geq 0$

Steiger, James H. (1980). "Tests for comparing elements of a correlation matrix.". In:

Psychological Bulletin 87.2, pp. 245–251. issn: 0033-2909. doi:

10.1037/0033-2909.87.2.245. url: <http://content.apa.org/journals/bul/87/2/245>

(visited on 12/02/2015).

Check for bivariate normality of the data

- Use a test for multivariate normality; see package mvn in R
- Use a test for univariate normality:
 (X_1, \dots, X_p) is multivariate normal \Rightarrow each X_i is normal
- Note: the inverse does not hold
- If a test rejects the normality on any X_i then (X_1, \dots, X_p) is NOT multivariate normal

⌚ Listing 11.8: example_steiger.R

```
1 library("MASS")    # for Boston Housing data  
2 cor.test(Boston$lstat, Boston$medv)
```

```
⌚ cor.test(x, y, alternative=c("two.sided", "less", "greater"))  
⌚ psych::r.test(n, r, twotailed=T)
```

Spearman's rank correlation

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R(x_i) - R(y_i))^2$$

- r_s is Bravais-Pearson correlation with $(R(x_i), R(y_i))$ instead of (x_i, y_i)
 - ▶ simplification follow, e.g. from $\overline{R(x_i)} = \frac{n+1}{2}$ etc.
- recognizes also non-linear monotone relationships

Spearman, C. (Jan. 1904). "The Proof and Measurement of Association between Two Things".
In: *The American Journal of Psychology* 15.1, p. 72. issn: 00029556. doi: 10.2307/1412159.
url: <http://www.jstor.org/stable/1412159?origin=crossref> (visited on 12/02/2015).

- With ties:

$$r_s = \frac{C_x + C_y - \sum_{i=1}^n (R(x_i) - R(y_i))^2}{2\sqrt{C_x C_y}}$$

- $C_x = \frac{n^3-n}{12} - \frac{1}{12} \sum_k (t_{x,k}^3 - t_{x,k})$
- $C_y = \frac{n^3-n}{12} - \frac{1}{12} \sum_k (t_{y,k}^3 - t_{y,k})$
- $t_{\bullet,k}$ number of observations with the same rank

i	1	2	3	4	5	6	7	8	Σ
x_i	2,0	3,0	3,0	5,0	5,5	8,0	9,0	9,0	
$R(x_i)$	1,0	2,5	2,5	4,0	5,0	6,0	7,5	7,5	
$t_{x,k}^3 - t_{x,k}$	0	6	-	0	0	0	6	-	12
y_i	1,5	1,5	4,0	3,0	1,0	5,0	5,0	9,5	
$R(y_i)$	2,5	2,5	5,0	4,0	1,0	6,5	6,5	8,0	
$t_{y,k}^3 - t_{y,k}$	6	-	0	0	0	6	-	0	12

Ties

- Ties appear if a sample observations have the same value, such that no unique rank can be assigned
- We distinguish
 - ▶ Ties in X , such that $x_i = x_j$, but $y_i \neq y_j$ for $i \neq j$
 - ▶ Ties in Y , such that $y_i = y_j$, but $x_i \neq x_j$ for $i \neq j$
 - ▶ Ties in X und Y , such that $x_i = x_j$ and $y_i = y_j$ for $i \neq j$
- Handling of ties
 - ▶ assign to each value the mean rank
 - ▶ assign to each value randomly a rank
 - ▶ assign to each value a rank such that a specific hypothesis is favored
 - ▶ use mid ranks and correct your coefficient

Example 11.20

- if possible then keep the property that the sum of all ranks is $\frac{n(n+1)}{2}$

Rank y x -value	3 105	1 125	2 170	4 170	6 200	5 215	8 220	10 220	7 220	9 315	r_s
Mean rank	1	2	3.5	3.5	5	6	8	8	8	10	0.906
Random rank	1	2	3	4	5	6	8	9	7	10	0.939
Small corr. favored	1	2	4	3	5	6	7	7	9	10	0.836
	every other ranking leads to higher r_s										
$t_{x,k}$	1	1	2	–	1	1	3	–	–	1	
$t_{y,k}$	1	1	1	1	1	1	1	1	1	1	

Correction with mean rank:

$$\left. \begin{array}{l} \sum_{i=1}^n (R(x_i) - R(y_i))^2 = 15.5 \\ C_x = \frac{10^3 - 10}{12} - \frac{30}{12} = 80.0 \\ C_y = \frac{10^3 - 10}{12} - \frac{0}{12} = 82.5 \end{array} \right\} \Rightarrow r_s = \frac{80.0 + 82.5 - 15.5}{2\sqrt{80.0 \cdot 82.5}} \approx 0.905$$

Test for Spearman's rank correlation

Hypotheses: $H_0 : \rho_s = 0$ vs. $H_1 : \rho_s \neq 0$

Test statistics: R uses (Algorithm AS 89) for

- $n < 10$ a permutation approach,
- $10 \leq n < 1290$ an approximation or
- otherwise $T = \frac{R_s \sqrt{n-2}}{\sqrt{1-R_s^2}} \approx t_{n-2}$

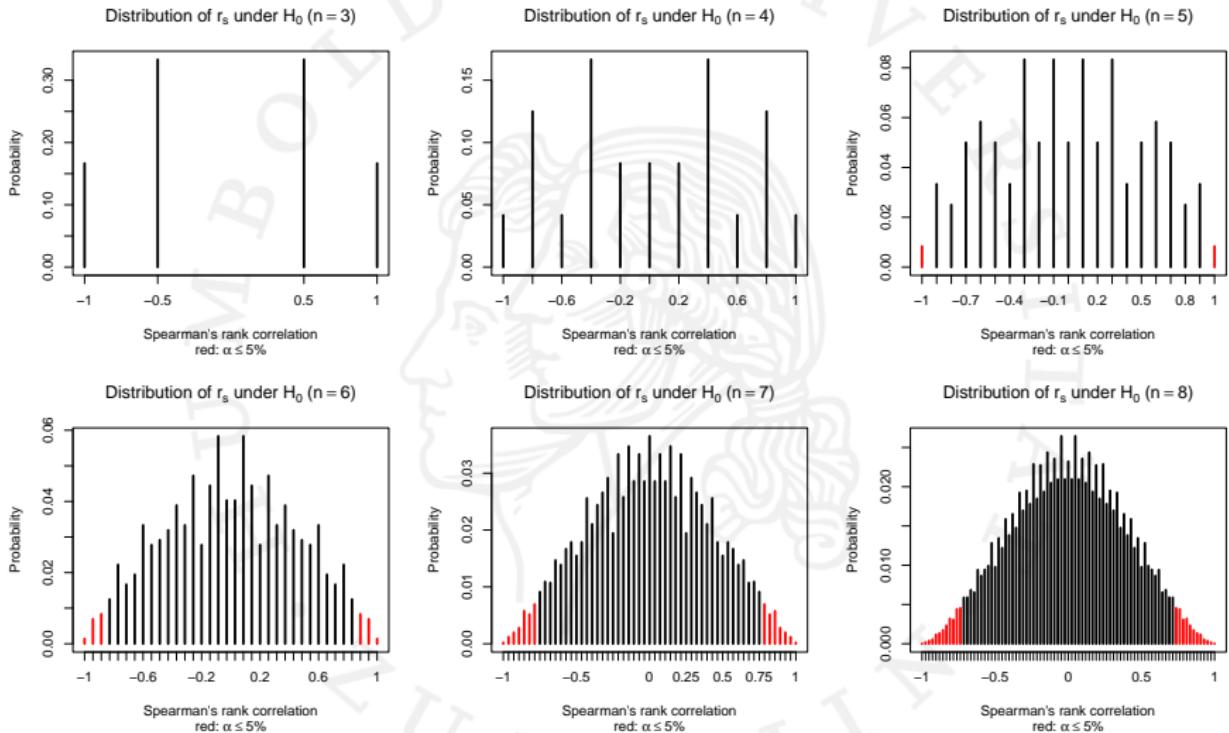
Pitman, E. J. G. (1937). "Significance Tests Which May be Applied to Samples From any Populations". In: *Supplement to the Journal of the Royal Statistical Society* 4.1, p. 119.
issn: 14666162. doi: 10.2307/2984124. url:
<http://www.jstor.org/stable/10.2307/2984124?origin=crossref> (visited on 12/02/2015).

Example 11.21

- Consider the case $n = 3$
- Renumber the data such $x_1 < x_2 < x_3$ and therefore $R(x_i) = i$
- Consider all possible permutations of ranks for y_i

i	1	2	3	r_s	t
$R(x_i)$	1	2	3		
$R(y_i)$	1	2	3	+1.0	$+\infty$
	1	3	2	+0.5	+0.577
	2	1	3	+0.5	+0.577
	2	3	1	-0.5	-0.577
	3	1	2	-0.5	-0.577
	3	2	1	-1.0	$-\infty$

- Under H_0 each permutation has the same probability (here 1/6)



R Listing 11.9: example_spearman.R

```
1 library("MASS")    # for Boston Housing data
2 cor.test(Boston$chas, Boston$rad, method="spearman")
```

- R `cor.test(x, y, method="spearman", alternative=c("two.sided", "less", "greater"))`
- R `pspearman::spearman.test(x, y, alternative=c("two.sided", "less", "greater"), approximation=c("exact", "AS89", "t-distribution"))`

 Both routines do not compute exact p-values in case of ties!

Best, D. J. and Roberts, D. E. (1975). "Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho". In: *Applied Statistics* 24.3, p. 377. issn: 00359254. doi: 10.2307/2347111. url: <http://www.jstor.org/stable/2347111?origin=crossref> (visited on 12/02/2015).

Kendall's τ

- Counts

C = number of pairs where $x_i < x_j; y_i < y_j$

D = number of pairs where $x_i < x_j; y_i > y_j$

T_x = number of pairs where $x_i = x_j; y_i < y_j$ (ties in X)

T_y = number of pairs where $x_i < x_j; y_i = y_j$ (ties in Y)

T_{xy} = number of pairs where $x_i = x_j; y_i = y_j$ (ties in X and Y)

- Number of possible pairs

$$C, D, T_x, T_y, T_{xy} \leq \sum_{i=1}^n (i-1) = \frac{n(n-1)}{2}$$

- Kendall τ_a (no ties),

$$\tau_a = \frac{C - D}{\frac{n(n-1)}{2}}$$

- Kendalls τ_b (only quadratic tables)

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

- Kendalls τ_c

$$\tau_c = \frac{2m(C - D)}{(m - 1)n^2}$$

- ▶ m = minimum of rows and columns in table

Kendall, M. G. (June 1938). "A New Measure of Rank Correlation". In: *Biometrika* 30.1, p. 81.
 issn: 00063444. doi: 10.2307/2332226. url:
<http://www.jstor.org/stable/2332226?origin=crossref> (visited on 12/02/2015).

Differences between rank correlations

- Mostly applies: $\tau \leq r_s$

$$-1 \leq 3\tau - 2r_s \leq 1$$

- r_s is based on Bravais-Pearson therefore r_s can be interpreted as a percentage of explained variability
- τ can be interpreted as a probability for a positive ($\tau > 0$) or a negative ($\tau < 0$) relationship
- if several τ 's can be computed then choose the lowest one
- If $T(x)$ is a monotone transformation, e.g. log, then Kendall's τ and Spearman's ρ are unchanged

Siegel, Sidney and Castellan, N. J. (Dec. 31, 1988). *Nonparametric Statistics for the Behavioral Sciences*. International 2 Revised ed Edition. New York: McGraw-Hill Professional. 330 pp.
isbn: 978-0-07-100326-1.

Somer's D

- asymmetric

$$D_{Y|X} = \frac{C - D}{C + D + T_y}$$

- symmetric

$$D = \frac{C - D}{C + D + (T_x + T_y)/2}$$

- between τ coefficients and Goodman and Kruskal's γ

Somers, Robert H. (1962). "A New Asymmetric Measure of Association for Ordinal Variables". In: *American Sociological Review* 27.6, pp. 799–811. issn: 00031224. url:
<http://www.jstor.org/stable/2090408>.



Listing 11.10: example_somerd.R

```
1 library("vcdExtra") # for Accident
2 library("devtools") # for source_...
3 Accident$mode <- ordered(Accident$mode, levels=levels(Accident$mode) [c(1, 3, 2, 4)])
4 tab <- xtabs(Freq~mode+age, data=Accident)
5 tab
6 source_url("http://gist.githubusercontent.com/marcshwartz/3665743/raw",
7 #source_gist("3665743",
8 #               sha1="9c4f2ccf91a88be5734b0a616156500add8acdb2")
9 calc.Sd(tab)
```

[gist.github.com/3665743](https://gist.github.com/marcshwartz/3665743)

Contingency tables

$X \setminus Y$	y_1	y_2	\dots	y_K	Marginal of X
x_1	h_{11}	h_{12}	\dots	h_{1K}	$h_{1\bullet}$
x_2	h_{21}	h_{22}	\dots	h_{2K}	$h_{2\bullet}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_J	h_{J1}	h_{J2}	\dots	h_{JK}	$h_{J\bullet}$
Marginal of Y	$h_{\bullet 1}$	$h_{\bullet 2}$	\dots	$h_{\bullet K}$	$h_{\bullet\bullet}$

- Common or joint distribution of X and Y h_{ij}
- Marginal distribution of X ($h_{i\bullet} = \sum_j h_{ij}$) and Y ($h_{\bullet j} = \sum_i h_{ij}$)
- Absolute ($h_{\bullet\bullet} = n$) and relative frequencies ($f_{\bullet\bullet} = 1$)

Conditional frequencies

$Y X$	y_1	y_2	\dots	y_K	
x_1	$f(y_1 x_1) = \frac{h_{11}}{h_{1\bullet}}$	$f(y_2 x_1) = \frac{h_{12}}{h_{1\bullet}}$	\dots	$f(y_K x_1) = \frac{h_{1K}}{h_{1\bullet}}$	1
x_2	$f(y_1 x_2) = \frac{h_{21}}{h_{2\bullet}}$	$f(y_2 x_2) = \frac{h_{22}}{h_{2\bullet}}$	\dots	$f(y_K x_2) = \frac{h_{2K}}{h_{2\bullet}}$	1
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_J	$f(y_1 x_J) = \frac{h_{J1}}{h_{J\bullet}}$	$f(y_2 x_J) = \frac{h_{J2}}{h_{J\bullet}}$	\dots	$f(y_K x_J) = \frac{h_{JK}}{h_{J\bullet}}$	1
$X Y$	y_1	y_2	\dots	y_K	
x_1	$f(x_1 y_1) = \frac{h_{11}}{h_{\bullet 1}}$	$f(x_1 y_2) = \frac{h_{12}}{h_{\bullet 2}}$	\dots	$f(x_1 y_K) = \frac{h_{1K}}{h_{\bullet K}}$	
x_2	$f(x_2 y_1) = \frac{h_{21}}{h_{\bullet 1}}$	$f(x_2 y_2) = \frac{h_{22}}{h_{\bullet 2}}$	\dots	$f(x_2 y_K) = \frac{h_{2K}}{h_{\bullet K}}$	
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_J	$f(x_J y_1) = \frac{h_{J1}}{h_{\bullet 1}}$	$f(x_J y_2) = \frac{h_{J2}}{h_{\bullet 2}}$	\dots	$f(x_J y_K) = \frac{h_{JK}}{h_{\bullet K}}$	
	1	1	\dots		1

- If X and Y independent then holds for all j, k :
 $f(y_k|x_1) = \dots = f(y_k|x_J)$ and $f(x_j|y_1) = \dots = f(x_j|y_K)$

χ^2 based coefficients

- χ^2 test value

$$\chi^2 = \sum_{ij} \frac{(h_{ij} - e_{ij})^2}{e_{ij}}$$

- ▶ expected frequency e_{ij} (under independence)
- ▶ observed frequency h_{ij}
- (Corrected) contingency coefficient ($m = \min(\text{rows, columns})$)

$$C = \sqrt{\frac{k^2}{k^2 + n}} < 1, \quad C^* = C \sqrt{\frac{m}{m-1}}$$

- Cramér's V (Φ for 2×2 tables)

$$V = \sqrt{\frac{k^2}{n(m-1)}}$$

$$\sum_{ij} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} = \sum_{ij} \frac{h_{ij}^2}{e_{ij}} - 2 \underbrace{\sum_{ij} \frac{h_{ij} e_{ij}}{e_{ij}}}_{=n} + \underbrace{\sum_{ij} \frac{e_{ij}^2}{e_{ij}}}_{=n} = \sum_{ij} \frac{h_{ij}^2}{e_{ij}} - n$$

$$\sum_{ij} \frac{h_{ij}^2}{e_{ij}} - n = \sum_{ij} \frac{n h_{ij}^2}{h_{i\bullet} h_{\bullet j}} - n \leq n \min(I-1, J-1)$$

$$\sum_{i=1}^I \frac{n}{h_{i\bullet}} \underbrace{\sum_{j=1}^J h_{ij} \frac{h_{ij}}{h_{\bullet j}}}_{\leq \max_j h_{ij}} - n \leq \sum_{i=1}^I n \underbrace{\frac{\max_j h_{ij}}{h_{i\bullet}}}_{\leq 1} - n \leq n(I-1)$$

$$\sum_{j=1}^J \frac{n}{h_{\bullet j}} \underbrace{\sum_{i=1}^I h_{ij} \frac{h_{ij}}{h_{i\bullet}}}_{\leq \max_i h_{ij}} - n \leq \sum_{j=1}^J n \underbrace{\frac{\max_i h_{ij}}{h_{\bullet j}}}_{\leq 1} - n \leq n(J-1)$$

④ Listing 11.11: example_chi2coeff.R

```
1 library("vcd")
2 dim(HairEyeColor)
3 assocstats(HairEyeColor)
4 tab <- apply(HairEyeColor, 1:2, sum)
5 chisq.test(tab)
```

④ chisq.test(table)\$statistic
④ vcd::assocstats(table)\$cont
④ vcd::assocstats(table)\$cramer
④ vcd::assocstats(table)\$phi

χ^2 independence test

Assumption(s): X, Y categorical or grouped, $e_{ij} > 5$ for all cells
 (other approximation conditions are possible)

Hypotheses:

$$H_0 : X, Y \text{ independent vs.}$$

$$H_1 : X, Y \text{ are not independent}$$

Test statistics:

$$K^2 = n \sum_{i=1}^I \sum_{j=1}^J \frac{(F_{ij} - p_{ij})^2}{p_{ij}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(H_{ij} - e_{ij})^2}{e_{ij}}$$

$$K^2 \approx \chi^2_{(I-1)(J-1)}$$

Reject H_0 :

$$|k^2| > \chi^2_{(I-1)(J-1); 1-\alpha}$$

h_{ij} observed observations, e_{ij} expected observations

f_{ij} observed frequency, p_{ij} expected frequency

I number of categories of X

J number of categories of Y

- Effect size $w = \sqrt{k^2/n}$ (0.1 small, 0.3 medium, 0.5 large)
- Derivation of test statistic (H_{ij} not independent)

$$H_{ij} \sim B(n; p_{ij}) \Rightarrow E(H_{ij}) = e_{ij}, \text{Var}(H_{ij}) = np_{ij} \underbrace{(1 - p_{ij})}_{\leq 1} \leq np_{ij} = e_{ij}$$

$$H_{ij} \approx N(e_{ij}; e_{ij}) \Rightarrow \frac{H_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \frac{H_{ij} - E(H_{ij})}{\sqrt{\text{Var}(H_{ij})}} \approx N(0; 1)$$

$$K^2 = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{H_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right)^2 \approx \chi^2_{(I-1)(J-1)}$$

Yates's correction for continuity

- χ^2 independence test for 2×2 tables with continuity correction ($n \leq 20$)

$$K_c^2 = \frac{n(|H_{11}H_{22} - H_{12}H_{21}| - 0.5n)^2}{(H_{11} + H_{12})(H_{21} + H_{22})(H_{11} + H_{21})(H_{12} + H_{21})} \approx \chi_1^2$$

- ▶ better approximation to χ_1^2 distribution
- ▶ too conservative
- for a general χ^2 test

$$\sum_{\bullet} \frac{(|H_{\bullet} - e_{\bullet}| - 0.5)^2}{e_{\bullet}}$$

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the χ^2 Test". In: *Supplement to the Journal of the Royal Statistical Society* 1.2, p. 217. issn: 14666162. doi: 10.2307/2983604. url: <http://www.jstor.org/stable/10.2307/2983604?origin=crossref> (visited on 12/02/2015).



Listing 11.12: example_indep.R

```
1 library("MASS") # for Boston Housing data
2 chisq.test(table(Boston$chas, Boston$rad))
```

⌚ chisq.test(x, rescale.p=FALSE, simulate.p.value=FALSE, B=2000,
correct=F)

Fisher's exact test

- Apply for 2×2 tables if

- ▶ $n < 30$
- ▶ at least one $e_{ij} < 5$ or
- ▶ strong asymmetry

$$H_{11} \sim Hyp(Np_{11}; N; n)$$

under the independence assumption all other frequencies are fixed

$X \setminus Y$	y_1	y_2	Marginal of X (fix)
x_1	h_{11}	$h_{12} = h_{1\bullet} - h_{11}$	$h_{1\bullet}$
x_2	$h_{21} = h_{\bullet 1} - h_{11}$	$h_{22} = h_{2\bullet} - h_{\bullet 1} - h_{11}$	$h_{2\bullet}$
Marginal of Y (fix)	$h_{\bullet 1}$	$h_{\bullet 2}$	$h_{\bullet\bullet}$

- ▶ too conservative since test statistics is discrete
- ▶ can be extended to general $I \times J$ tables

Fisher, R. A. (Jan. 1922). "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P". In: *Journal of the Royal Statistical Society* 85.1, p. 87. issn: 09528385. doi: 10.2307/2340521. url: <http://www.jstor.org/stable/2340521?origin=crossref> (visited on 12/02/2015).



Listing 11.13: example_fisher.R

```
1 # acacia ants (Hand (1994), Handbook of small datasets)
2 aa <- data.frame(acacia=c("A", "B", "A", "B"),
3                     invaded=c("No", "No", "Yes", "Yes"),
4                     Freq=c(2,10,13,3))
5 tab <- xtabs(Freq~acacia+invaded, data=aa)
6 tab
7 fisher.test(tab)
```



```
fisher.test(table, hybrid=F, alternative="two.sided",
            simulate.p.value=F, B=2000)
```

Linear-by-Linear association test

Assumptions: X and Y are (ordinal) metric discrete variables

Hypotheses: H_0 : X and Y are independent vs.

H_1 : X and Y are dependent

Test statistics: $M = (n - 1)R^2 \approx \chi_1^2$

Reject H_0 : $m > \chi_{1;1-\alpha}^2$

Remark: test takes order into account

Example 11.22

# Traumatic Events	0	1	2	3	4+	
Dropout from study	25	13	9	10	6	63
Remain in study	31	21	6	2	3	63
	56	34	15	12	9	126

$$k^2 = 9.459 < 9.488 = \chi_{4;0.95}^2 \Rightarrow \text{can not reject } H_0$$

$$m = 5.757 > 3.842 = \chi_{1;0.95}^2 \Rightarrow \text{reject } H_0$$

④ Listing 11.14: example_linear_by_linear.R

```
1 library("coin")
2 # create data
3 tab <- as.table(rbind(c(25,13,9,10,6), c(31,21,6,2,3)))
4 # Chi-square test
5 chisq.test(tab)
6 # Linear-by-linear association
7 lbl_test(tab)
```

④ `coin::lbl_test(table, distribution=c("asymptotic", "approximate"))`

PRE measures

- Prediction Reduction Error measures
- measure the association by comparing the errors of two prediction methods
 - ▶ method 1 *does not* take the association into account
 - ▶ method 2 *does* take the association into account
 - ▶ compare both errors

$$PRE = \frac{E_1 - E_2}{E_1} = 1 - \frac{E_2}{E_1}$$

- same interpretation for each level of measurement
 - ▶ $0 \leq PRE \leq 1$
 - ▶ it may happen $PRE_{Y|X} \neq PRE_{X|Y}$
 - ▶ unsymmetric and symmetric measures are possible

- metric: R^2

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\bar{y} - y_i)^2}$$

- ▶ \bar{y} best prediction without knowledge of the association
- ▶ \hat{y}_i best prediction with knowledge of the association
- ordinal: absolute values of Somers D (with ties), Goodman and Kruskals γ (without ties), Kendalls τ_b (quadratic) and τ_c (arbitrary)
- nominal: Goodman and Kruskals τ (based on distribution) and λ (based on max. category), uncertainty coefficient (bivariate entropy)
- dependent metric, independent nominal or ordinal: η^2

Goodman and Kruskal's λ

- for E_1 compute the error when predicting the mode of the dependent variable
- for E_2 compute the error when predicting the mode for each column/row of the dependent variable

$$\lambda_{Y|X} = \frac{E_1^{Y|X} - E_2^{Y|X}}{E_1^{Y|X}}$$

$$\lambda = \frac{(E_1^{Y|X} - E_2^{Y|X}) + (E_1^{X|Y} - E_2^{X|Y})}{E_1^{Y|X} + E_1^{X|Y}}$$

Goodman, Leo A and Kruskal, William H (1979). *Measures of Association for Cross Classifications.* (A book with the four landmark papers from 1954, 1959, 1963 and 1972 by Goodman and Kruskal). New York, NY: Springer New York. isbn: 978-1-4612-9995-0. url: <http://dx.doi.org/10.1007/978-1-4612-9995-0> (visited on 12/09/2015).

Wirtschaftslage des Befragten heute	Wirtschaftslage in der BRD heute					Summe
	Sehr gut	Gut	Teils teils	Schlecht	Sehr schlecht	
Sehr gut (++)	60	88	36	5	1	190
Gut (+)	95	918	759	113	12	1897
Teils teils (0)	19	226	557	190	25	1017
Schlecht (-)	9	46	121	106	14	296
Sehr schlecht (—)	1	10	20	19	13	63
Summe	184	1288	1493	433	65	3463

$$E_1 = 3463 - 1897 = 1566$$

$$E_2 = 89 + 370 + 734 + 243 + 40 = 1476$$

$$\lambda = \frac{1566 - 1476}{1566} = 0,0575$$



Listing 11.15: example_gkl.R

```
1 library("MASS")      # for Boston Housing data
2 library("DescTools")
3 tab <- table(Boston$chas, Boston$rad)
4 Lambda(tab)
5 #
6 library("ryouready")
7 nom.lambda(tab)
```

② DescTools::Lambda(x, y=NULL, direction=c("symmetric", "row",
 "column"), conf.level=NA)

② ryouready::nom.lambda(tab)

Goodman and Kruskal's τ

- use a random draw from the distribution for predicting rather than the mode

	++	+	0	-	—	
Wirtschaftslage des Befragten heute	190 5.5%	1897 54.8%	1017 29.4%	296 8.5%	63 1.8%	3463 100.0%
Error E_1	94.5% 32.6% 67.4%	45.2% 51.6% 48.4%	70.6% 10.3% 89.7%	91.5% 4.9% 95.1%	98.2% 0.5% 99.5%	
Wirtschaftslage in der BRD heute (++)	60 32.6% 67.4%	95 51.6% 48.4%	19 10.3% 89.7%	9 4.9% 95.1%	1 0.5% 99.5%	184 100.0%

analogous for the other columns

$$E_1 = 0.055 \cdot 0.945 + \dots + 0.018 \cdot 0.982 = 0.603$$

$$E_{2.1} = 0.326 \cdot 0.674 + \dots + 0.005 \cdot 0.995 = 0.614$$

$$E_2 = 0.053 \cdot 0.614 + \dots + 0.019 \cdot 0.731 = 0.557$$

$$\tau = \frac{0.603 - 0.557}{0.603} = 0.077$$

R Listing 11.16: example_gkt.R

```
1 library("MASS")      # for Boston Housing data
2 library("DescTools")
3 tab <- table(Boston$chas, Boston$rad)
4 GoodmanKruskalTau(tab)
```

R DescTools::GoodmanKruskalTau(x, y=NULL, direction=c("row",
 "column"), conf.level = NA)

Eta squared

- Y metric, X nominal or ordinal

$$\eta^2 = 1 - \frac{E_2}{E_1}$$

- E_1 error when predicting \bar{y}

$$E_1 = \sum_{i=1}^n (y_i - \bar{y})^2$$

- E_2 error when predicting \bar{y}_k for each category

$$E_2 = \sum_{k=1}^p \sum_{i=1}^n (y_i - \bar{y}_k)^2 I(\text{Obs. } i \text{ in group } k)$$

- η^2 can be used as effect size for one-way ANOVA
0.02 - small effect, 0.13 - medium effect, 0.26 - large effect

Cohen's κ

- for interrater agreement (high association \neq high agreement)

$$\kappa = \frac{f_a - f_r}{1 - f_r} \quad \text{with } f_a = \frac{1}{n} \sum_{i=1}^I h_{ii}, \quad f_r = \frac{1}{n^2} \sum_{i=1}^I h_{i\bullet} h_{\bullet i}$$

- from a $I \times I$ contingency table
 - ▶ f_a percentage of identical ratings from the raters
 - ▶ f_r percentage of identical ratings if raters judge randomly
- judgement
 - ▶ Landis & Koch: 0-0.2 as slight, 0.2-0.4 as fair, 0.4-0.6 as moderate, 0.6-0.8 as substantial, and 0.8-1 as almost perfect
 - ▶ Fleiss's: 0-0.4 as poor, 0.4-0.75 as fair to good, 0.75-1 as excellent
- can be extended to m raters
 - ▶ Fleiss's κ
 - ▶ Light's κ

- Cohen's κ for 2 rates

- R vcd::Kappa(table)
 - R psych::cohen.kappa(table, alpha=.05)
 - R epibasix::epiKappa(table, alpha=.05, k0=0.4)
 - R epiR::epi.kappa(a, b, c, d, conf.level=0.95)
 - R psy::ckappa(ratings)
 - R irr::kappa2(ratings)

- for m raters

- R psy::lkappa(ratings, type="Cohen")
 - R epicalc::kap(ratings)
 - R irr::kappam.fleiss(ratings)
 - R irr::kappam.light(ratings)

Cohen's κ test

Assumptions: $n \min(f_a, 1 - f_a) > 5$

Hypotheses: $H_0 : \kappa = \kappa_0$ vs. $H_1 : \kappa \neq \kappa_0$

Test statistics: $Z = \frac{\kappa - \kappa_0}{\sqrt{\frac{f_a(1-f_a)}{n(1-f_r)^2}}} \approx N(0; 1)$

Reject H_0 : $z > z_{1-\alpha/2}$

Derivation of
test statistics $F_a \sim Hyp(N; M = f_r + \kappa_0(1 - f_r); n)$
 $\Rightarrow K = \frac{F_a - f_r}{1 - f_r} \approx N \left(\kappa_0; \sigma_K^2 = \frac{f_a(1-f_a)}{n(1-f_r)^2} \right)$

McNemar's test

Assumptions:

X and Y are binary and dependent, $\hat{e} > 5$

Hypotheses:

$H_0 : \pi_{10} = \pi_{01}$ vs. $H_1 : \pi_{10} \neq \pi_{01}$

Test statistics:

$$V = \frac{(H_{01} - e)^2}{e} + \frac{(H_{10} - e)^2}{e} \approx \chi^2_1$$

$$v = \frac{(h_{01} - h_{10})^2}{h_{01} + h_{10}} \quad (\hat{e} = \frac{h_{01} + h_{10}}{2})$$

$$v = \frac{(|h_{01} - h_{10}| - 0.5)^2}{h_{01} + h_{10}}$$

$$v = \frac{(|h_{01} - h_{10}| - 1)^2}{h_{01} + h_{10}} \quad (h_{01} + h_{10} < 30)$$

with Yates correction

$$|v| > \chi^2_{1;1-\alpha}$$

with Edwards correction

Reject H_0 :

Example 11.23 (Opinion about a car free sunday)

before / after	yes	no	
yes	8	5	13
no	16	11	27
	24	16	40

- Does the general opinion change after the respondents experienced a car free sunday?

$$\nu = \frac{(16 - 5)^2}{16 + 5} = 5.762$$

$$v_Y = \frac{(|16 - 5| - 0.5)^2}{16 + 5} = 5.250$$

$$v_E = \frac{(|16 - 5| - 1)^2}{16 + 5} = 4.762$$

$$\chi^2_{1;0.95} = 3.84 \Rightarrow \text{reject } H_0$$

McNemar, Quinn (June 1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02295996. url:

<http://link.springer.com/10.1007/BF02295996> (visited on 12/09/2015).

Edwards, Allen L. (Sept. 1948). "Note on the "correction for continuity" in testing the significance of the difference between correlated proportions". In: *Psychometrika* 13.3, pp. 185–187. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02289261. url:

<http://link.springer.com/10.1007/BF02289261> (visited on 12/09/2015).



Listing 11.17: example_mc nemar.R

```
1 tab <- matrix(c(8,16,5,11), nrow=2)
2 colnames(tab) <- c("after_yes", "after_no")
3 rownames(tab) <- c("before_yes", "before_no")
4 tab
5 # uses Edwards correction
6 mcnemar.test(tab)
```



```
mcnemar.test(table, correct=T)
```

Cochran-Mantel-Haenszel Test

Assumptions: X, Y binary variable, Z categorical variable

Hypotheses: $H_0 : OR_{XY|z_1} = \dots = OR_{XY|z_K} = 1$ vs.

$H_1 : \text{at least for one } i \text{ holds } OR_{XY|z_i} \neq 1$

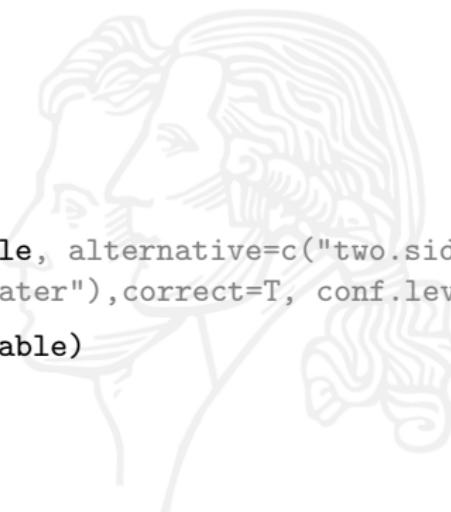
Test statistics: $M^2 = \frac{\left(\left| \sum_{k=1}^K h_{11k} - \sum_{k=1}^K \hat{e}_{11k} \right| - 0.5 \right)^2}{\sum_{k=1}^K \hat{\sigma}^2(h_{11k})} \approx \chi_1^2$

$$\hat{e}_{11k} = \frac{h_{1\bullet k} h_{\bullet 1k}}{h_{\bullet\bullet k}}, \hat{\sigma}(h_{11k}) = \frac{h_{1\bullet k} h_{2\bullet k} h_{\bullet 1k} h_{\bullet+2k}}{h_{\bullet\bullet k}^2 (h_{\bullet\bullet k} - 1)}$$

Reject H_0 : $m^2 > \chi_{1:1-\alpha}^2$

Remark: Association should not change in each 2×2 table

- Cochran, William G. (1954). "Some Methods for Strengthening the Common χ^2 Tests". In: *Biometrics* 10.4, pp. 417–451. issn: 0006341X, 15410420. url: <http://www.jstor.org/stable/3001616> (visited on 09/01/2022).
- Mantel, Nathan and Haenszel, William (Apr. 1959). "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease". In: *JNCI: Journal of the National Cancer Institute* 22.4, pp. 719–748. issn: 0027-8874. doi: 10.1093/jnci/22.4.719. eprint: <https://academic.oup.com/jnci/article-pdf/22/4/719/2674652/22-4-719.pdf>. url: <https://doi.org/10.1093/jnci/22.4.719>.



```
R mantelhaen.test(table, alternative=c("two.sided", "less",
                                         "greater"), correct=T, conf.level=0.95)
R lawstat::cmh.test(table)
```

Relative risk

	factor variable			
Event did	I_1	\dots	I_m	
not occur	h_{01}	\dots	h_{0m}	$h_{0\bullet}$
occur	h_{11}	\dots	h_{1m}	$h_{1\bullet}$
consider only $m = 2$				

- Cohort study or prospective study
 - ▶ Study design: at begin two groups, one with factor, one without factor, event has not occurred, observe how many events occur
- Incidence rates: $I_m = \frac{h_{1m}}{h_{0m} + h_{1m}}$
 - ▶ percentage how often event occurred in each group
- Relative risk ($m = 2$): $RR = I_2/I_1$

Example 11.24

	non-smoker	smoker	
no cancer	30	2	32
cancer	8	10	18
	38	12	50

- cohort study: representative for the smoker/non-smoker in the population at study begin

$$I_{smoker} = 0.83, I_{non-smoker} = 0.21, rr = 3.96$$

Example 11.25 (PIP implant scandal)

- PIP silicon gel filled implants are placed in 400.000 women in 65 countries
 - ▶ 2000 TÜV Rheinland gave a CE mark
 - ▶ 2006 multiple ruptures are reported
 - ▶ 2010 french regulator announced that unauthorized silicon gel has been used
 - ▶ 2011 french government recommends removal of PIP implants
- 2012 National Health Service wants to assess the risk
 - ▶ rupture rates ranges from 1% - 7% (8 out of 108 women)
 - ▶ problem: no data since no register available
 - ▶ retrospective study to find out how many implants from which brand are used
 - ▶ prospective study to surgeons who remove PIP implants

Test on relative risk

Assumptions: $n > 30$

Hypotheses: $H_0 : RR = 1$ vs. $H_1 : RR \neq 1$

Test statistics: $Z = \frac{\log(RR)}{\sigma(\log(RR))} \approx N(0; 1)$

$$\hat{\sigma}(\log(RR)) = \sqrt{\left(\frac{1}{h_{12}} + \frac{1}{h_{11}}\right) - \left(\frac{1}{h_{02}+h_{12}} + \frac{1}{h_{01}+h_{11}}\right)}$$

Reject H_0 : $|z| > z_{1-\alpha/2}$

- usually confidence intervals are computed
- test relies on asymptotics ($n \rightarrow \infty$)

- ⌚ `epiR::epi.2by2(table, method="case.control", conf.level=0.95)`
- ⚠ Other methods are `cohort.count` (default), `cohort.time`, or `cross.sectional`

Odds ratio

	factor variable	
Event did	$l_1 \dots l_m$	
not occur	$h_{01} \dots h_{0m}$	$h_{0\bullet}$
occur	$h_{11} \dots h_{1m}$	$h_{1\bullet}$

consider only $m = 2$

- Case control study or retrospective study
 - ▶ Study design: two groups, one with event occurred, one without event occurred, observe factor level
- Odds: $o_m = \frac{h_{1m}}{h_{0m}}$
 - ▶ for each group observe proportion event occurred and event not occurred
- Odds-ratio ($m = 2$): $or = o_2/o_1$

Example 11.26

	non-smoker	smoker	
no cancer	30	2	32
cancer	8	10	18
	38	12	50

- case control study: representative for the cancer/non-cancer in the current population

$$o_{smoker} = 5.00, o_{non-smoker} = 0.27, or = 18.75$$

Example 11.27 (PIP implant scandal)

- PIP silicon gel filled implants are placed in 400.000 women in 65 countries
 - ▶ 2000 TÜV Rheinland gave a CE mark
 - ▶ 2006 multiple ruptures are reported
 - ▶ 2010 french regulator announced that unauthorized silicon gel has been used
 - ▶ 2011 french government recommends removal of PIP implants
- 2012 National Health Service wants to assess the risk
 - ▶ rupture rates ranges from 1% - 7% (8 out of 108 women)
 - ▶ problem: no data since no register available
 - ▶ retrospective study to find out how many implants from which brand are used

- retrospective study: 230,000 implants in 131,000 women

	Not-PIP	PIP	
Explant with clinical signs	215	246	677
Other	104737	25736	130257
	104952	25982	130934
Odds	0.2%	0.9%	<i>or</i> = 4.6
	Not-PIP	PIP	
Explant with implant failure	245	432	677
Other	104707	25550	130257
	104952	25982	130934
Odds	0.2%	1.7%	<i>or</i> = 7.1

- severe uncertainty, substantial underestimation

Test on odds ratio

Assumptions: $h_{ij} > 10$ and $n > 30$

Hypotheses: $H_0 : OR = 1$ vs. $H_1 : OR \neq 1$

Test statistics: $Z = \frac{\log(OR)}{\sigma(\log(OR))} \approx N(0; 1)$

$$\hat{\sigma}(\log(OR)) = \sqrt{\frac{1}{h_{12}} + \frac{1}{h_{11}} + \frac{1}{h_{01}} + \frac{1}{h_{02}}}$$

Reject H_0 : $|z| > z_{1-\alpha/2}$

- usually confidence intervals are computed
- test relies on asymptotics ($n \rightarrow \infty$)

- ⌚ `fisher.test(table, or=1, conf.int=T, conf.level=0.95)`
uses the conditional MLE rather than the unconditional MLE (the sample odds ratio)
- ⌚ `vcd::oddsratio(table, log=T)`
- ⌚ `epiR::epi.2by2(table)`
- ⌚ `epitools::oddsratio(table, method=c("midp", "fisher", "wald", "small"), conf.level=0.95, correction=F)`

Common odds ratio

$$or_{XY; MH} = \frac{\sum_{k=1}^K \frac{h_{11k} h_{22k}}{h_{\bullet\bullet k}}}{\sum_{k=1}^K \frac{h_{12k} h_{21k}}{h_{\bullet\bullet k}}}$$

- Breslow-Day test and Tarone test are tests on homogeneity:
 - ▶ $H_0 : OR_{XY|z_1} = \dots = OR_{XY|z_K}$ vs.
 - ▶ $H_1 : \text{least for one pair } (i, j) \text{ holds } OR_{XY|z_i} \neq OR_{XY|z_j}$

Tarone, Robert E. (1985). "On heterogeneity tests based on efficient scores". In: *Biometrika* 72.1, pp. 91–95. issn: 0006-3444, 1464-3510. doi: 10.1093/biomet/72.1.91. url: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/72.1.91> (visited on 09/01/2022).

Breslow, N. E. (Mar. 1996). "Statistics in Epidemiology: The Case-Control Study". In: *Journal of the American Statistical Association* 91.433, pp. 14–28. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1996.10476660. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476660> (visited on 09/01/2022).

```
R metafor::rma.mh(ai, bi, ci, di, n1i, n2i)$QE  
R metafor::rma.mh(ai, bi, ci, di, n1i, n2i)$QEp  
R metafor::rma.mh(ai, bi, ci, di, n1i, n2i)$TA  
R metafor::rma.mh(ai, bi, ci, di, n1i, n2i)$TAp
```

Multivariate Graphics

November 3, 2022

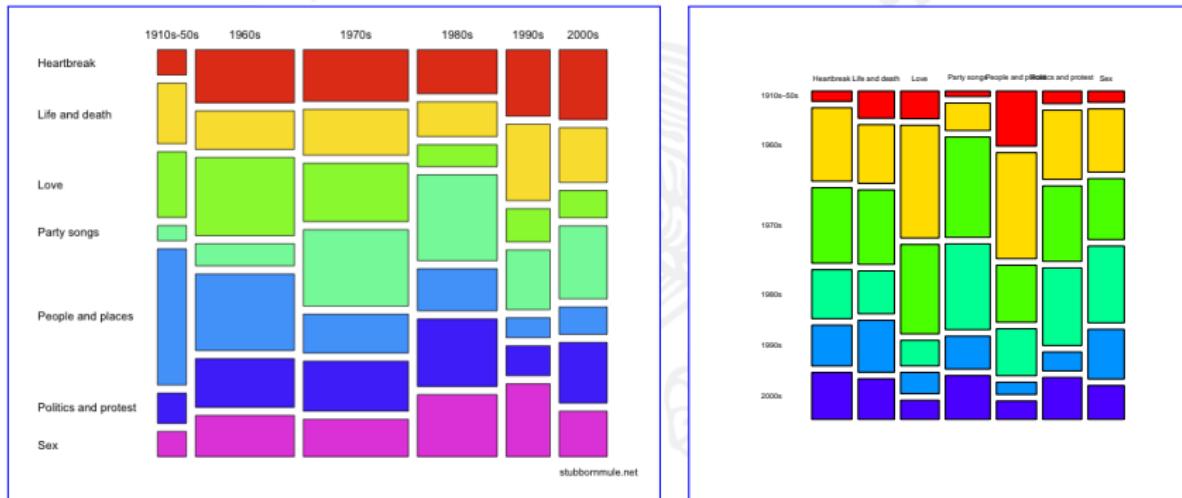
- Mosaic plot • Scatterplot matrix • Scagnostics • Trellis display • 3D plots
- Parallel coordinates • Andrews curves • Radar chart • Chernoff faces

Mosaic plot

- Hartigan, Kleiner, (1981), Minard (1844)
- Visualization: association
 - ▶ Variables: two (or more) categorical variables
 - ▶ Observations: unlimited
- Plot for two variables:
 - ▶ width of a block: proportional to song frequency in a decade
 - ▶ height of a block: proportional to conditional song frequency theme given by decade
 - ▶ area of block: proportional to the song frequency by theme and decade
- Problems: order of the variables

Minard, C.J. (1844). *Tableaux figuratifs de la circulation de quelques chemins de fer.*

Hartigan, J. A. and Kleiner, B. (1981). "Mosaics for Contingency Tables". In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. Ed. by William F. Eddy. New York, NY: Springer US, pp. 268–273. isbn: 978-0-387-90633-1 978-1-4613-9464-8. url: http://link.springer.com/10.1007/978-1-4613-9464-8_37 (visited on 08/26/2015).



Joint and marginal frequency distributions

$X \setminus Y$	0	1	
0	f_{11}	f_{12}	$f_{1\bullet}$
1	f_{21}	f_{22}	$f_{2\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	1

Conditional frequency distributions (conditioned on X)

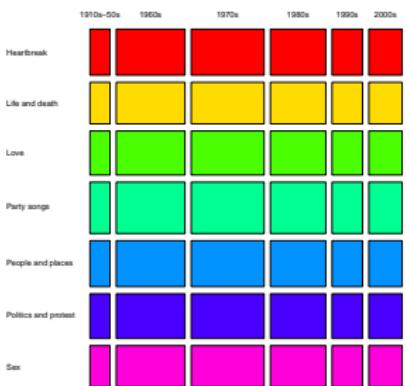
	0	1	
0	$f(y_1 x_1) = f_{11}/f_{1\bullet}$	$f(y_2 x_1) = f_{12}/f_{1\bullet}$	1
1	$f(y_2 x_1) = f_{21}/f_{2\bullet}$	$f(y_2 x_2) = f_{22}/f_{2\bullet}$	1

Conditional frequency distributions (conditioned on Y)

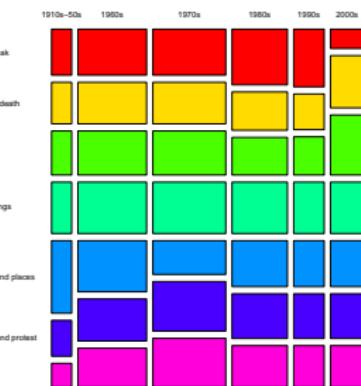
	0	1	
0	$f(x_1 y_1) = f_{11}/f_{\bullet 1}$	$f(x_1 y_2) = f_{12}/f_{\bullet 2}$	
1	$f(x_2 y_1) = f_{21}/f_{\bullet 1}$	$f(x_2 y_2) = f_{22}/f_{\bullet 2}$	
	1	1	



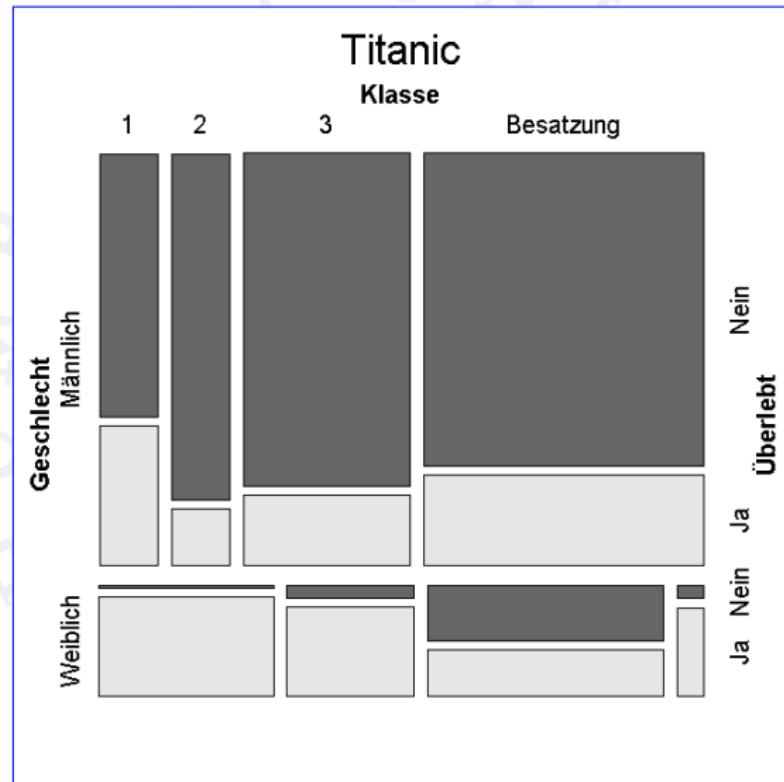
- area = height \times width
- $f_{ij} = f(y_j|x_i) \times f_{i\bullet}$
- if X and Y are independent it holds $f(y_i|x_1) = f(y_i|x_2)$
- if X and Y are independent the horizontal gaps should be at same height
- two conditional frequencies are possible (conditioned on X and Y)
- order of categorical variables plays a role



Full independence



Partial independence



⌚ Listing 12.1: example_mosaic_plot.R

```
1 plot(Titanic)
```

⌚ Listing 12.2: example_mosaic_graphics.R

```
1 mosaicplot(Titanic)
```

⌚ plot(table)

⌚ mosaicplot(table, type=c("pearson", "deviance", "FT"))

② Listing 12.3: example_mosaic_vcd.R

```
1 library("vcd")
2 mosaic(Titanic)
```

② Listing 12.4: example_mosaic_cotabplot.R

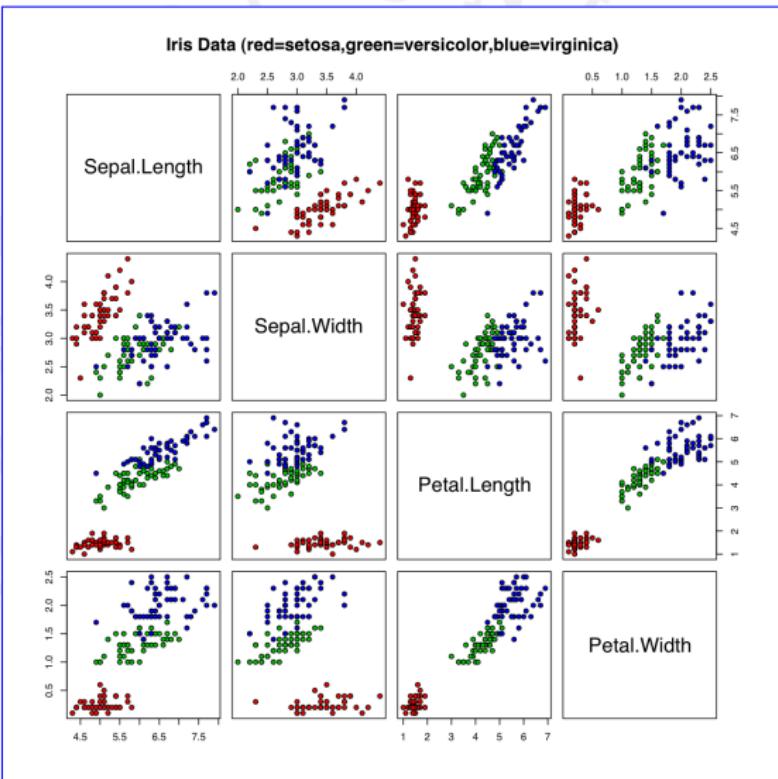
```
1 library("vcd")
2 cotabplot(Titanic)
```

② vcd::mosaic(table, direction)
② vcd::cotabplot(table)

Scatterplot matrix

- Visualization: association
 - ▶ Variables: three (or more) continuous variables
 - ▶ Observations: small
- Problems:
 - ▶ specific projections
 - ▶ limited number of variables
 - ▶ limited number of observations

Hartigan, J.A. (1975). "Printer graphics for clustering". In: *Journal of Statistical Computation and Simulation* 4.3, pp. 187–213.



R Listing 12.5: example_splographics.R

```
1 data(Boston, package="MASS")
2 pairs(~rm+lstat+medv, data=Boston, cex=0.5)
```

R Listing 12.6: example_splo_ggplot.R

```
1 library("MASS")
2 library("ggplot2")
3 library("GGally")
4 ggpairs(Boston, columns=c(6,13,14))
```

- R pairs(x)
- R pairs(formula, data)
- R car::scatterplot.matrix(x)
- R gclus::cpairs(x)
- R GGally::ggpairs(x, columns)

Scagnostics

Tukey and Tukey (1985) proposed measures to characterize a point cloud in a scatterplot (**Scatterplot diagnostics**)

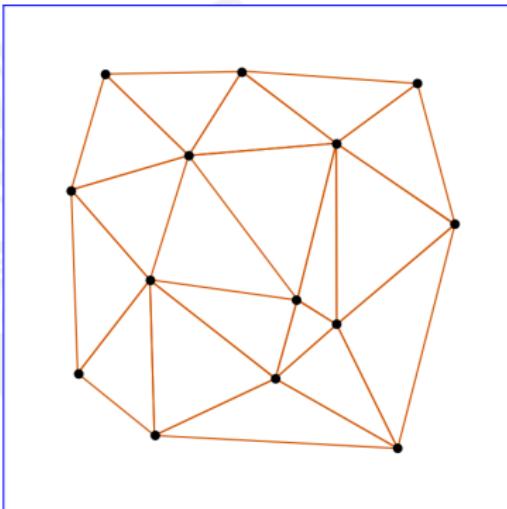
- based on peeled convex hull
- kernel density estimation \Rightarrow continuous variables
- principal curves
- computation complexity $O(n^3)$

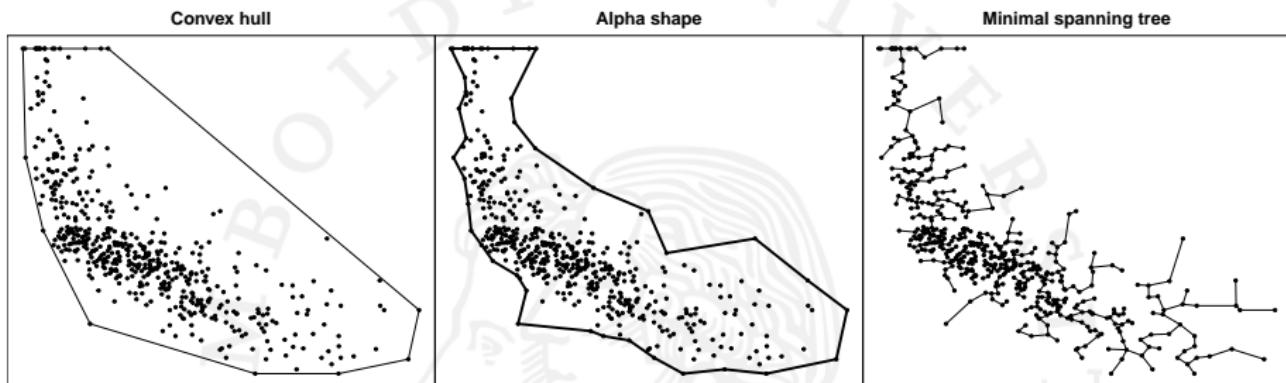
outlier	association
outlyingness	monotonicity
density	shape
skewness	convexity
lumpiness	skinnyness
sparseness	stringiness
striation	

Tukey, J.W. and Tukey, P.A. (1985). "Computer Graphics and Exploratory Data Analysis: An Introduction". In: *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics '85*. Vol. 3. Fairfax, VA, pp. 773–785.

Wilkinson, Anand, Grossman (2005, 2006)
worked out measurements based on graph
theoretical ideas

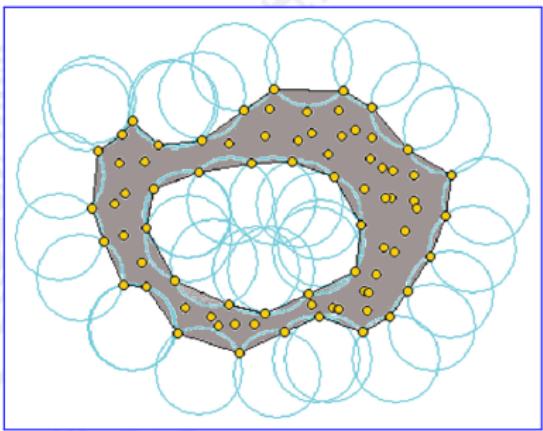
- Delaunay-Triangulation
- Minimal spanning tree (MST)
- Convex hull
- α shape
- Preprocessing
 - ▶ all variables are rescaled to [0; 1]
 - ▶ first compute outlyingness
 - ▶ delete outliers
 - ▶ compute all other coefficients





- Wilkinson, L., Anand, A., and Grossman, R. (2005). "Graph-theoretic scagnostics". In: IEEE, pp. 157–164. isbn: 978-0-7803-9464-3. doi: 10.1109/INFVIS.2005.1532142. url: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1532142> (visited on 08/27/2015).
- (Nov. 2006). "High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions". In: *IEEE Transactions on Visualization and Computer Graphics* 12.6, pp. 1363–1372. issn: 1077-2626. doi: 10.1109/TVCG.2006.94. url: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1703359> (visited on 08/27/2015).

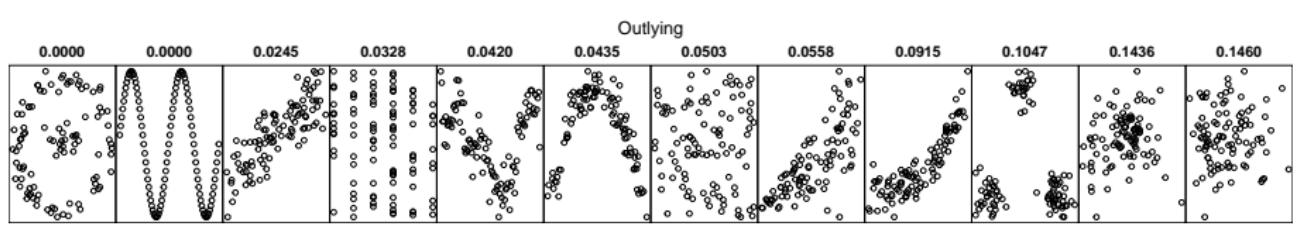
- α shape computation
 - ▶ a datapoint is called α extreme if he lies on a border of a circle with radius α and the circle contains no other data point inside or on the border
 - ▶ all α extreme data points form the alpha shape
 - ▶ choose α as $\min(0.1, 90\% \text{ quantile of edgelengths of MST})$
- all coefficients
 - ▶ are in the interval $[0, 1]$
 - ▶ can be computed in $O(n \log(n))$



Outlyingness

$$c_{outlying} = \frac{\text{length of all "long" edges in MST}}{\text{length of all edges in MST}}$$

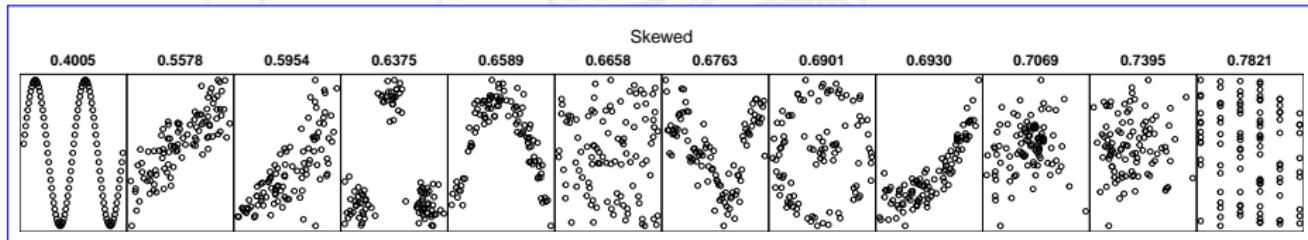
- compute 25% quantile $q_{0.25}$ and 75% quantile $q_{0.75}$ of edgelengths in MST
- a “long” edge is an edge which is longer than $q_{0.75} + 1.5 * (q_{0.75} - q_{0.25})$



Skewness

$$c_{\text{skewed}} = \frac{q_{0.9} - q_{0.5}}{q_{0.9} - q_{0.1}}$$

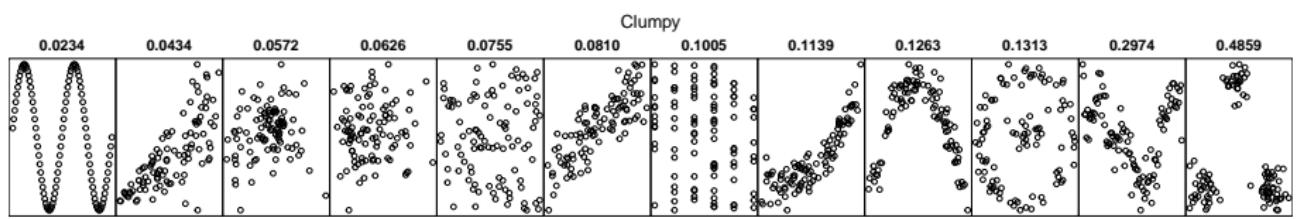
- compute $x\%$ quantiles q_x of edgelengths in MST
- a robust measure for a right-skewness of a distribution
- large values reflect that $q_{0.5} \approx q_{0.1} \Rightarrow$ some dense data region(s)



Lumpiness

$$c_{clumpy} = \max_j \left(1 - \frac{\max_k \text{ in smaller subgraph } e_k}{e_j} \right)$$

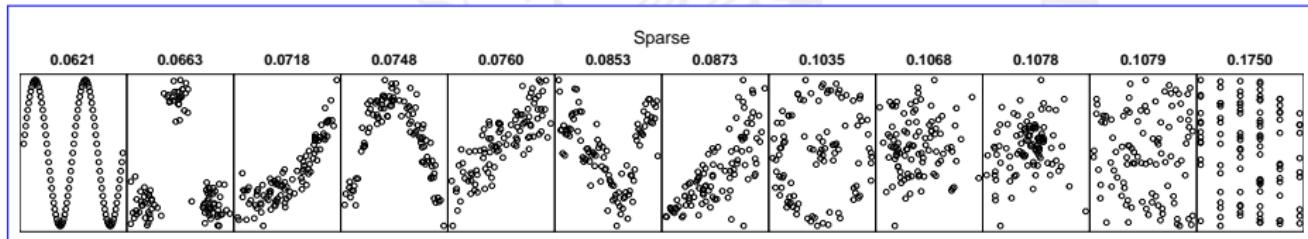
- compute edgelengths e_j in MST
- if edge j is taken away from then the MST decomposes in two subgraph
- look for the longest edge k in the subgraph with less data points
- if cloud point is clustered and we take an edge between two clusters then c_{clumpy} becomes large



Sparseness

$$c_{sparse} = \min(1, q_{0.9})$$

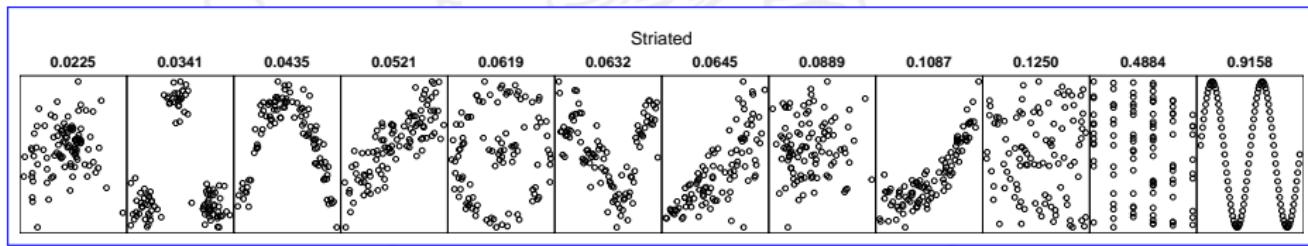
- compute 90% quantile $q_{0.9}$ of edgelengths in MST
- large values of c_{sparse} hint a highly discretized point cloud



Striation

$$c_{\text{striated}} = \frac{\#\text{"striated" vertices in MST}}{\#\text{all vertices in MST}}$$

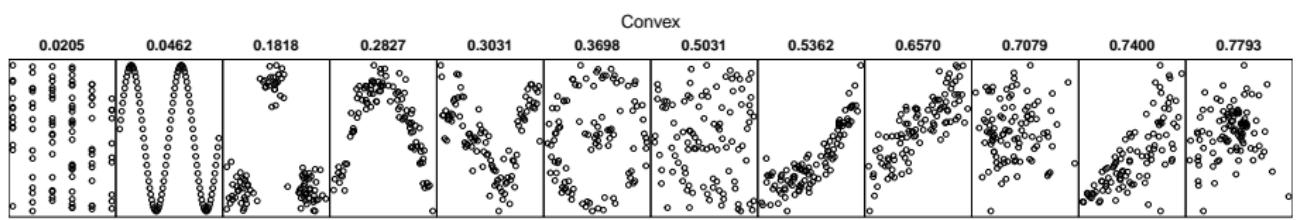
- a vertex is called “striated” if the angle between the edges is larger than 138,5 degrees ($\cos(\text{angle}) < -0.75$)
- functions, striated patterns etc. can be detected



Convexity

$$c_{\text{convex}} = \frac{\text{area of } \alpha \text{ shape}}{\text{area of convex hull}}$$

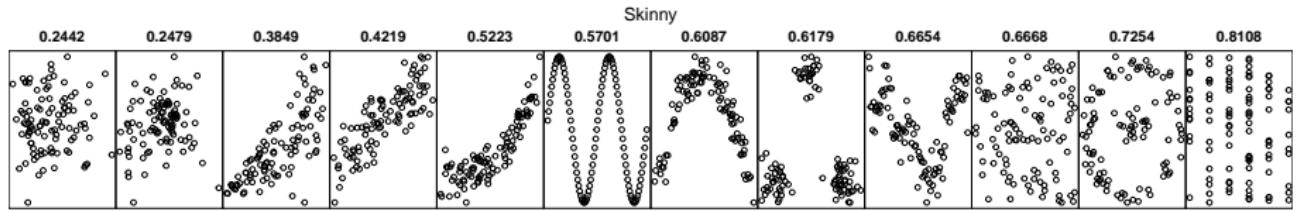
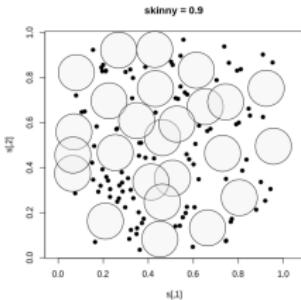
- large values of c_{convex} hint to a point cloud with no or only small "holes"
- gives information how the data fill out the convex hull



Skinnyness

$$c_{\text{skinny}} = 1 - \frac{\sqrt{4\pi \text{ area of } \alpha \text{ shape}}}{\text{length of border(s) of } \alpha \text{ shape}}$$

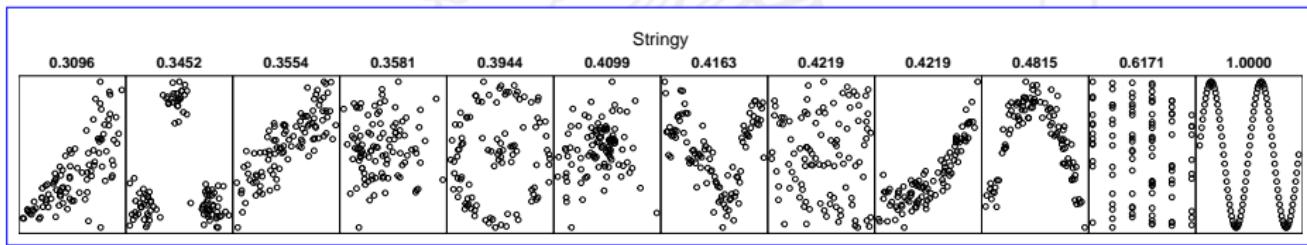
- large values of c_{skinny} hint to a point cloud with "holes"
- if the α shape is a circle then $c_{\text{skinny}} = 0$ since $A_{\text{circle}} = \pi r^2$
- the α shape may decompose in several parts



Stringiness

$$c_{stringy} = \frac{\text{perimeter of MST}}{\text{length of all edges in MST}}$$

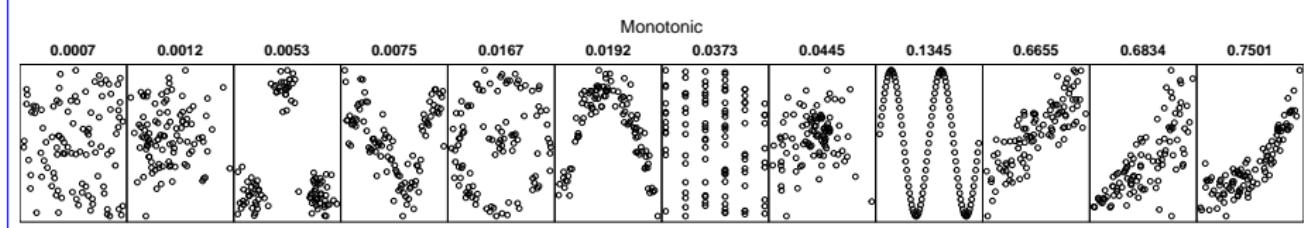
- the perimeter is defined as the longest path without branches
- approximate the MST by a curve



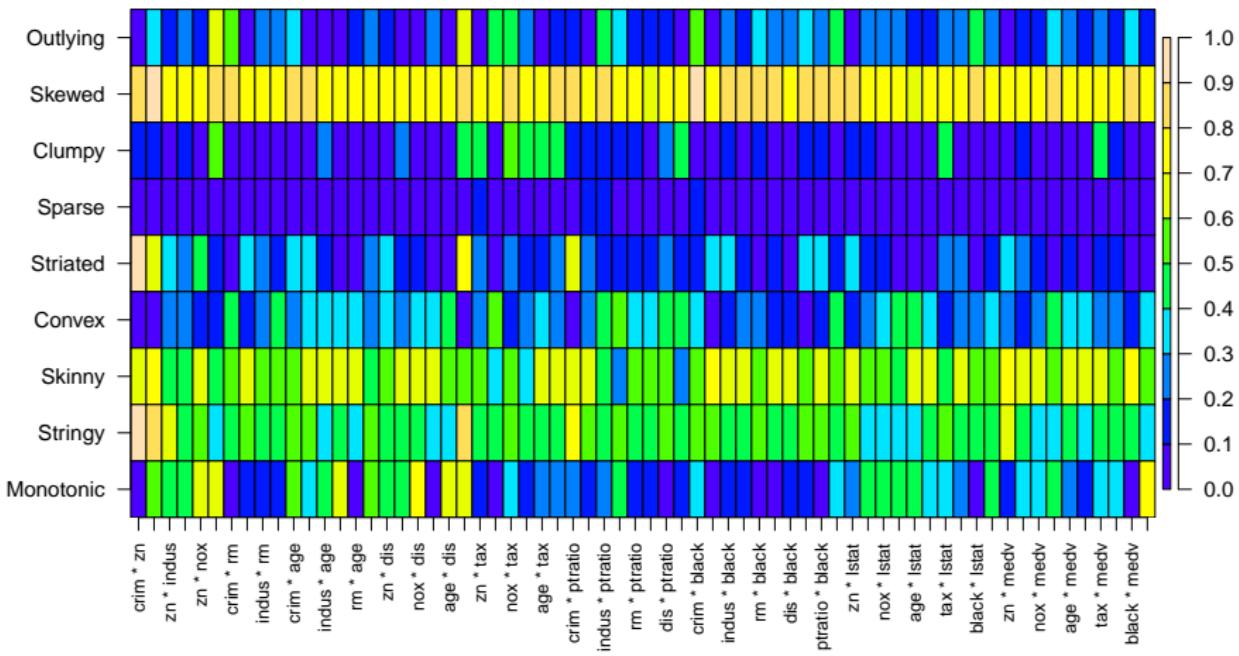
Monotonicity

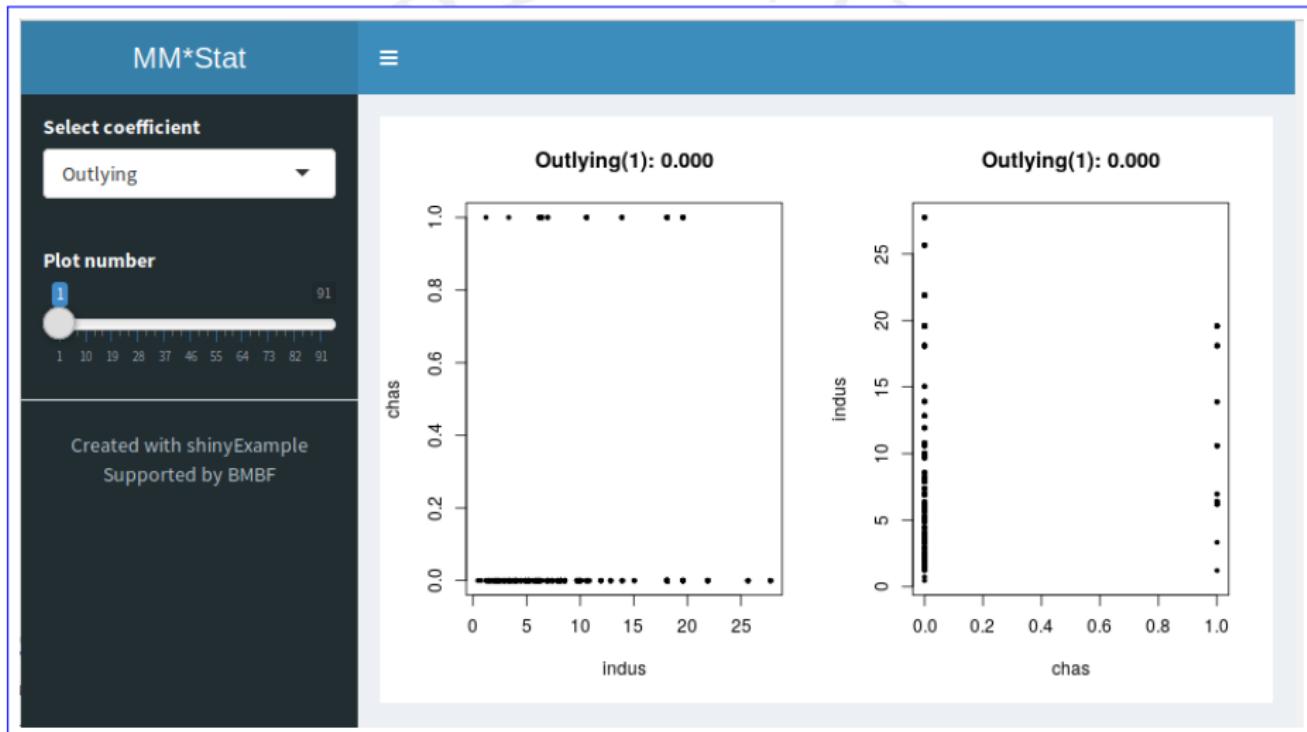
$$c_{\text{monotonic}} = r_s^2$$

- r_s is Spearman's rank correlation coefficient



Scagnostic measures of Boston Housing data





 Listing 12.7: example_scagnostics.R

```
1 library("MASS") # for Boston Housing data
2 library("scagnostics")
3 data <- Boston[,c(-4, -9)]
4 s     <- scagnostics(data)
5 s
6 g     <- scagnosticsGrid(s)
7 # extract max. scagnostics value and plot
8 plots <- apply(s==max(s), 2, any)
9 gps   <- g[plots,]
10 for (i in 1:nrow(gps)) {
11   xi <- gps$x[i]
12   yi <- gps$y[i]
13   plot(data[,xi], data[,yi],
14         xlab=names(data)[xi], ylab=names(data)[yi])
15 }
```

☞ scagnostics::scagnostics(x)
☞ scagnostics::scagnosticsGrid(s)

Trellis display

- Tukey, Tukey (1981), Roesle (1911)
- Visualization: depends on statistical graph
 - ▶ Variables: two (or more) variables
 - ▶ Observations: depends on statistical graph
- Construction:
 - ▶ condition for one or more variables
 - ▶ create one plot for each conditional
- Problems: order of the variables

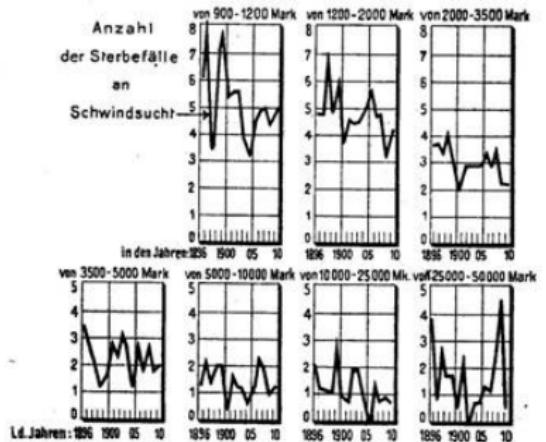
Roesle, E.E., ed. (1911). *Sonderkatalog für die Gruppe Statistik der wissenschaftlichen Abteilung der Internationalen Hygiene-Ausstellung*. Dresden, Germany: Verlag der Internationalen Hygiene-Ausstellung.

Tukey, Paul and Tukey, John (1981). "Graphical Display of Data Sets in 3 or More Dimensions". In: *Interpreting multivariate data. Looking at Multivariate Data* (Mar. 24–27, 1980). Ed. by Vic Barnett. Wiley, pp. 189–275. isbn: 9780471280392.

Einfluß der Wohlhabenheit auf die Häufigkeit der Tuberkulose

auf Grund der Erhebung über die Einkommensverhältnisse der Gestorbenen in der Stadt Hamburg seit 1896

Anzahl der Sterbefälle an Lungentuberkulose, die auf je 1000 Steuerzahler oder deren Angehörige treffen bei einem Einkommen



Reproduktion Nr. 9
Graphische Darstellung Nr. 92

Die Säuglingssterblichkeit in den ersten vier Lebensquartalen

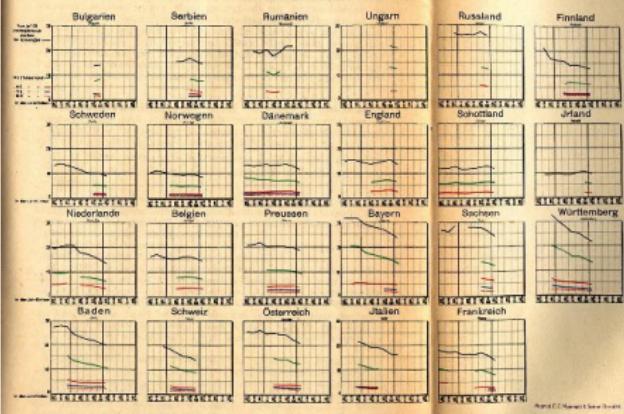
in den europäischen Staaten im Durchschnitt von je 5 Jahren seit 1861

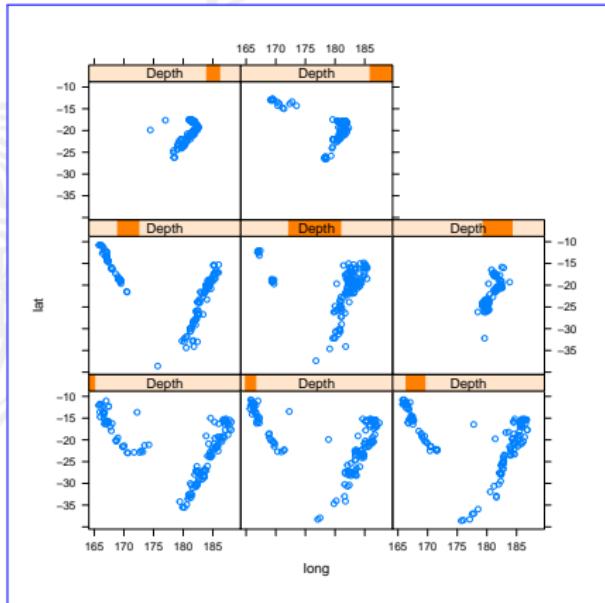
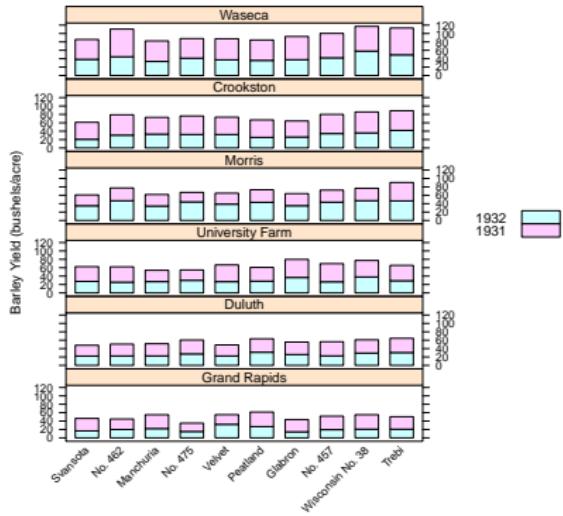
Mittel der infantilen sterblichkeit je 5 jahre verschieden nach der zeit des lebensquartals - Neugeborene gestorben ab 1861

Von je 100 Lebendgeborenen sterben im Durchschnitt jährlich

— im 1., — im 2., — im 3., — im 4. Lebensquartal

— im 1.-4. Lebensquartal = 1 Lebensjahr





 Listing 12.8: example_trellis_xyplot.R

```
1 library("MASS")
2 library("lattice")
3 # scatterplots of lstat and medv by rad
4 xyplot(medv~lstat|factor(rad), data=Boston)
```

 Listing 12.9: example_trellis_xyplot2.R

```
1 library("MASS")
2 library("lattice")
3 # scatterplots of lstat and medv by rm strips
4 rmstrip <- equal.count(Boston$rm, number=9)
5 xyplot(medv~lstat|rmstrip, data=Boston)
```

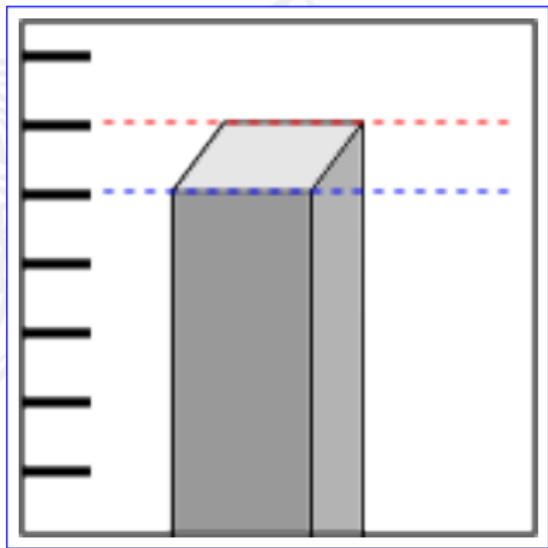
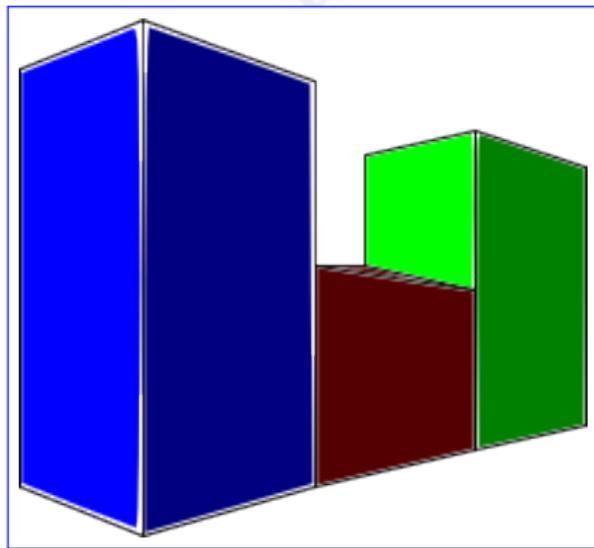
⌚ Listing 12.10: example_trellis_boxplot.R

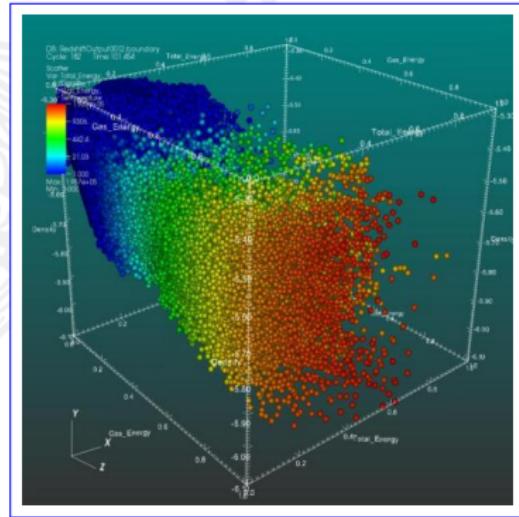
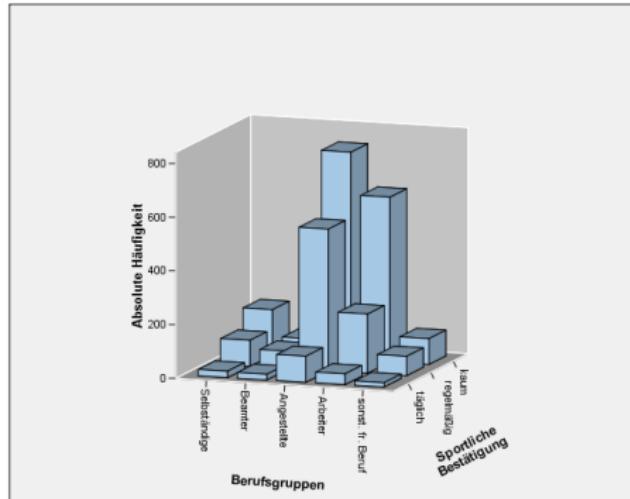
```
1 library("MASS")
2 library("lattice")
3 # boxplots of medv for chas and by rad
4 bwplot(medv~factor(chas)|factor(rad), data=Boston)
```

⌚ lattice::....(formula, data)

3D plots

- Variables: two (or more) variables
- Observations: small
- Visualization: association, frequency
- Problems 3D barchart:
 - ▶ comparison of heights
 - ▶ backward bars
- Problems 3D scatterplot:
 - ▶ backward structure





R Listing 12.11: example_3dscatter.R

```
1 library("MASS") # for Boston Housing data
2 library("scatterplot3d")
3 scatterplot3d(Boston$rm, Boston$lstat, Boston$medv)



---

R persp(x=seq(0, 1, length.out=nrow(z)), y = seq(0, 1,
      length.out=ncol(z)), z, theta=0, phi=15, r=sqrt(3))
R plot3D::persp3D(x=seq(0, 1, length.out=nrow(z)), y = seq(0, 1,
      length.out=ncol(z)),z, colvar = z, phi = 40, theta =
      40)
R plot3D::ribbon3D(params as before)
R plot3D::hist3D(params as before)
R scatterplot3d::scatterplot3d(x, y=NULL, z=NULL, scale.y=1,
      angle=40)
R lattice::cloud(formula, data, perspective=T)
R rgl::plot3d(x, y, z)
```

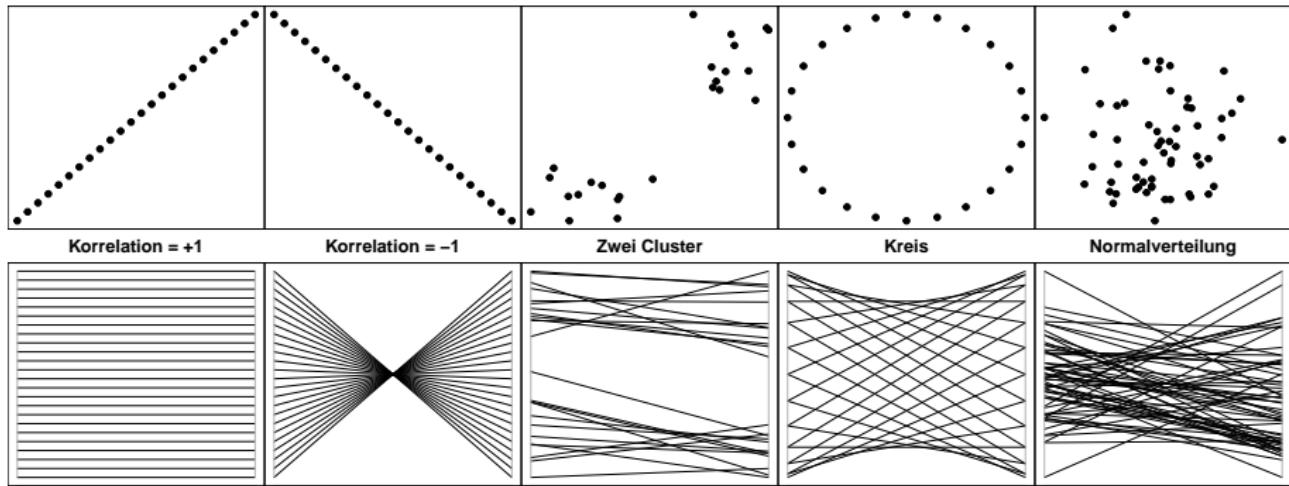
Parallel coordinates

- Visualization: structure
 - ▶ Variables: three (or more) continuous variables
 - ▶ Observations: small
- Construction:
 - ▶ express each observation as a line
 - ▶ give up orthogonality between coordinates
- Problems:
 - ▶ order of the variables
 - ▶ recognition of patterns

d'Ocagne, Maurice (1885). *Coordonnées parallèles et axiales : Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles.*

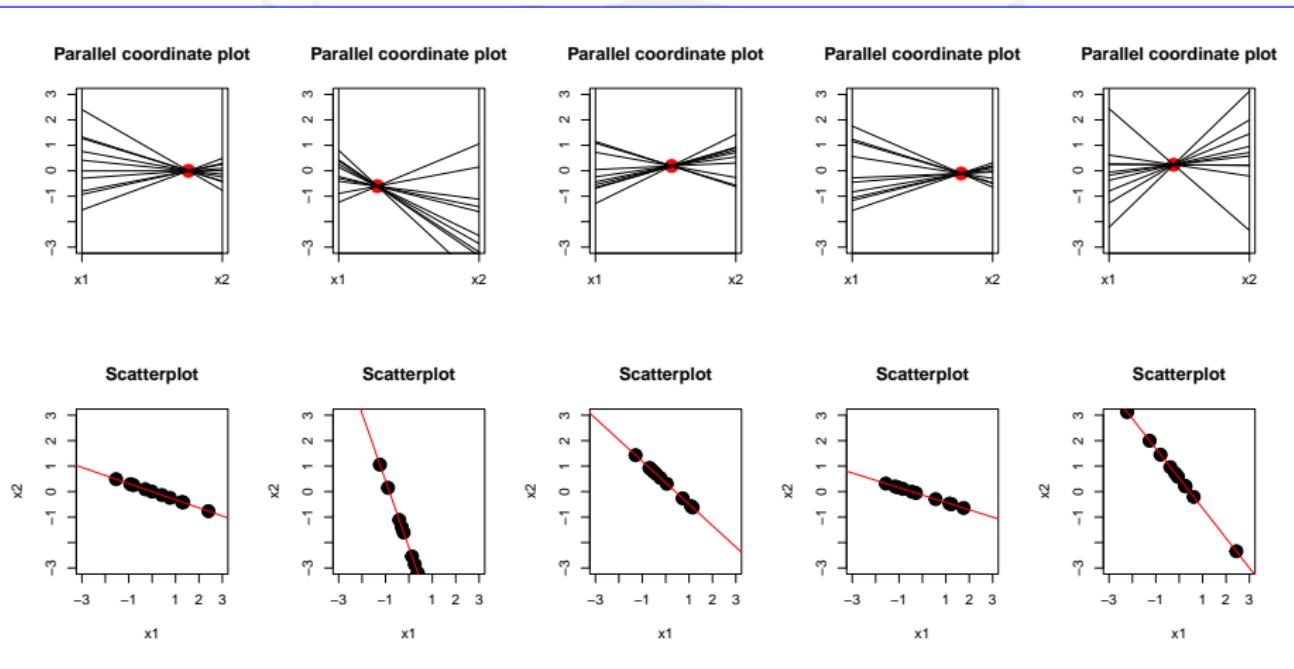
Inselberg, Alfred (Aug. 1985). "The plane with parallel coordinates". In: *The Visual Computer* 1.2, pp. 69–91. issn: 0178-2789, 1432-2315. doi: 10.1007/BF01898350. url: <http://link.springer.com/10.1007/BF01898350> (visited on 08/27/2015).

- translation of scatterplot patterns into parallel coordinate patterns

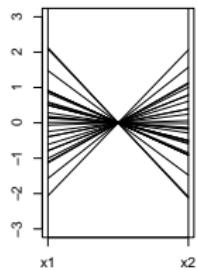


Point line duality

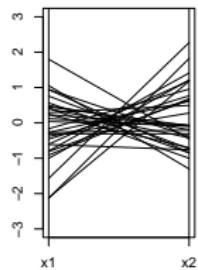
- each point in 2D corresponds to a line in parallel coordinates (black)
- each line in 2D corresponds to a point in parallel coordinates (red)



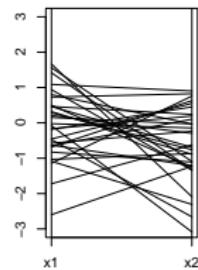
Pcp



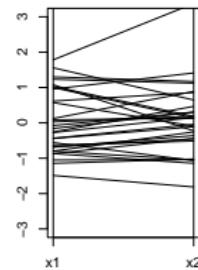
Pcp



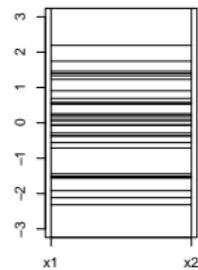
Pcp



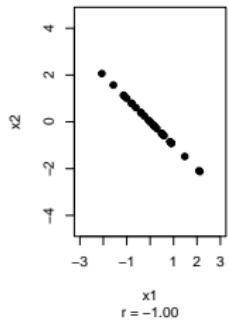
Pcp



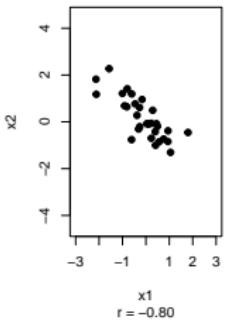
Pcp



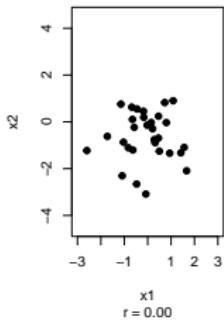
2D



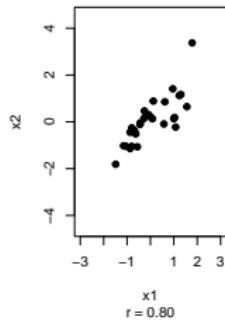
2D



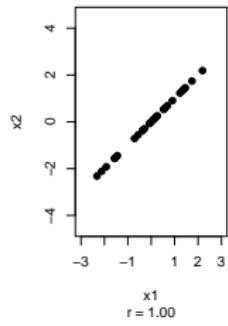
2D

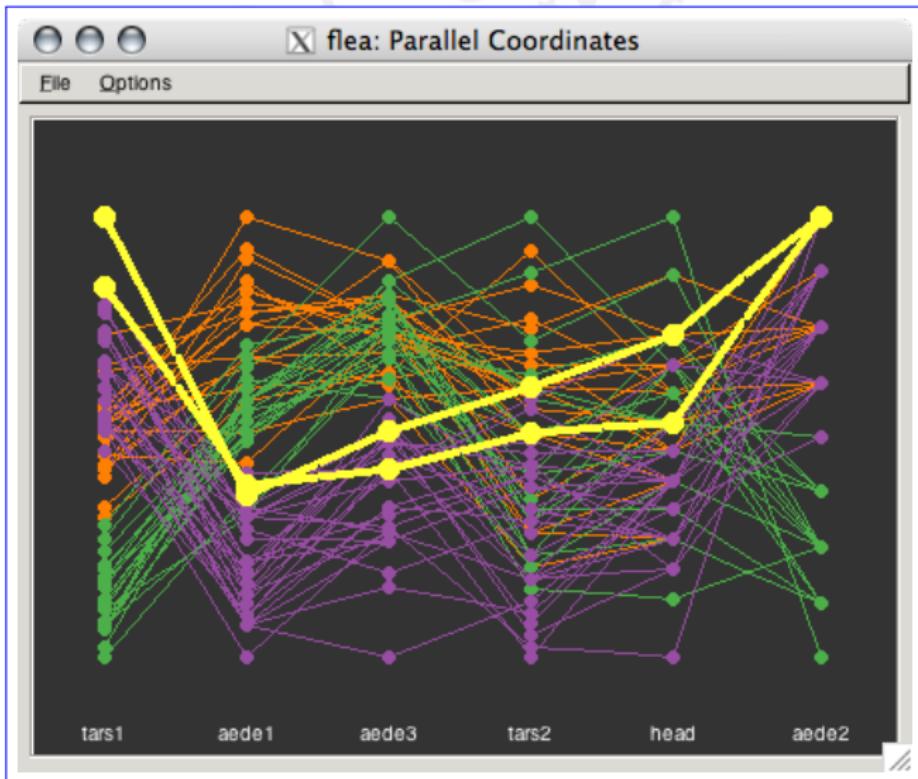


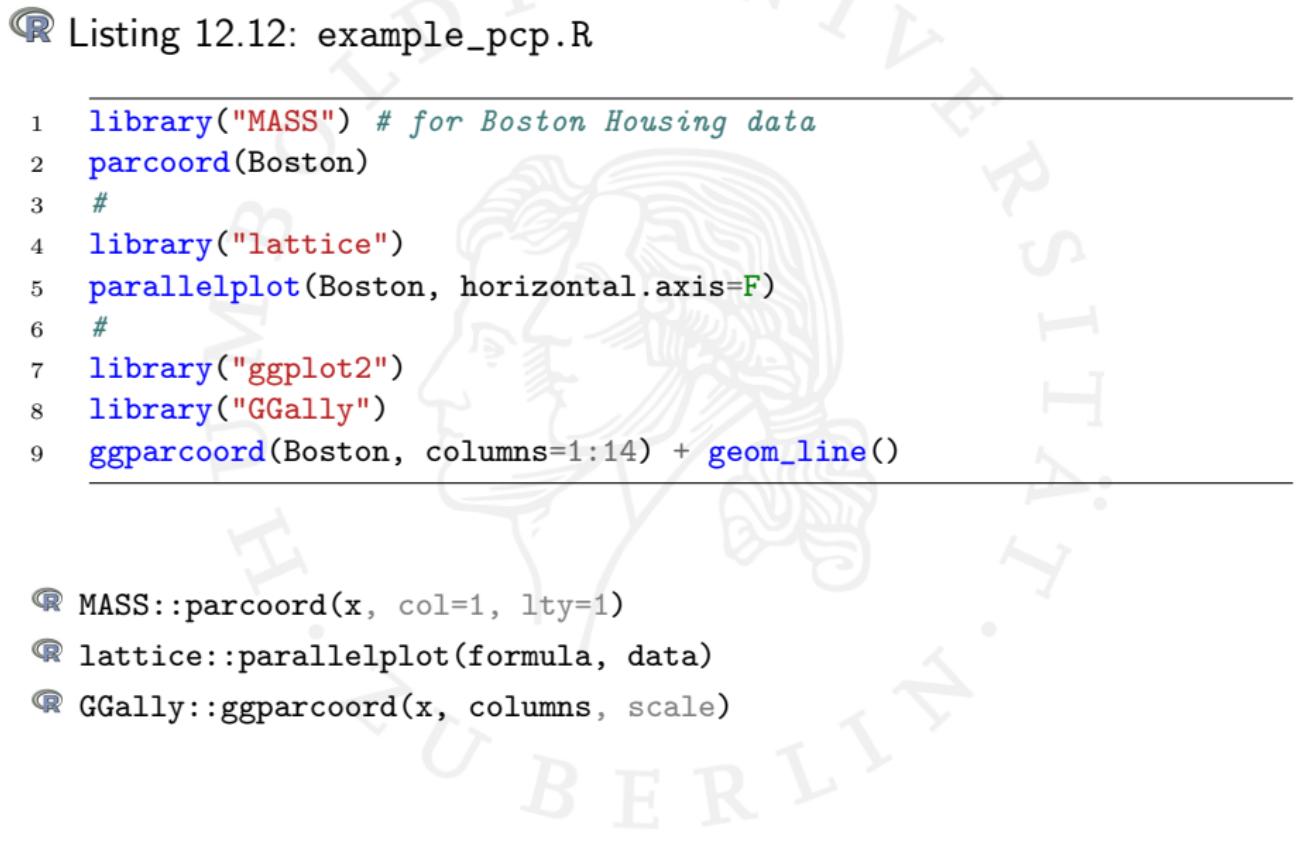
2D



2D





The background of the slide features a large, faint watermark of the HU Berlin logo, which consists of a stylized profile of a person's head and shoulders.

R Listing 12.12: example_pcp.R

```
1 library("MASS") # for Boston Housing data
2 parcoord(Boston)
3 #
4 library("lattice")
5 parallelplot(Boston, horizontal.axis=F)
6 #
7 library("ggplot2")
8 library("GGally")
9 ggparcoord(Boston, columns=1:14) + geom_line()
```

- R MASS::parcoord(x, col=1, lty=1)
- R lattice::parallelplot(formula, data)
- R GGally::ggparcoord(x, columns, scale)

 Listing 12.13: example_pcp_graphics.R

```
1 library("MASS") # for Boston Housing data
2 parcoord(Boston)
```

 Listing 12.14: example_pcp_lattice.R

```
1 library("MASS") # for Boston Housing data
2 library("lattice")
3 parallelplot(Boston, horizontal.axis=F)
```

 Listing 12.15: example_pcp_ggplot.R

```
1 library("MASS") # for Boston Housing data
2 library("ggplot2")
3 library("GGally")
4 ggparcoord(Boston, columns=1:14) + geom_line()
```

 MASS:::parcoord(x, col=1, lty=1)

 lattice:::parallelplot(formula, data)

 GGally:::ggparcoord(x, columns, scale)

Andrews curves

- Visualization: outliers, clusters
 - ▶ Variables: three (or more) continuous variables
 - ▶ Observations: small
- Problem: order of variables!

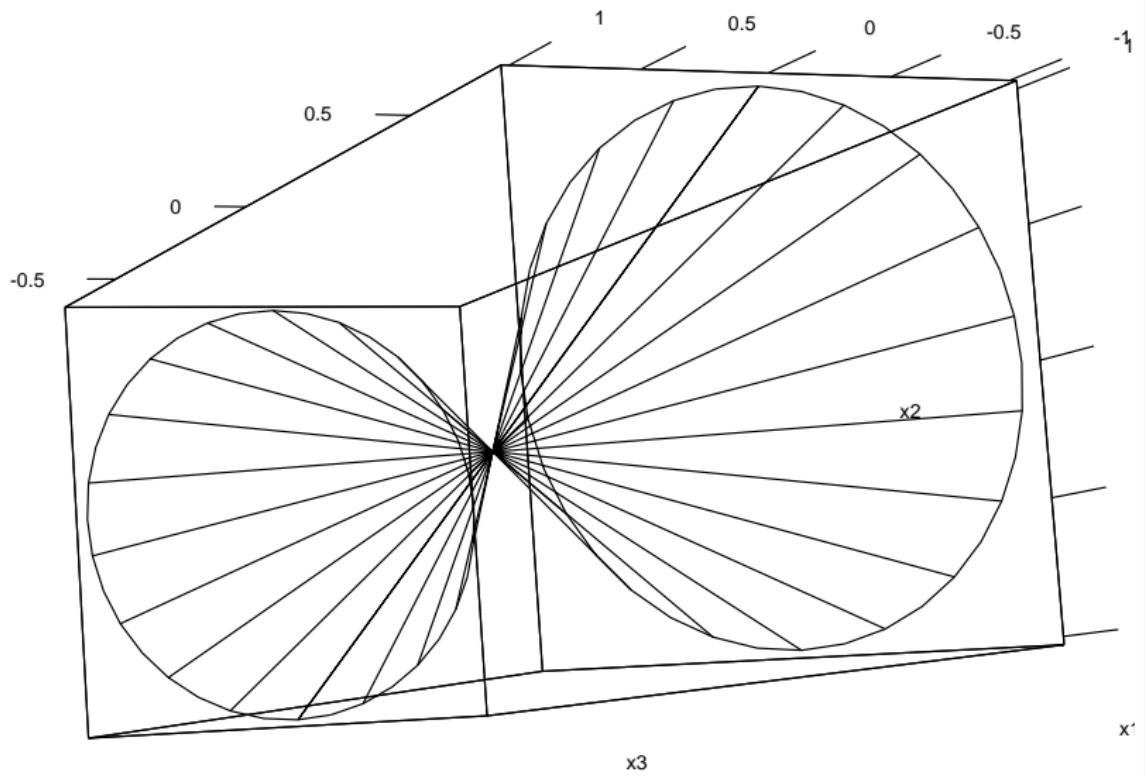
Andrews, D. F. (Mar. 1972). "Plots of High-Dimensional Data". In: *Biometrics* 28.1, p. 125.
issn: 0006341X. doi: 10.2307/2528964. url:
<http://www.jstor.org/stable/2528964?origin=crossref> (visited on 06/23/2016).

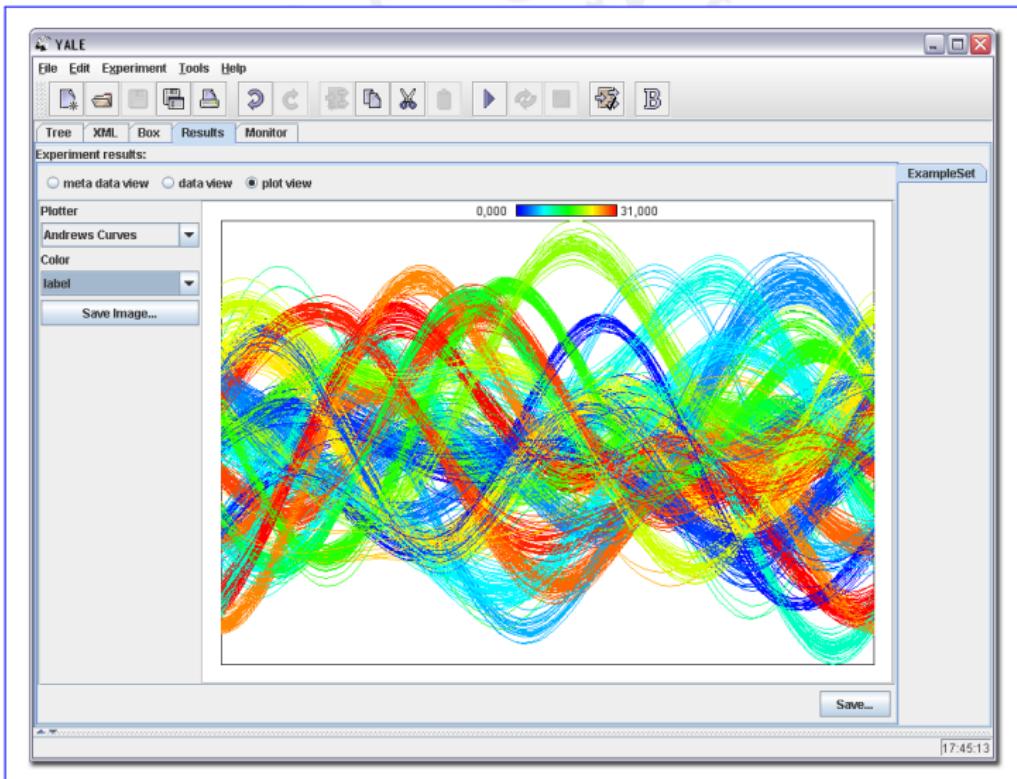
- Each observation $x_i = (x_{i,1}, \dots, x_{i,p})$ will be represented by a line

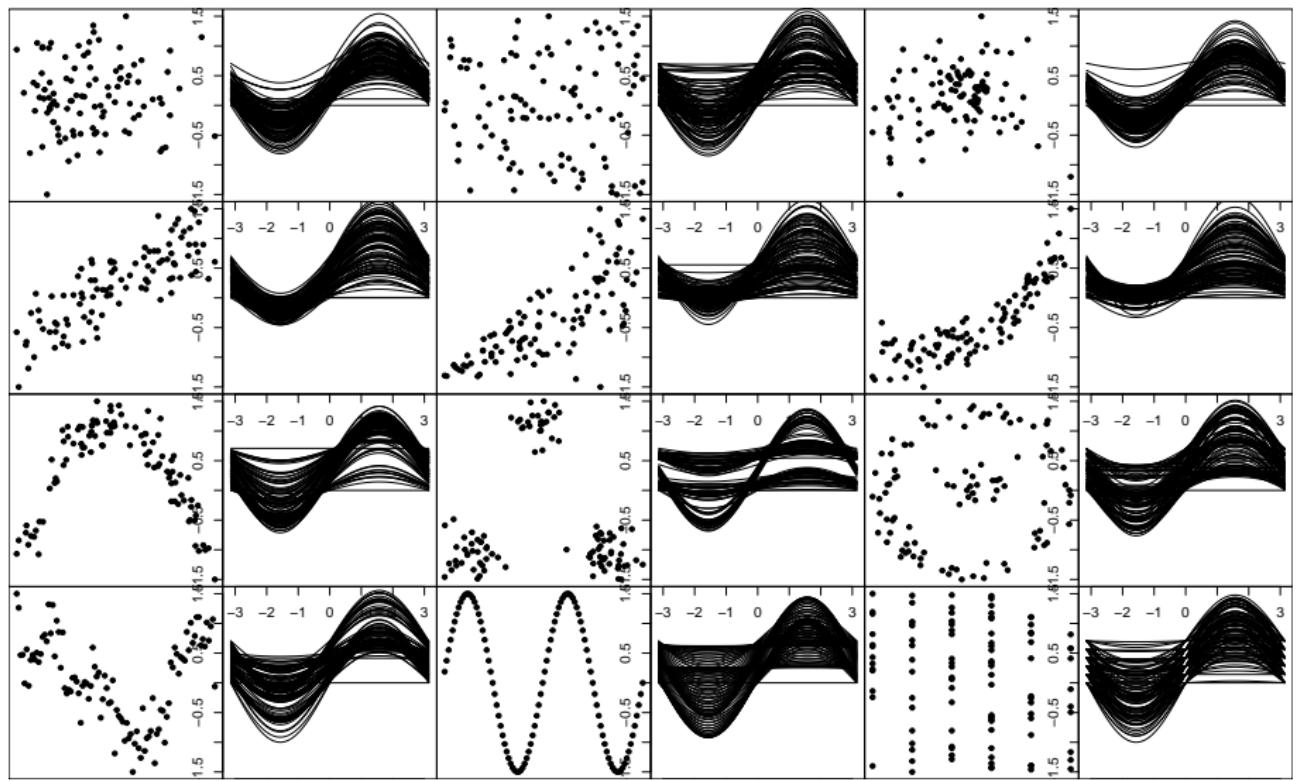
$$f_i(t) = \frac{x_{i,1}}{\sqrt{2}} + x_{i,2} \cos(t) + x_{i,3} \sin(t) + x_{i,4} \cos(2t) + x_{i,5} \sin(2t) + \dots \quad t \in [-\pi; \pi]$$

- It holds

$$\pi \underbrace{\sum_{k=1}^p (x_{i,k} - x_{j,k})^2}_{= \text{squared euclidian distance}} = \pi \|x_i - x_j\|^2 = \underbrace{\int_{-\pi}^{\pi} (f_i(t) - f_j(t))^2 dt}_{\approx \text{area between curves}}$$







 Listing 12.16: example_andrews.R

```
1 library("MASS") # for Boston Housing data
2 library("andrews")
3 andrews(Boston)
```

 andrews::andrews(df, type=1)

type=2 $f_i(t) = x_{i,1} \cos(t) + x_{i,2} \sin(t) + x_{i,3} \cos(2t) + x_{i,4} \sin(2t) + \dots$

type=3 $f_i(t) = x_{i,1} \cos(t) + x_{i,2} \cos(\sqrt{2}t) + x_{i,3} \cos(\sqrt{3}t) + \dots$

type=4
$$\begin{aligned} f_i(t) = \frac{1}{\sqrt{2}} & [x_{i,1} + x_{i,2}(\sin(t) + \cos(t)) + x_{i,3}(\sin(t) - \cos(t)) \\ & + x_{i,4}(\sin(2t) + \cos(2t)) + x_{i,5}(\sin(2t) - \cos(2t)) + \dots] \end{aligned}$$

Radar chart

- Visualization: comparison of observations
 - ▶ Variables: three (or more) continuous variables
 - ▶ Observations: small
- Other names: star diagram/plot, spider chart
- Problems: requires scaling

Mayr, Georg von (1877). *Die Gesetzmässigkeit im Gesellschaftsleben*. R. Oldenbourg.



⌚ Listing 12.17: example_radar.R

```
1 library("MASS") # for Boston Housing data
2 bmin <- apply(Boston, 2, min)
3 bmax <- apply(Boston, 2, max)
4 sBoston <- scale(Boston, center=bmin, scale=bmax-bmin)
5 stars(sBoston[1:25,], scale=F, lwd=2)
```

⌚ stars(x, full=T, scale=T, radius=T)

Chernoff faces

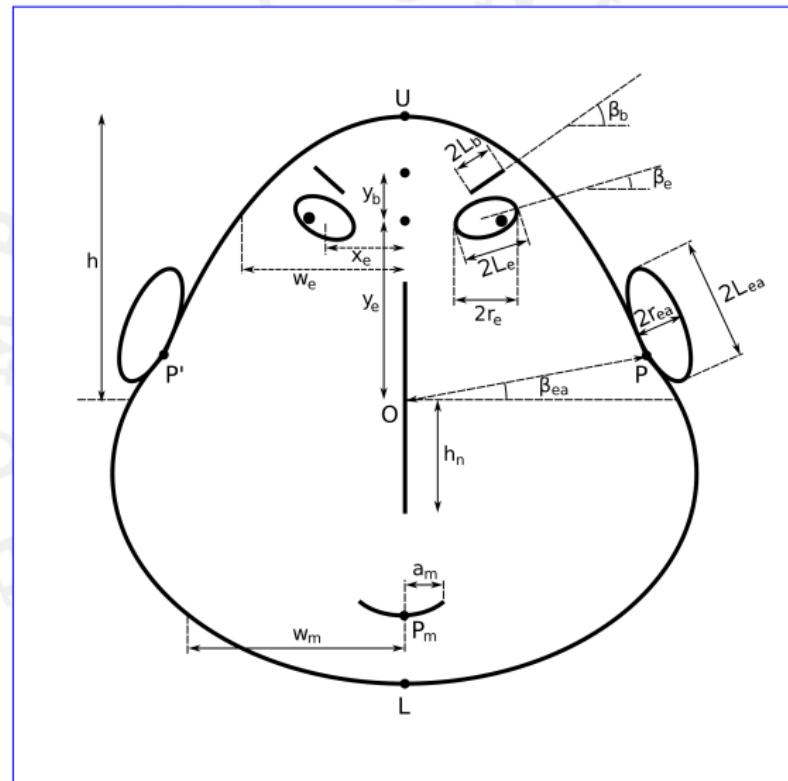
- Visualization: comparison of observations
 - ▶ Variables: three (or more) continuous variables (up to 36)
 - ▶ Observations: small
- Construction: use ability of human brain to distinguish faces
- Problem: coding of variables to face parts
- has been transferred to area specific drawings

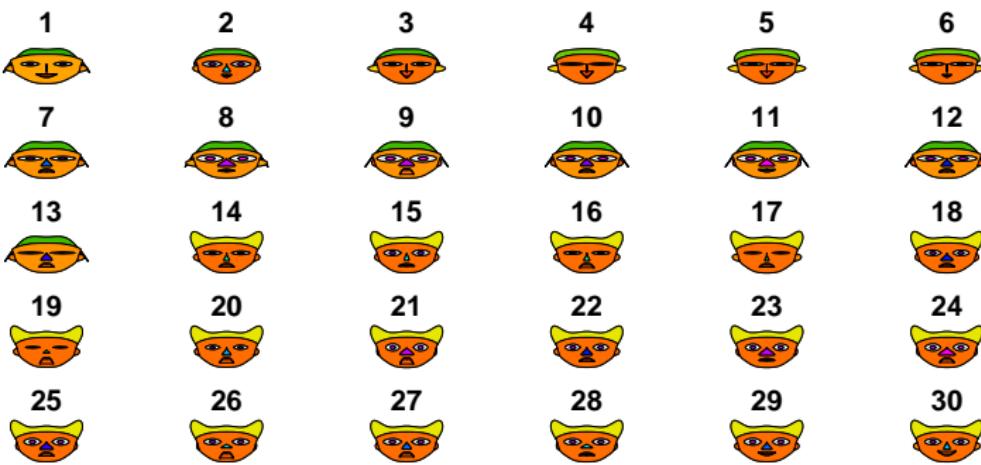
Chernoff, Herman (June 1973). "The Use of Faces to Represent Points in K-Dimensional Space Graphically". In: *Journal of the American Statistical Association* 68.342, p. 361. issn: 01621459. doi: 10.2307/2284077. url:

<http://www.jstor.org/stable/2284077?origin=crossref> (visited on 12/17/2015).

Flury, Bernhard and Riedwyl, Hans (Dec. 1981). "Graphical Representation of Multivariate Data by Means of Asymmetrical Faces". In: *Journal of the American Statistical Association* 76.376, p. 757. issn: 01621459. doi: 10.2307/2287565. url:

<http://www.jstor.org/stable/2287565?origin=crossref> (visited on 12/17/2015).





⌚ Listing 12.18: example_faces.R

```
1 library("MASS") # for Boston Housing data
2 library("aplypack")
3 zBoston = scale(Boston)
4 faces(zBoston[1:30,], scale=F)
```

⌚ aplypack:faces(xy, face.type=1, scale=F)

Principal component analysis

November 3, 2022

- Best fitting line
- Maximizing variance
- Component computation
- Dimension reduction
- Swiss Banknote data
- Covariance vs. correlation
- Correlation between PCs and variables
- Scree plot
- How many components
- Parallel analysis of Horn
- Asymptotic properties

Best fitting line

- For p dimensional dataset find the “best” fitting line

$$\min_{(a_1, b_1)} \sum_{i=1}^n \| (a_1 t_{1i} + b_1) - x_i \|^2$$

with $a_1, b_1 \in \mathbb{R}^p$ and $t_{1i} \in \mathbb{R}$

- repeat the process to find a_i orthogonal to a_j ($j = 1, \dots, i-1$)
- a_1, \dots, a_p represent a rotation of the coordinate system
- for identification we need
 - ▶ $\| a_j \|^2 = 1$
 - ▶ $\| b_j \|^2 \rightarrow \min.$

Pearson, Karl (Nov. 1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine Series 6* 2.11, pp. 559–572. issn: 1941-5982, 1941-5990. doi: 10.1080/14786440109462720. url: <http://www.tandfonline.com/doi/abs/10.1080/14786440109462720> (visited on 12/17/2015).

MM*Stat



Line angle (in degree)

0 180

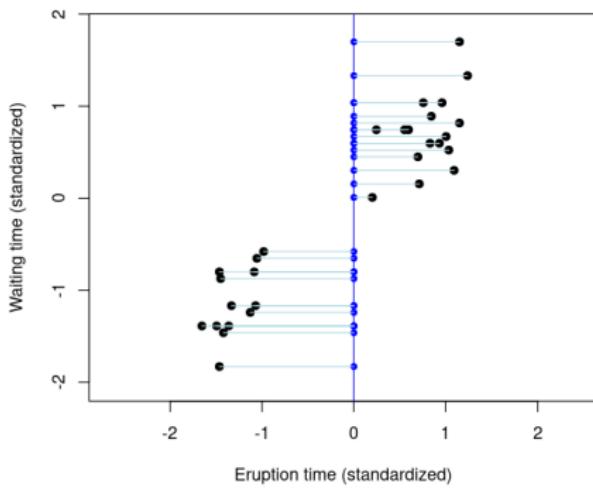
0 18 36 54 72 90 108 144 180

Total distance

30.94

PCA - Best line

Subsample of Old Faithful geyser data



Maximizing variance

- For p dimensional dataset find the line which maximize the variance of

$$\max_{a_1} \text{Var}(a_1^T x)$$

with $a_1 \in \mathbb{R}^p$

- repeat the process to find a_i orthogonal to a_j ($j = 1, \dots, i - 1$)
- a_1, \dots, a_p represent a rotation of the coordinate system
- for identification we need
 - ▶ $\|a_j\|^2 = 1$

MM*Stat

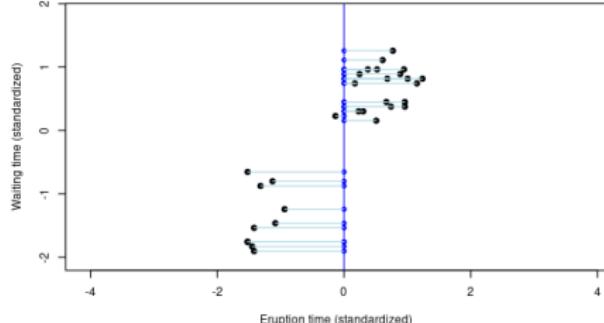
Line angle (in degree)



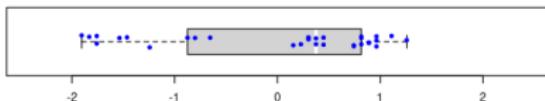
≡

PCA - Maximizing variance

Subsample of Old Faithful geyser data



Projected data points



Total distance

26.43

Variance of
projected data
points

1.11

Explained total
variance

53%

Component computation

- the vectors a_j can be found by computing the p eigenvalues and -vectors of the covariance matrix S_x
 - ▶ $S_x u = \lambda_u u$
 - ▶ u eigenvector
 - ▶ λ_u eigenvalue
- the “rotated” principal components can be computed by

$$PC = S_x^{-1/2}(x - \bar{x})$$

- the eigenvalues λ_j are

$$\begin{aligned}Var(PC_j) &= \lambda_j \\Cov(PC_i, PC_j) &= 0 \text{ for } i \neq j\end{aligned}$$

 Listing 13.1: example_pca.R

```
1 library("rio")
2 data("bank2", package="mmstat4")
3 # do PCA (covariance)
4 pc <- prcomp(bank2)
5 pc
6 # what R delivers
7 summary(pc)
8 par(mfrow=c(1,2))
9 plot(pc, main="Scree plot as bar chart")
10 plot(pc$sdev^2, type="b", main="Scree plot")
```

⌚ `prcomp(x, center=F, scale=F)`

⌚ `princomp(xcenter=F, scale=F)`

⚠ `prcomp` uses singular value decomposition (more stable) and `princomp` eigenvalues from the covariance matrix

Dimension reduction

Measuring the variability of a data set

- univariate: variance, interquartile range

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- bivariate: covariance, correlation (indirect)

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- generalized variance $GV = \det(S)$

\sqrt{GV} = volume of space occupied by data points

$$GV_2 = \begin{vmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{vmatrix} = s_x^2 s_y^2 - s_{xy}^2, \text{ std. variables: } 1 - r_{xy}^2$$

- reduce the dimensionality of the dataset without “loss” of information
- the information criteria is the total variance

$$T = \sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \text{Var}(PC_i)$$

- explained total variance by $q < p$ principal components

$$\frac{1}{T} \sum_{i=1}^q \text{Var}(PC_i)$$

- the total variance is invariant under rotations

$$\begin{aligned} T &= \sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \\ &= \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^p (x_{ij} - \bar{x}_i)^2 = \frac{1}{n} \sum_{j=1}^n d^2(x_j, \bar{x}) \end{aligned}$$

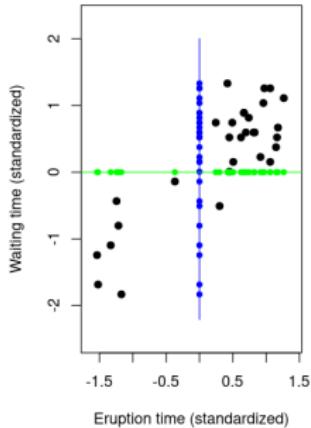
MM*Stat

Line angle (in degree)

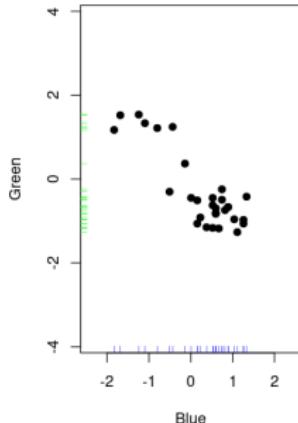


PCA - Dimension reduction

Subsample of Old Faithful geyser data



Projected data points



Variance of projected data points (blue)

0.74

Explained total variance (blue)

48%

Variance of projected data points (green)

0.82

Explained total variance (green)

52%

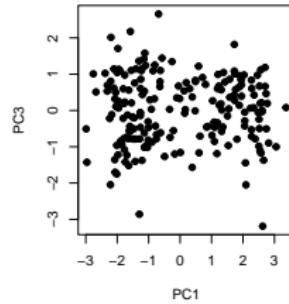
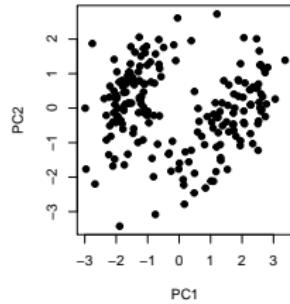
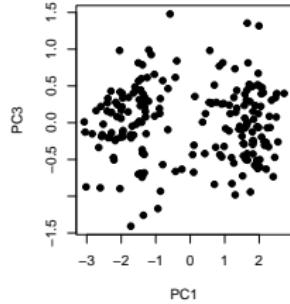
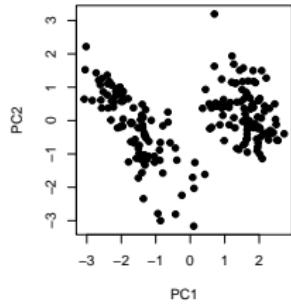
Swiss Banknote data

- Variables
 - ▶ width (X_1), left height (X_2), right height (X_3), bottom frame (X_4), upper frame (X_5) and diagonal (X_6)
- 100 genuine and 100 counterfeit banknotes
 - ▶ can we distinguish genuine and counterfeit banknotes?

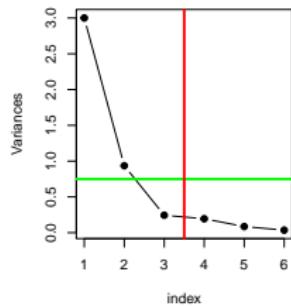


Reverse of the Swiss 1000 franc note of the second banknote series (1911)

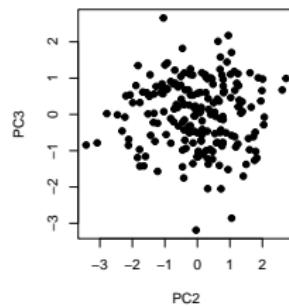
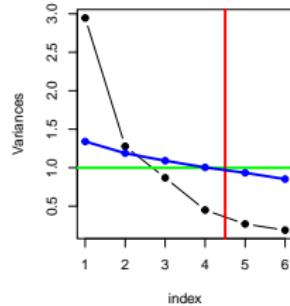
Covariance vs. correlation



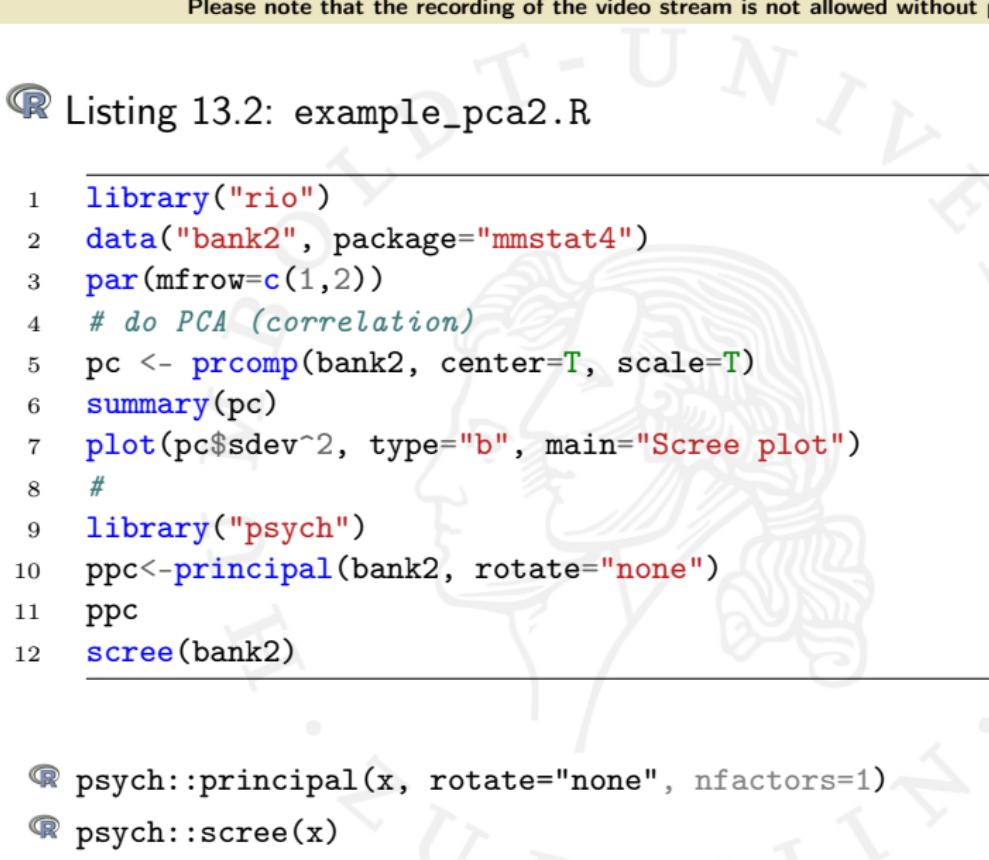
Scree (covariance)



Scree (correlation)



- Use covariance matrix
 - ▶ all variables in PCA are measured in the same units
 - ▶ the variances of all variables is of the same magnitude
 - ▶ default in R (base)
- Use correlation matrix
 - ▶ if variables in PCA are not measured in the same units
 - ▶ if the variance of a set of variables dominate the other variables
 - ▶ default in SPSS
 - ▶ default in R (psych)

A large, faint watermark of a stylized brain in profile, facing right, is centered behind the text.

R Listing 13.2: example_pca2.R

```
1 library("rio")
2 data("bank2", package="mmstat4")
3 par(mfrow=c(1,2))
4 # do PCA (correlation)
5 pc <- prcomp(bank2, center=T, scale=T)
6 summary(pc)
7 plot(pc$sdev^2, type="b", main="Scree plot")
8 #
9 library("psych")
10 ppc<-principal(bank2, rotate="none")
11 ppc
12 scree(bank2)
```

R psych::principal(x, rotate="none", nfactors=1)
R psych::scree(x)

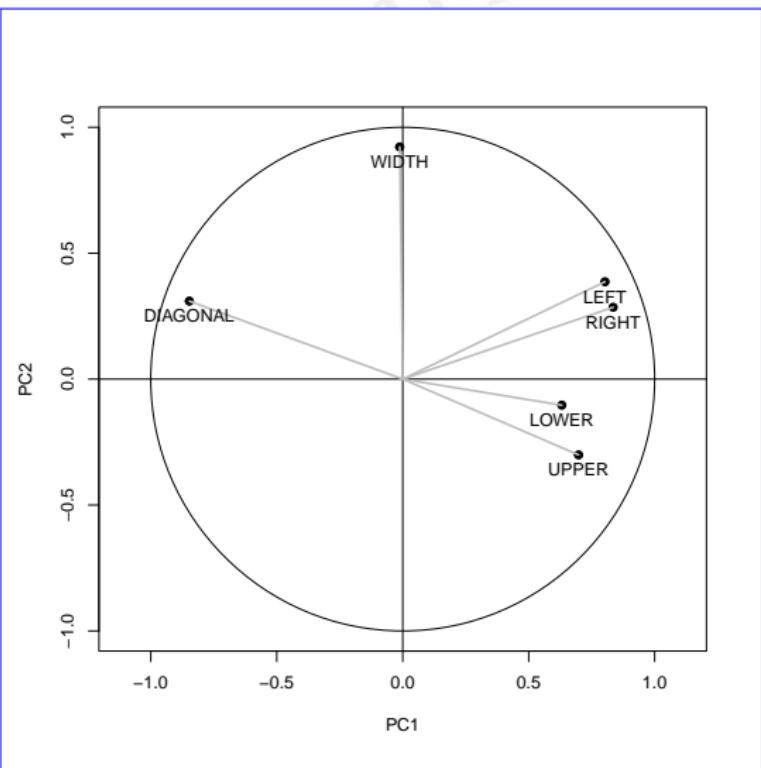
Correlation between PCs and variables

For each variable X_j (with $j=1, \dots, p$) holds

$$X_j = a_{j1}PC_1 + \dots + a_{jp}PC_p$$

$$\begin{aligned} \text{cor}(X_j, PC_k) &= \text{cor}(a_{j1}PC_1 + \dots + a_{jp}PC_p, PC_k) \\ &= a_{j1} \underbrace{\text{cor}(PC_1, PC_k)}_{=0} + \dots + a_{jk} \underbrace{\text{cor}(PC_k, PC_k)}_{=1} \\ &\quad + \dots + a_{jp} \underbrace{\text{cor}(PC_p, PC_k)}_{=0} \\ \text{cor}(X_j, PC_k) &= a_{jk} \end{aligned}$$

Rule-of-thumb: if $|a_{jk}| \geq 0.5$ then PC_k is (strongly) related to X_j

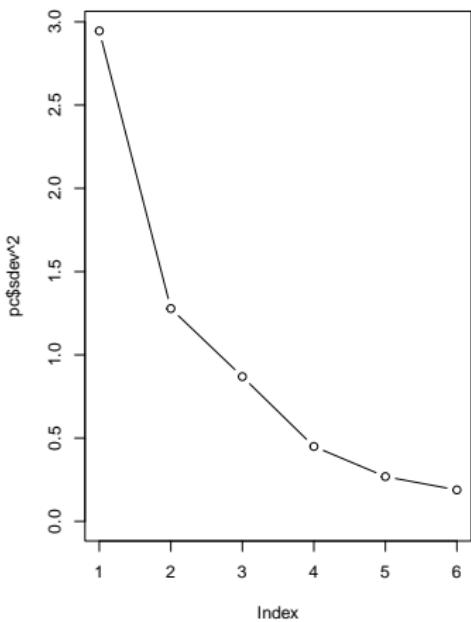


	PC1	PC2
WIDTH	-0.01	0.92
LEFT	0.80	0.39
RIGHT	0.84	0.29
UPPER	0.70	-0.30
LOWER	0.63	-0.10
DIAGONAL	-0.85	0.31

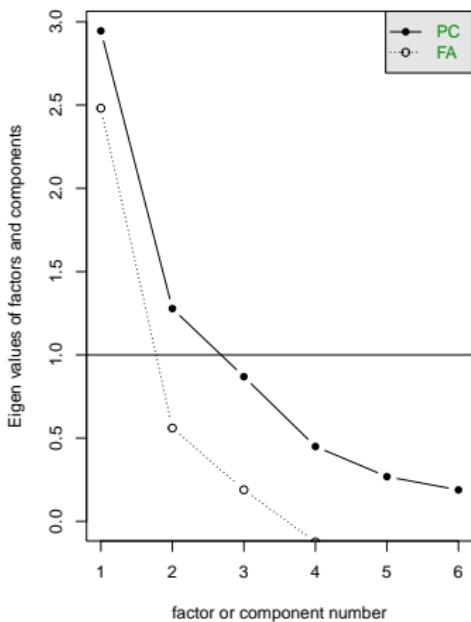
PCA with standardized data

Scree plot

Scree plot (handmade)



Scree plot (psych)



Cattell, Raymond B. (Apr. 1966). "The Scree Test For The Number Of Factors". In: *Multivariate Behavioral Research* 1.2, pp. 245–276. issn: 0027-3171, 1532-7906. doi: 10.1207/s15327906mbr0102_10. url:

How many components

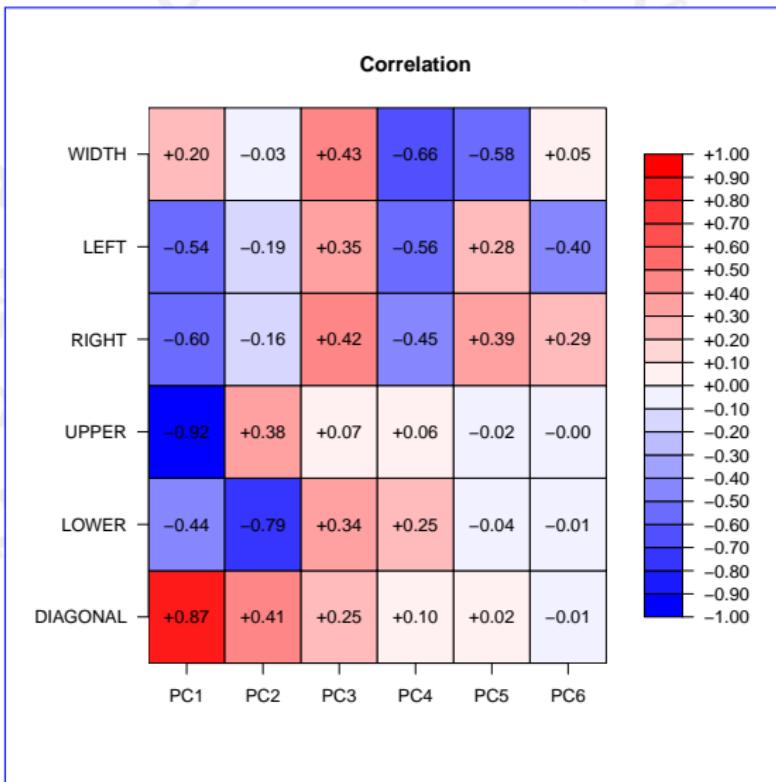
- Ellbow criterium in the Scree plot
- 90% criterium

$$\frac{\lambda_1 + \dots + \lambda_q}{T} \geq 0.9$$

- ▶ How much of the distance between the center of the data and the data points are captured by q components?
- ▶ How much of the *variability* of the data are captured?
- Kaiser criterium
- Parallel analysis of Horn

$$\lambda_i \geq \frac{T}{p}$$

- Interpretability

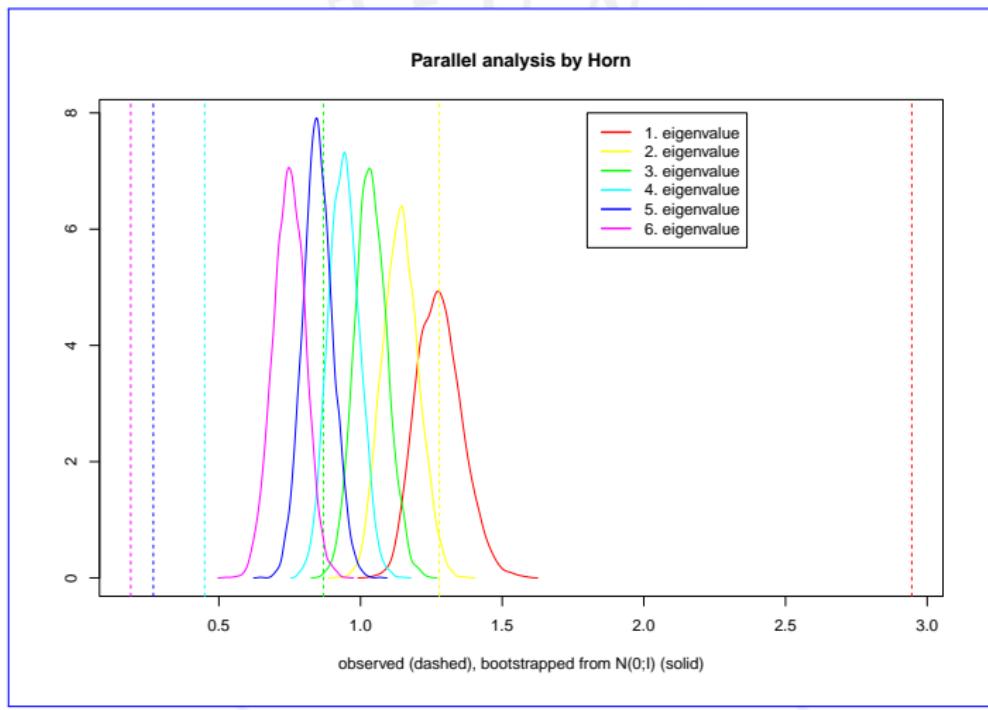


Parallel analysis of Horn

- extension of Kaiser criterium (for standardized data)
- sample B times a dataset from $N_p(0, I)$
- compute eigenvalues l_i for each sample
- sampling error implies that largest eigenvalue $l_1 \geq 1$
- choose $\lambda_i \geq \text{mean}(l_i)$'s or $\lambda_i \geq 95\%$ -quantile of l_i 's

	ev.data	ev.horn.mean	ev.horn.q95
1	2.946	1.279	1.417
2	1.278	1.141	1.247
3	0.869	1.037	1.134
4	0.450	0.942	1.031
5	0.269	0.849	0.937
6	0.189	0.748	0.838

n=200, p=6, B=5000



Horn, John L. (June 1965). "A rationale and test for the number of factors in factor analysis". In: *Psychometrika* 30.2, pp. 179–185. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02289447. url: <http://link.springer.com/10.1007/BF02289447> (visited on 12/17/2015).



Listing 13.3: example_paran.R

```
1 library("rio")
2 data("bank2", package="mmstat4")
3 par(mfrow=c(1,2))
4 # do parallel analysis
5 library("paran")
6 paran(bank2, centile=95, all=T, graph=T)
7 # adjusted ev = unadjusted - bias (random ev-1)
8 library("psych")
9 fa.parallel(bank2, fa="pc")
```

R `paran::paran(x, iterations=0, centile=0, quietly=FALSE, mat=NA,
n=NA, all=F, graph=F)`

R `psych::fa.parallel(x, fa="pc", n.iter=20)`

Asymptotic properties

If $(X_1, \dots, X_p) \sim N_p(\mu, \Sigma)$ then holds

$$\sqrt{n-1}(\hat{\lambda}_j - \lambda_j) \rightarrow N(0; 2\lambda_j)$$

$$\sqrt{\frac{n-1}{2}}(\log(\hat{\lambda}_j) - \log(\lambda_j)) \rightarrow N(0; 1)$$

$$\sqrt{n-1}(\hat{\psi} - \psi) \rightarrow N(0; \omega^2)$$

$$\omega = \frac{2\text{tr}(\Sigma)}{\text{tr}^2(\Sigma)}(\psi^2 - 2\beta\psi + \beta)$$

$$\psi = \frac{\lambda_1 + \dots + \lambda_q}{\lambda_1 + \dots + \lambda_p}$$

$$\beta = \frac{\lambda_1^2 + \dots + \lambda_q^2}{\lambda_1^2 + \dots + \lambda_p^2}$$

Exploratory factor analysis

November 3, 2022

- Exploratory factor analysis
- Factor analysis vs. Principal components
- Basic equations
- Invariance of scale
- Suitability of factor analysis
- Factor analysis methods
- How many factors to choose?
- Rotations
- Pattern and structure matrix
- A case study: the WiWi evaluation forms
- Scores

Exploratory factor analysis

- Belief: behind a large number of observed variables is a small number of latent variables
 - ▶ observed variables are also called items
 - ▶ latent variables are also called unobserved variables, factors
- Correlation between observed variables stems (mainly) from the latent variable

Example 14.28

- How to measure "intelligence"?
- Make a lot of tests and extract a latent variable

Spearman, C. (Apr. 1904). ""General Intelligence," Objectively Determined and Measured". In: *The American Journal of Psychology* 15.2, p. 201. issn: 00029556. doi: 10.2307/1412107. url: <http://www.jstor.org/stable/1412107?origin=crossref> (visited on 12/17/2015).

Factor analysis vs. Principal components

- Principal components
 - ▶ PC starts with "best" fitting line or maximizing variance (\rightarrow descriptive)
 - ▶ and ends up with a linear model
- Factor analysis
 - ▶ starts with a linear model (\rightarrow model based)
 - ▶ PC is one possible model
 - ▶ rotation for interpretability

Latent/Observed	Metrical	Categorical
Metrical	Factor analysis	Latent trait
Categorical	Latent profile	Latent class

Basic equations

Z_i observed variables

$A = (a_{ij})$ loadings

F_j common factors

U_i specific factors

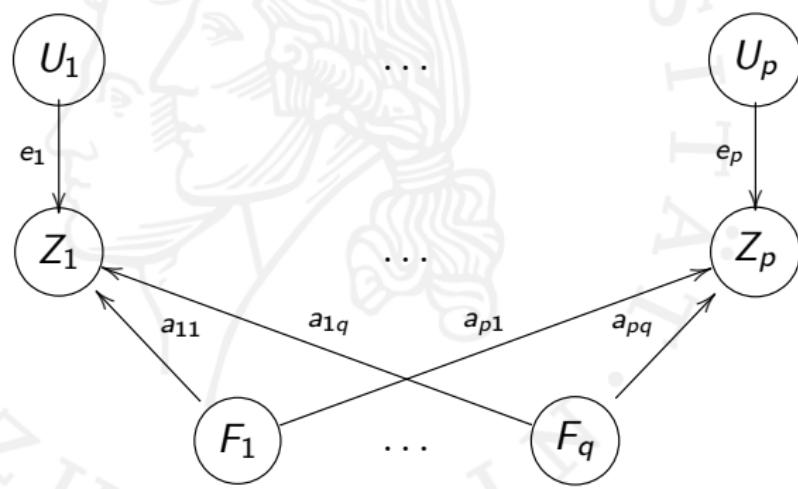
(factors are standardized)

E diagonal matrix

specific
factors

observed
variables

common
factors



$$Z = FA^T + UE$$

$$z_{ij} = a_{j1}f_{i1} + \dots + a_{jq}f_{iq} + e_j u_{ij}$$

with

$$\bar{z}_j = 0, s_{z_j}^2 = 1 \quad j = 1, \dots, p \text{ (standardized variables)}$$

$$\bar{f}_k = 0, s_{f_k}^2 = 1 \quad k = 1, \dots, q \text{ (common factors)}$$

$$\bar{u}_j = 0, s_{u_j}^2 = 1 \quad j = 1, \dots, p \text{ (specific factors)}$$

Further assumptions:

- independent observations and metric variables
- large number of observations and approximately normally distributed variables (*)

If the F_j , the U_i and all F_j and U_i are uncorrelated then follows for the correlation matrix R

$$\begin{aligned}
 R &= \frac{Z^T Z}{n - 1} \\
 &= \frac{(FA^T + UE)^T (FA^T + UE)}{n - 1} \\
 &= \frac{(AF^T + EU^T)(FA^T + UE)}{n - 1} \\
 &= A \underbrace{\frac{F^T F}{n - 1}}_{=I} A^T + A \underbrace{\frac{F^T U}{n - 1}}_{=0} E + E \underbrace{\frac{U^T F}{n - 1}}_{=0} A^T + E \underbrace{\frac{U^T U}{n - 1}}_{=I} E^T \\
 &= AA^T + EE
 \end{aligned}$$

For the diagonal elements of R holds

$$1 = r_{jj} = \underbrace{\sum_{i=1}^q a_{ji}^2}_{h_j^2 \text{ communality}} + e_j^2$$

- Communality h_j^2
 - ▶ how much of the variability is captured by q common factors
 - ▶ Heywood case: one or more communalities are one
 - ▶ Ultra-Heywood case: one or more communalities are larger than one (invalid model !)
 - ★ bad prior communality estimates, too many/few common factors, not enough data for stable estimates, model is not appropriate for data
 - ▶ like R^2 in linear regression
- Computation of factor scores f_{ij} (if $E = 0$)

$$F = ZA(A^T A)^{-1}$$

Invariance of scale

$$\begin{aligned} X &= A_X^T F + U E_X \\ Y &= CX \quad (C \text{ diagonal matrix}) \\ &= \underbrace{CA_X^T}_{{A_Y}^T} F + \underbrace{U C E_X}_{{E_Y}} \\ &= A_Y^T F + U E_Y \end{aligned}$$

- rescaling the data changes
 - ▶ factor loadings
 - ▶ the variance of the specific factors
 - ▶ but not the factors itself
- using the correlation matrix is basically a change of scale

Suitability of factor analysis

- Assumption: the model assumes that some factors linearly influence the observed model
- Check beforehand
 - ▶ inverse correlation matrix
 - ▶ Anti-image matrix
 - ▶ Kaiser-Meyer-Olkin-Criteria
 - ▶ Bartlett's test of sphericity
- Check afterwards
 - ▶ communalities h_j^2
 - ▶ reproduced correlation matrix \hat{R}

Assumption(s): X multivariate normal

Hypotheses: $H_0 : R = I_p$ vs. $H_1 : R \neq I_p$

Test statistics: $V = -\left(n - 1 - \frac{2p+5}{6}\right) \log(|R|) \approx \chi^2_{p(p-1)/2}$
 $|R|$ determinant of R

n, p number of observations and variables

Reject H_0 : $v > \chi^2_{p(p-1)/2; 1-\alpha}$

Note:
sensitive against violation of normality
Bartlett's test of sphericity

- inverse correlation matrix $P = R^{-1}$
 - ▶ Guttman: if the model holds then the non-diagonal elements of R^{-1} must be close to zero (relative to the diagonal element)
- partial correlation $r_{ij,\bullet} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}$
 - ▶ regress X_i and X_j by all other variables
 - ▶ partial correlation is the correlation of the residuals
 - ▶ correlation of X_i and X_j by excluding the influence of all other variables
 - ▶ Note: if a factor influences at least three variables then $r_{ij,\bullet}$ must be small

Guttman, Louis (Dec. 1953). "Image theory for the structure of quantitative variates". In: *Psychometrika* 18.4, pp. 277–296. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02289264. url: <http://link.springer.com/10.1007/BF02289264> (visited on 12/17/2015).

- Anti-image matrix

- ▶ off-diagonal: negative partial correlation coefficients $-r_{ij,\bullet}$
- ▶ diagonal: measure of sampling adequacy (Kaiser 1970)

$$MSA_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} r_{ij,\bullet}^2}$$

unacceptable 0.5 <	miserable 0.5-0.6	mediocre 0.6-0.7
middling 0.7-0.8	meritorious 0.8-0.9	marvelous >0.9

- Kaiser-Meyer-Olkin-criteria

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} r_{ij,\bullet}^2}$$

Kaiser, Henry F. (Dec. 1970). "A second generation little jiffy". In: *Psychometrika* 35.4, pp. 401–415. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02291817. url: <http://link.springer.com/10.1007/BF02291817> (visited on 12/17/2015).

Kaiser, H. F. and Rice, J. (Apr. 1, 1974). "Little Jiffy, Mark IV". In: *Educational and Psychological Measurement* 34.1, pp. 111–117. issn: 0013-1644. doi: 10.1177/001316447403400115. url: <http://epm.sagepub.com/cgi/doi/10.1177/001316447403400115> (visited on 12/17/2015).

 Listing 14.1: example_efa_suitability.R

```
1 library("psych")
2 library("lattice")
3 # bfi: 25 personality self report items taken from
4 # the International Personality Item Pool
5 names(bfi)
6 bfi2 <- na.omit(bfi[,1:25])
7 # inverse and partial correlations & anti-image
8 p <- solve(cor(bfi2, use="complete.obs"))
9 print(levelplot(p, main="Inverse & partial correlations"), split=c(1,1))
10 pr <- -p/sqrt(outer(diag(p), diag(p)))
11 print(levelplot(pr, main="Anti-Image correlation"), split=c(2,1,2,1), r)
12 # Kaiser-Meyer-Olkin & MSA
13 KMO(bfi2)
14 # Bartlett test of sphericity
15 cortest.bartlett(bfi2)
```

④ psych::KMO(r)

④ psych::cortest.bartlett(r)

Factor analysis methods

- Principal component
 - ▶ the unexplained variance is a result of missing factors and not specific factors
 - ▶ the factors can completely explain the dataset
- Maximum likelihood
 - ▶ data needs to be normal distributed and $\Sigma = AA^T + EE$

$$I(\mu; \Sigma) = -\frac{n}{2} \log |2\pi\Sigma| - \frac{n}{2} \text{tr} (\Sigma^{-1} S_x) - \frac{n}{2} (\bar{x} - \mu) \Sigma^{-1} (\bar{x} - \mu)^T$$

- Unweighted least squares
 - ▶ minimize the difference between the empirical correlation matrix R and the reproduced correlation matrix ($\hat{R} = AA^T$)

$$\min_A \sum_{i=1}^p \sum_{j=1}^p (r_{ij} - \hat{r}_{ij})^2$$

- Principal axis

- ▶ the unexplained variance is a result of missing factors and specific factors
- ▶ besides the factors are other external variables necessary to explain the dataset

- 1a choose as estimate for h_j^2 the squared multiple correlation coefficient of X_j with all other variables
- 1b choose as estimate for h_j^2 the maximum of the absolute correlation coefficient of X_j with all other variables
- 2 estimate \hat{e}_j^2 from \hat{h}_j^2
- 3 compute $R - EE$ and compute eigenvalues and -vectors
- 5 compute the reproduced correlation matrix
- 6 goto 2 and repeat until estimate stabilizes

Note: Sometimes this leads to negative estimates for e_j^2 !

R Listing 14.2: example_efa_extraction.R

```
1 library("psych")
2 scree(bfi[,1:25])
3 # principal component extraction
4 principal(bfi[,1:25], nfactors=5, rotate="none")
5 # principal axis extraction
6 fa(bfi[,1:25], nfactors=5, rotate="none", fm="pa")
7 # maximum likelihood extraction
8 fa(bfi[,1:25], nfactors=5, rotate="none", fm="ml")
9 # unweighted least squares extraction
10 fa(bfi[,1:25], nfactors=5, rotate="none")
```

```
④ psych::principal(r, nfactors=1, residuals=FALSE, rotate="varimax",
                     scores=TRUE, oblique.scores=TRUE,
                     method="regression")
④ psych::fa(r, nfactors=1, residuals=FALSE, rotate="varimax",
            scores=TRUE, oblique.scores=TRUE, method="regression",
            fm="minres")
```

How many factors to choose?

Criteria:

- 90%: minimum of explained total variance
- Kaiser: $\lambda_j > 1$
- Parallel analyse: $\lambda_j > \lambda_j^*$ with λ_j^* bootstrapped from $N(0, I)$
- Stability: $s_j = 0,934 - \frac{1,1}{\sqrt{n}} + 0,12 * \min_{i*} \|a_{i*j}\|$
 - ▶ Choose if $s_j \geq 0,9$
 - ▶ Do not choose if $s_j < 0,8$ (otherwise replication required)
- Communality: large increase of explained variance in at least two or three items
- Interpretation: useful interpretation

- From Guadagnoli and Velicer (1988)

- ▶ If a factor loads on ten or more items then $n \approx 150$ is sufficient
- ▶ If at least four absolute loadings above 0.6 then n does not matter
- ▶ The same holds if more than 12 loadings of a factor are above 0.4
- ▶ In all other cases we should have at least 300 observations and for less than 300 observations we rely on replications and/or theory

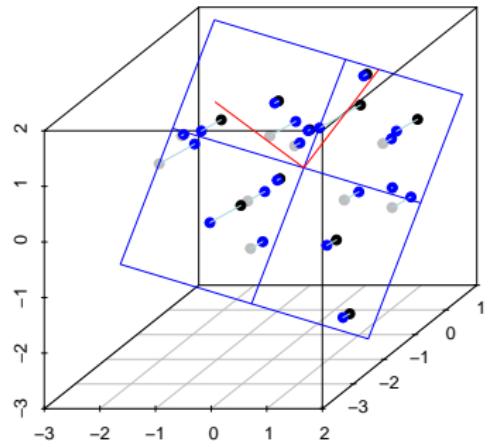
Guadagnoli, Edward and Velicer, Wayne F. (1988). "Relation to sample size to the stability of component patterns.". In: *Psychological Bulletin* 103.2, pp. 265–275. issn: 1939-1455, 0033-2909. doi: 10.1037/0033-2909.103.2.265. url:
<http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.103.2.265> (visited on 08/24/2016).

Rotations

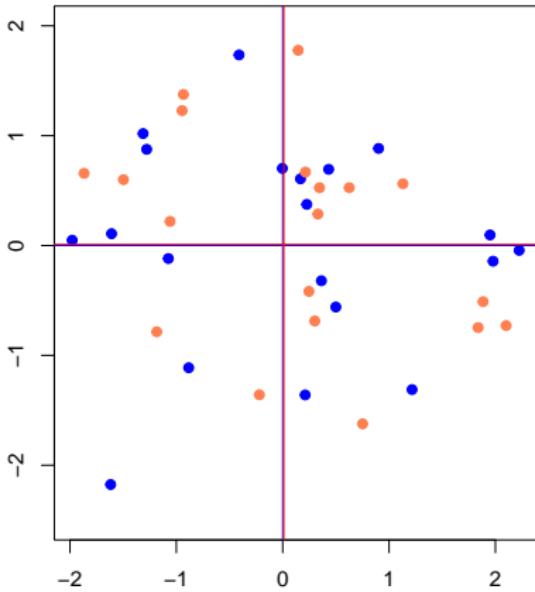
$$\begin{aligned} X &= A_X^T F_X + U E \\ &= \underbrace{A_X^T G^T}_{A_Y^T} \underbrace{G F_X}_{F_Y} + U E \quad (G \text{ orthogonal matrix: } G^T G = I) \\ &= A_Y F_Y + U E \end{aligned}$$

- if we find a solution then any rotation in the space spanned by the factors is also a solution
- basically we are estimating a low-dimensional subspace
- this gives room to think about “interpretable” factors

Principal component plane ($q=2$)



Projected data points

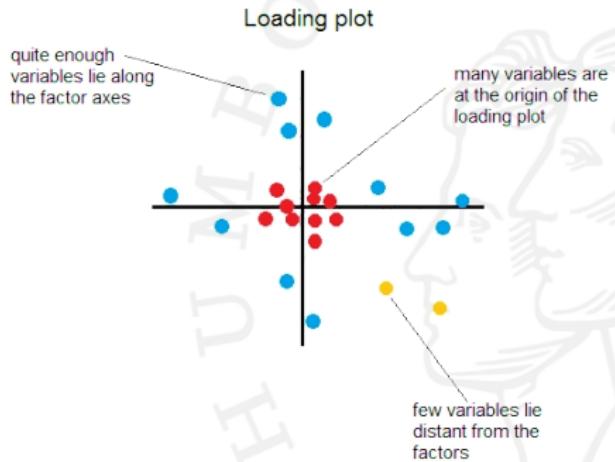


black: before plane, grey: behind plane, blue: projected on plane

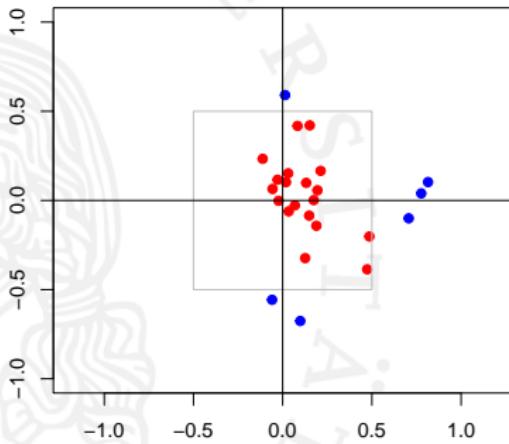
“Simple structure” of loadings (Thurstone 1944)

1. Each row of a loadings matrix is to contain one zero at least, i.e. each feature is described by a maximum of $q - 1$ factors.
2. Each column of a loadings matrix contains q zero-loadings at least, i.e. each factor contributes to the description of a maximum of $p - q$ of the p features.
3. In each pair of columns of the loadings matrix there are some features which have high loadings on one factor, no loadings on the other.
4. If more than four factors were extracted each pair of columns whichever is to contain zero in both columns for a large number of features.
5. For each pair of columns only few features should have high loadings in both columns.

Simple structure: each plot of two factor loadings looks like



First two factor loadings for BFI data



Source: stats.stackexchange.com

Orthogonal rotations

- Kaiser (or Horst) normalization
 - ▶ Idea: weight all variables equally and not by communalities
 - ▶ rescale loadings before rotation such that $\sum_j \tilde{a}_{ij}^2 = 1$
 - ▶ denormalize after rotation
- Varimax (Kaiser 1958):
 - ▶ each factor has small number of large loadings and a large number of small loadings
 - ▶ maximize the variance of the squared loadings in each factor
 - ▶ tends to produce multiple group factors
- Quartimax (Neuhaus und Wrigley 1954)
 - ▶ minimize the number of factors needed to explain each variable
 - ▶ maximize the variance of the squared loadings in each variable
 - ▶ tends to produce a general factor and additional smaller multiple group factors

- Equimax (Saunders 1962)
 - ▶ compromise between Varimax and Quartimax

Thurstone, L. L. (June 1944). "Second-order factors". In: *Psychometrika* 9.2, pp. 71–100. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02288715. url:

<http://link.springer.com/10.1007/BF02288715> (visited on 08/24/2016).

Neuhaus, Jack O. and Wrigley, Charles (Nov. 1954). "THE QUARTIMAX METHOD: AN ANALYTIC APPROACH TO ORTHOGONAL SIMPLE STRUCTURE1". In: *British Journal of Statistical Psychology* 7.2, pp. 81–91. issn: 0950561X. doi: 10.1111/j.2044-8317.1954.tb00147.x. url:

<http://doi.wiley.com/10.1111/j.2044-8317.1954.tb00147.x> (visited on 08/24/2016).

Kaiser, Henry F. (Sept. 1958). "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23.3, pp. 187–200. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02289233. url: <http://link.springer.com/10.1007/BF02289233> (visited on 08/24/2016).

Saunders, D.R. (1962). "Trans-varimax: some properties of the ratiomax and equamax criteria for blind orthogonal rotation". In: *American psychologist* 17. Abstract, pp. 395–396.

Horst, Paul (1965). *Factor analysis of data matrices*. Holt, Rinehart and Winston. isbn: 978-0030502507.

Non-orthogonal rotations

- expected correlation between factors > 0.15
- compare oblique and orthogonal rotations
- if factors have a correlation < 0.3 then you may stay with orthogonal rotation
- Direct oblimin (Jennrich & Simpson 1966)
 - ▶ Quartimin: minimize the sum of squared product loadings between two columns
 - ▶ Kovarimin: minimize the covariance between two columns
 - ▶ balances between Quartimin and Kovarimin
 - ▶ tends to produce varimax-looking factors, but which are oblique
 - ▶ δ controls the extent of correlation between factors (< 0.8)

- Promax (Hendrickson & White 1964)
 - ▶ compute a target solution as a power, usually 2, 4 or 6, of the varimax solution
 - ▶ compute a linear transformation as near as possible to the target solution

Hendrickson, Alan E. and White, Paul Owen (May 1964). "PROMAX: A QUICK METHOD FOR ROTATION TO OBLIQUE SIMPLE STRUCTURE". In: *British Journal of Statistical Psychology* 17.1, pp. 65–70. issn: 0950561X. doi: 10.1111/j.2044-8317.1964.tb00244.x. url: <http://doi.wiley.com/10.1111/j.2044-8317.1964.tb00244.x> (visited on 08/24/2016).

Jennrich, R. I. and Sampson, P. F. (Sept. 1966). "Rotation for simple loadings". In: *Psychometrika* 31.3, pp. 313–323. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02289465. url: <http://link.springer.com/10.1007/BF02289465> (visited on 08/24/2016).



Listing 14.3: example_efa_rotation.R

```
1 library("psych")
2 bfi2 <- na.omit(bfi[,1:25])
3 # ML with Kaiser normalization
4 factanal(bfi2, factors=5)
5 # oblimin rotation without Kaiser normalization
6 fa1 <- fa(bfi2, nfactors=5)
7 fa1
8 # apply Kaiser normalization
9 fa2 <- fa(bfi2, nfactors=5, rotate="none")
10 fa2 <- kaiser(fa2)
11 fa2
12 # compare loading sets (vector cosines)
13 factor.congruence(fa1, fa2)
```

- with Kaiser normalization
 - ⌚ varimax(loadings)
 - ⌚ promax(loadings, m=4)
- without (or with) Kaiser normalization
 - ⌚ GPArotation::Varimax(loadings, normalize=FALSE)
 - ⌚ GPArotation::oblimin(loadings, normalize=FALSE)
 - ⌚ GPArotation::quartimin(loadings, normalize=FALSE)
 - ⌚ GPArotation::::(loadings)
- Kaiser normalization
 - ⌚ psych::kaiser(fa, rotate="oblimin")
- comparing loadings
 - ⌚ psych::factor.congruence(x, y=NULL,digits=2)

Pattern and structure matrix

- pattern matrix: matrix of loadings A
- structure matrix: matrix of correlations between variables and factors
- orthogonal rotation
 - ▶ factors are uncorrelated

$$C = \text{corr}(F_j, F_k) = \delta_{jk} = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases}$$

- ▶ structure matrix is equal to pattern matrix

$$\text{corr}(Z_j, F_k) = \sum_{i=1}^q a_{ij} \underbrace{\text{corr}(F_i, F_k)}_{=\delta_{ik}} = a_{kj}$$

- non-orthogonal rotation
 - ▶ factors are correlated

$$\text{corr}(Z_j, F_k) = \sum_{i=1}^q a_{ij} \underbrace{\text{corr}(F_i, F_k)}_{\neq \delta_{ik}}$$

- ▶ structure matrix is *not equal* to pattern matrix
- ▶ for interpretation of factors analyze structure matrix

	Unrotated		Pattern		Structure	
	F1	F2	F1	F2	F1	F2
WIDTH	-0,01	0,92	-0,18	0,93	-0,04	0,90
LEFT	0,80	0,39	0,73	0,42	0,79	0,53
RIGHT	0,84	0,29	0,78	0,32	0,83	0,43
UPPER	0,70	-0,30	0,75	-0,28	0,71	-0,17
LOWER	0,63	-0,10	0,65	-0,08	0,63	0,01
DIAGONAL	-0,85	0,31	-0,90	0,29	-0,86	0,15

Extraction: Principal component, Rotation: Promax

A case study: the WiWi evaluation forms

- Structure of the form
 - ▶ General information: major, gender, global overall ratings of the course, ...
 - ▶ Lecturer's characteristic: explain ability, transparency quality, ...
 - ▶ Lecture's Concept: aspects covered, topic structure, choice and availability of lecture notes ...
 - ▶ Course characteristics: speed, mathematical level ...
 - ▶ Self assessment of students: interest degree, attention span ...
 - ▶ Course atmosphere: stress level, disciplined degree ...
- Most items are using a 5 point scale ranging from 1 (very good or too high) to 5 (very bad or too low)

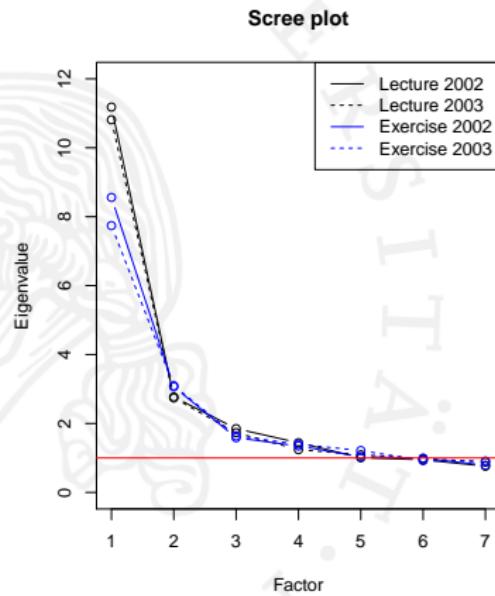
Data

- Covers 2 summer semesters: 2002 and 2003
- Comprises 164 individual undergraduate and graduate courses taught by more than 35 instructors
- Three kinds of evaluation forms used : lecture class, exercise class, seminar class
- Data for lecture class and exercise class has been used

Zhou, Yilan (July 6, 2004). "Basic Statistical Analysis and Modelling of Evaluation Data for Teaching". Master thesis. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät. url:
<http://edoc.hu-berlin.de/master/zhou-yilan-2004-07-06/PDF/zhou.pdf> (visited on 11/15/2016).

Explained total variance with Kaiser criterium:

Factors	Lect 03	Lect 02
1	37%	37%
2	47%	47%
3	53%	54%
4	58%	62%
5	62%	65%
6	66%	
Factors	Exer 03	Exer 02
1	34%	36%
2	46%	49%
3	54%	56%
4	59%	61%
5	64%	66%
6		



	Dozent/Übungsleitung								Konzept der Veranstaltung						Anford.	Selbsteinsch.			Atmosphäre											
	B1	B2	B3	B4	B5	B6	B7	B8	C1	C2	C3	C4	C5	C6	C7	C8	C9	D1	D2	D3	E1	E2	E3	E4	E5	F1	F2	F3	F4	Erklärte Varianz
VL 2003	,8	,8	,6	,8	,7	,7	,8	,2	,7	,8	,7	,7	,6	,6	,7	,3	-,1	-2	-3	,6	,6	,8	-,1	-,1	,4	,8	,5	,8	0,38	
VL 2002	,8	,8	,7	,8	,6	,7	,8	,2	,7	,7	,6	,7	,6	,6	,6	,6	,4	-3	-4	-5	,5	,6	,7	,0	-2	,5	,8	,6	,8	0,38
UE 2003	,8	,8	,6	,8	,6	,7	,8	,2	,7	,8	,6	,6	,6	,5				-2	-3	-3	,5	,7	,1	-2	,4	,7	,5	,8	0,35	
UE 2002	,8	,8	,6	,8	,6	,6	,7	,2	,7	,7	,6	,7	,7	,6				-4	-5	-6	,4	,7	,0	-4	,5	,7	,5	,8	0,39	

- 1st factor: lecturer's general abilities of to teach
highly correlated lowly correlated

variable	loading	variable	loading
explain ability	0,83	preparation level	0,10
content clarity	0,81	lecture speed	0,13
aspects covered	0,75	time allowed after class	-0,22
topic structure	0,75	interest degree	0,23

	Dozent/Übungsleitung								Konzept der Veranstaltung									Anford.	Selbsteinsch.					Atmosphäre																	
	Erläuterungsvermögen	Verständlichkeit	Qualität Folien	Didakt. Kompetenz	Eigenständiges Denken	Bereit. Zwischenfragen	Qualität Zwischenfragen	Schwerpunkte									Gliederung	Interdisziplinarität	Veranschaulichung	Skripte	Verfügbarkeit	Internet	Aktualität	Abstimmung	Geschwindigkeit	Formalisierung	Schwierigkeitsgrad	Interesse	E1	E2	E3	E4	E5	Lernzuwachs	Überfordert	Vorbereitung	Stressfrei	Interessant	Diszipliniert	Motivierend	Erklärte Varianz
VL 2003	B1	B2	B3	B4	B5	B6	B7	B8	C1	C2	C3	C4	C5	C6	C7	C8	C9	D1	D2	D3	E1	E2	E3	E4	E5	F1	F2	F3	F4												
Lehrbefähigung	,7	,7	,6	,6	,4	,3	,4	,1	7	7	,5	,4	,4	,2	,2	,4	,0	,0	,0	-1	,2	,2	,4	-1	,0	,2	,4	,3	,5	0,17											
Lehrmaterialien	,1	,2	,3	,1	,1	,1	,1	,2	,3	,3	,2	,3	,6	,8	,7	,4	,1	,0	,0	,0	,1	,1	,2	-1	,0	,1	,0	,1	,0	0,08											
Leistungsanforderung	,1	,2	,0	,0	,0	,0	,0	,0	,0	,0	,0	,0	,1	,0	,0	,0	,0	,1	,7	,7	,9	,1	-1	,0	,3	,6	,3	,1	,1	,1	0,09										
Zwischenfragen	,4	,3	,2	,4	,3	8	,8	,2	,1	,1	,2	,3	,1	,2	,2	,3	,3	,0	,0	,0	,1	,2	,2	,1	,0	,2	,3	,3	,3	0,09											
Selbsteinschätzung	,3	,3	,1	,3	,4	,2	,2	,0	,3	,2	,3	,3	,2	,1	,1	,4	,1	,0	-1	-1	,7	,7	,7	-2	,0	,2	,7	,3	,6	0,13											
VL 2002																																									
Lehrbefähigung	,8	,7	,5	,7	,4	,2	,4	,1	6	,7	,5	,4	,3	,1	,2	,3	,1	,0	-1	-1	,1	,2	,4	,1	,0	,2	,5	,3	,5	0,16											
Lehrmaterialien	,1	,2	,4	,1	,1	,1	,2	,1	,3	,4	,3	,4	,7	,8	,7	,5	,1	,0	-1	-1	,1	,1	,2	-1	,0	,1	,1	,2	,2	0,10											
Leistungsanforderung	,2	,3	,1	,1	,0	,1	,1	,0	,0	,1	,1	,2	,1	,1	,0	,1	,3	-8	-8	-9	,1	-1	,1	,3	,5	,4	,1	,1	,2	0,10											
Zwischenfragen	,3	,2	,1	,4	,4	8	,8	,3	,1	,1	,2	,3	,1	,2	,1	,3	,3	,1	,0	,0	,1	,2	,2	,1	,0	,2	,2	,2	,3	0,08											
Selbsteinschätzung	,3	,2	,2	,3	,4	,2	,2	,0	,2	,2	,3	,3	,2	,1	,1	,3	,1	,0	-1	-1	,7	,7	,7	-3	,0	,2	,7	,4	,6	0,12											
UE 2003																																									
Lehrbefähigung	,8	,7	,6	,7	,5	,7	,8	,2	,4	,4	,3	,2	,1	,2	,	,0	,0	,0	,0	,0	,1	,2	,0	,0	,2	,2	,2	,3	,5	0,17											
Leistungsanforderung	,0	,2	,0	-1	-1	-1	-1	,1	,1	,1	,1	,0	,0			-8	-7	-9	-3	-1	,3	-7	,3	,0	,1	,1	,1	0,11													
Lehrmaterialien	,1	,1	,2	,1	,1	,2	,2	,2	,3	,3	,2	,6	,9	,7		,0	,0	,0	,0	,1	,1	,0	,0	,1	,1	,2	,1	0,09													
Konzept	,3	,4	,3	,3	,2	,0	,0	,0	7	,7	,4	,3	,1	,1	,	,0	-1	-1	,4	,5	-2	,0	-1	,3	,0	,3	0,09														
Selbsteinschätzung	,2	,2	,1	,2	,4	,3	,3	,1	,2	,2	,3	,1	,1	,1	,	,0	-1	-1	,5	,6	-1	,0	,5	,7	,6	,7	0,11														
UE 2002																																									
Lehrbefähigung	,8	,6	,5	,7	,5	,7	,8	,3	,3	,4	,4	,2	,2	,2	,2	,	-1	,0	,0	,2	,3	,0	,0	,2	,2	,3	,3	0,17													
Leistungsanforderung	,1	,2	,1	,0	-1	,0	,0	,0	,1	,1	,2	,2	,1	,0		-8	-7	-9	-3	,0	,3	-7	,3	,1	,0	,2	0,12														
Lehrmaterialien	,0	,1	,3	,1	,1	,2	,2	,2	,2	,3	,2	,7	,9	,7		,0	-1	-1	,0	,1	,0	-1	,2	,1	,2	,2	0,10														
Selbsteinschätzung	,3	,3	,2	,3	,4	,3	,2	,0	,3	,2	,3	,2	,2	,2	,2	,	-1	-1	-1	,6	,6	-2	,0	,5	,8	,5	,8	0,14													
Konzept	,3	,4	,3	,3	,2	,0	,0	,1	7	,7	,3	,2	,1	,1	,	-1	-1	-1	,2	,3	-1	,0	,0	,2	,0	,1	0,07														

Scores

- Scales $S_j = \frac{1}{p_j} \sum_{i=1}^{p_j} x_{i(j)}$
 - ▶ requires $\text{Var}(X_{i(j)})$ and/or the range of $X_{i(j)}$ approx. the same range
 - ▶ easy to calculate and interpret
 - ▶ acceptable for most EFA situations
- Component scores $C_j = \sum_{i=1}^p a_{ij}x_i$
 - ▶ uses the loadings as coefficients
 - ▶ turns out to be monotone (non-linear) transformation of the factor scores in certain circumstances
 - ▶ differences in loadings may depend on extraction and rotation method
- Factor scores
 - ▶ usually computed with a regression model
- Scales and component scores thought to be more stable across samples than factor scores

- (Unweighted) Regression scores (Thurstone, 1935)

$$\sum_{i=1}^p \varepsilon_{ij}^2 = \sum_{i=1}^p (z_{ij} - \sum_{j=1}^q a_{ij} f_i)^2 \rightarrow \min.$$

- (Weighted) Regression scores (Barlett, 1937)

$$\sum_{i=1}^p \frac{\varepsilon_{ij}^2}{s_{U_i}^2} = \sum_{i=1}^p \frac{(z_{ij} - \sum_{j=1}^q a_{ij} f_i)^2}{s_{U_i}^2} \rightarrow \min.$$

- Modification of Bartlett method to ensure orthogonality of factors (Anderson and Rubin, 1956)

- Regression factor scores
 - ▶ have mean zero, variance equals squared multiple correlation between variables and factor
 - + maximizes correlation between true and estimated scores
 - are not unbiased estimates of true scores
 - might be correlated even when factors are orthogonal
- Bartlett scores
 - ▶ have mean and variance as under regression scores
 - + maximize the correlation between true and estimated scores
 - + are unbiased estimates of true scores
 - might be correlated even when factors are orthogonal
- Anderson-Rubin scores
 - ▶ mean 0 and variance 1
 - + reach a reasonably high correlation between true and estimated scores
 - are not unbiased estimates of true scores
 - might be correlated even when factors are orthogonal



Listing 14.4: example_efa_scores.R

```
1 library("psych")
2 bfi2 <- na.omit(bfi[,1:25])
3 # ML with Kaiser normalization
4 fa1 <- factanal(bfi2, factors=5, scores="regression")
5 head(fa1$scores)
6 # oblimin rotation without Kaiser normalization
7 fa2 <- fa(bfi2, nfactors=5)
8 head(fa2$scores)
9 # compare scores
10 cor(fa1$scores, fa2$scores)
```

¶ `factanal(x, factors, scores=c("none", "regression", "Bartlett"))`

¶ `psych::factor.scores(x, f, method=c("Thurstone", "tenBerge",
 "Anderson", "Bartlett", "Harman",
 "components"))`

Reliability

November 3, 2022

Reliability • Cronbach's α • Other reliability coefficients • Tukey's test of additivity

Reliability

- if a set of items for a latent variable/construct found then we have to ask for
 - ▶ Reliability: can we replicate the measurements?
 - ▶ Validity: does the latent variable measure what we want to measure?
 - ▶ Objectivity: are the measurements independent from the surrounding framework?

It is the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores.

Scores that are highly reliable are precise, reproducible, and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test takers, essentially the same results would be obtained. Various kinds of reliability coefficients, with values ranging between 0 (much error) and 1 (no error), are usually used to indicate the amount of error in the scores.

National Council on Measurement in Education (Wikipedia)

Cronbach's α

- Cronbachs α

$$\alpha = \frac{p}{p-1} \left(1 - \frac{\sum Var(X_i)}{Var(Y)} \right)$$

$$Y = \sum X_i \Rightarrow Var(Y) = \sum Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$$

- ▶ measure how much the X_i belong to one underlying construct (internal consistency)
- ▶ $\alpha = 0 \Rightarrow Var(Y) = \sum Var(X_i)$ and $\sum_{i \neq j} Cov(X_i, X_j) = 0$
- Standardized Cronbachs α (X_i standardized)

$$\alpha_z = \frac{p\bar{r}}{1 + (p-1)\bar{r}}$$

- ▶ \bar{r} average correlation between X_i 's

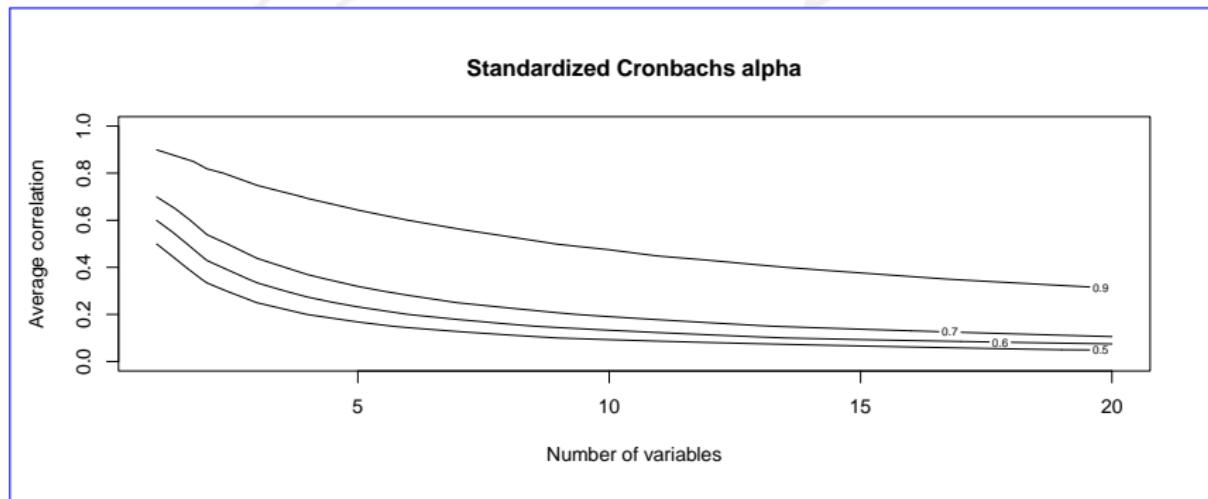
- Cronbachs α usually underestimates the true reliability
- rule of thumb: $\alpha \geq 0.7$

<0.5	0.5-0.6	0.6-0.7	0.7-0.9	>0.9
unacceptable	poor	acceptable	good	excellent

- check deleted Cronbachs α
 - ▶ does α increase when you delete one variable from the sum score?

Cronbach, Lee J. (Sept. 1951). "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16.3, pp. 297–334. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02310555. url: <http://link.springer.com/10.1007/BF02310555> (visited on 08/25/2016).

- trade-off between rule-of-thumb and number of variables



- α is not a measure for unidimensionality

$$R = \begin{pmatrix} 1 & 0.95 & 0.1 & 0.1 \\ 0.95 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.95 \\ 0.1 & 0.1 & 0.95 & 1 \end{pmatrix} \Rightarrow \alpha_z \approx 0.71$$

R Listing 15.1: example_reliability.R

```
1 library("psych")
2 bfi2 <- na.omit(bfi[,1:25])
3 # extract first factor
4 fa <- fa(bfi2)
5 vars <- abs(fa$loadings)>0.5
6 # Cronbachs alpha (items not reversed)
7 alpha(cor(bfi2[,vars]))
8 # Cronbachs alpha (items reversed)
9 alpha(cor(bfi2[,vars]), check.keys=T)
10 # Cronbachs alpha and sum scores
11 keys <- vars*sign(fa$loadings)
12 si   <- scoreItems(keys, bfi2)
```

R psych::alpha(x, keys=NULL, check.keys=FALSE)

R psych::scoreItems(keys, items, min=NULL, max=NULL)

Other reliability coefficients

Split-half model

- split the variables in two sets with $p_A + p_B = p$ variables
- compute two sum scores Y_A and Y_B
- analyze Bravais-Pearson correlation r_{AB} of Y_A and Y_B
- Spearman-Brown coefficient (reliability of two “variables”)

$$SB = \begin{cases} \frac{2r_{AB}}{1+r_{AB}} & \text{if } p_A = p_B \\ \frac{-r_{AB}^2 + \sqrt{r_{AB}^4 + 4r_{AB}^2(1-r_{AB}^2)\frac{p_A p_B}{p_A + p_B}}}{2(1-r_{AB}^2)\frac{p_A p_B}{p_A + p_B}} & \text{if } p_A \neq p_B \end{cases}$$

- Spearman-Brown assumes equal reliabilities/variances for Y_A and Y_B
- Guttman coefficient (reliability of two “variables”)

$$\lambda_4 = 2 \left(1 - \frac{S_{Y_A}^2 + S_{Y_B}^2}{S_Y^2} \right)$$

- Cronbachs α is the mean of all possible split-half coefficients

Brown, William (Oct. 1910). "SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES¹". In: *British Journal of Psychology*, 1904-1920 3.3, pp. 296–322.

issn: 09505652. doi: 10.1111/j.2044-8295.1910.tb00207.x. url:

<http://doi.wiley.com/10.1111/j.2044-8295.1910.tb00207.x> (visited on 08/25/2016).

Spearman, C. (Oct. 1910). "CORRELATION CALCULATED FROM FAULTY DATA". In:

British Journal of Psychology, 1904-1920 3.3, pp. 271–295. issn: 09505652. doi:

10.1111/j.2044-8295.1910.tb00206.x. url:

<http://doi.wiley.com/10.1111/j.2044-8295.1910.tb00206.x> (visited on 08/25/2016).

Guttman, Louis (Dec. 1945). "A basis for analyzing test-retest reliability". In: *Psychometrika*

10.4, pp. 255–282. issn: 0033-3123, 1860-0980. doi: 10.1007/BF02288892. url:

<http://link.springer.com/10.1007/BF02288892> (visited on 08/25/2016).

⌚ Listing 15.2: example_splithalf.R

```
1 library("psych")
2 bfi2 <- na.omit(bfi[,1:25])
3 # various coefficients
4 splitHalf(bfi2)
```

⌚ psych::splitHalf(r, keys=NULL, check.keys=TRUE)

Tukey's test of additivity

Assumption(s) $E_{ij} \sim N(0; \sigma)$

Hypotheses: $H_0 : \lambda = 0$ vs. $H_1 : \lambda \neq 0$

$$\text{with } X_{ij} = \mu + f_i + \mu_j + \lambda f_i \mu_j + E_{ij}$$

Test statistics: $V = SS_{AB}/SS_E$

$$SS_E = SS_T - SS_A - SS_B - SS_{AB}$$

$$SS_T = \sum_{ij} (X_{ij} - \bar{X}_{\bullet\bullet})^2$$

$$SS_A = p \sum_i (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2$$

$$SS_B = n \sum_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2$$

$$SS_{AB} = \frac{\sum_{ij} X_{ij}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})}{\sum_{ij} (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})}$$

Reject H_0 : $v > F_{1;np-p-n;1-\alpha}$

Note: Test is conservative (if normality assumption is violated)

Tukey, John W. (Sept. 1949). "One Degree of Freedom for Non-Additivity". In: *Biometrics* 5.3, p. 232. issn: 0006341X. doi: 10.2307/3001938. url: <http://www.jstor.org/stable/3001938?origin=crossref> (visited on 08/25/2016).



Listing 15.3: example_additivity_test.R

```
1 #install.packages("devtools")
2 #library("devtools")
3 #install_github("simecek/additivityTests")
4 library("psych")
5 # extract first factor
6 bfi2 <- na.omit(bfi)
7 fa <- fa(bfi2)
8 vars <- (abs(fa$loadings)>0.5)
9 # create corrected items matrix
10 keys <- keys[vars*sign(fa$loadings)]
11 items <- reverse.code(keys, bfi2[,vars])
12 # additivity test
13 library("additivityTests")
14 tukey.test(items)
```

⌚ psych::reverse.code(keys, items, mini=NULL, maxi=NULL)

⌚ additivityTests::tukey.test(data, alpha=0.05)

⚠ <https://github.com/rakosnicek/additivityTests>

Cluster analysis

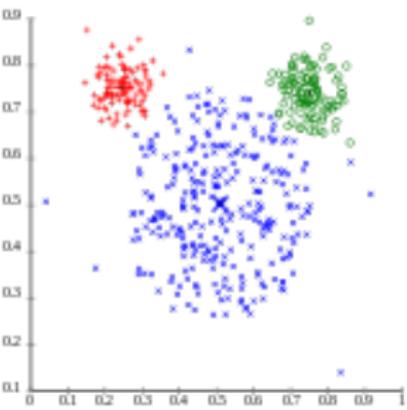
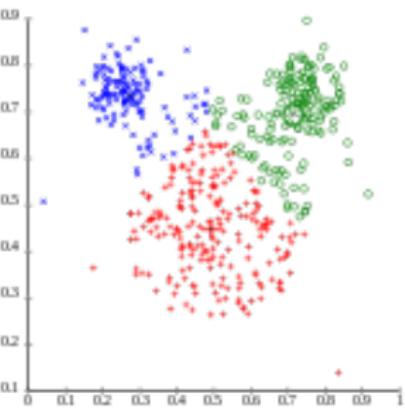
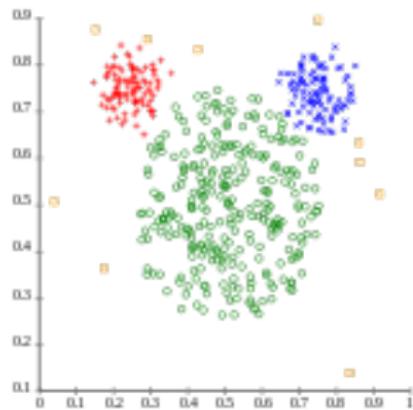
November 3, 2022

- Cluster analysis
- *k*-means
- Hierarchical clustering
- Diana
- Distance
- Similarity
- Agglomeration
- Dendrogram
- Dendrogram II
- EM-Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Mixed clustering
- Measures to compare two classifications
- Two-Step-Clustering in SPSS
- Number of clusters
- Silhouette
- Cluster visualization

Cluster analysis

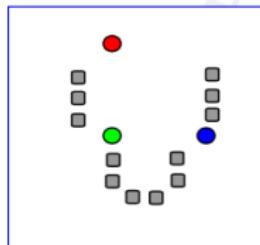
- Goals
 - ▶ reduce the number of observation
 - ▶ find a representative (observation)
 - ▶ find similar observations
- observations with similar “properties” belong together
 - ▶ similar “properties” translates to “near in a distance”
 - ▶ how to measure distances between observations?
- Methods
 - ▶ hard clustering - each observation belongs only to one cluster
 - ▶ soft clustering - each observation to some degree a cluster
 - ▶ R: cran.r-project.org/web/views/Cluster.html

Clustering-Ergebnisse auf dem "Maus" Datensatz: Originaldaten k-Means Clustering EM Clustering

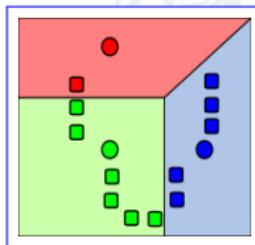


k-means

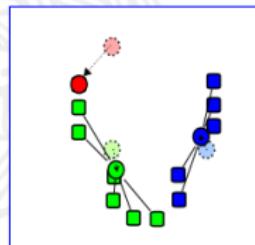
- Assumption: the number of clusters k must be known in advance
- Lloyd algorithm:



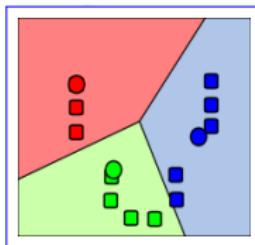
distribute
randomly
 k cluster
centers



assign each
observation
to the “near-
est” cluster
center



recompute
cluster cen-
ters



stop if clus-
ters do not
change or
max iteration
reached

 Listing 16.1: example_kmeans1.R

```
1 library("mclust") # for Swiss banknote data
2 data(banknote)
3 x <- banknote[,2:7]
4 z <- scale(x)
5 #
6 kx <- kmeans(x, c=2)
7 plot(x$Bottom, x$Diagonal, col=kx$cluster,
8       main="Unstandardized", pch=19)
9 kx$centers
10 scale(kx$centers, center=colMeans(x), scale=apply(x, 2, sd))
11 #
12 kz <- kmeans(z, c=2)
13 plot(x$Bottom, x$Diagonal,
14       col=kz$cluster, main="Standardized", pch=19)
```

 `kmeans(x, centers, iter.max=10, nstart=1,`
`algorithm=c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))`

- k -median
 - ▶ use L_1 distance instead of euclidean distance
 - ▶ use median instead of mean
- k -means++
 1. choose randomly an observation as first cluster center
 2. compute the minimal (squared) distance d_i to all cluster centers
 3. choose as next cluster center randomly an observation such that the choosing probability is proportional to d_i
 4. repeat 2. and 3. until all k cluster centers determined
 5. apply k -means clustering

- **k -medoid: Partitioning Around Medoids**
 1. choose k of n observations as cluster centers (medoids)
 2. assign each observation to the closest medoid
 - 3a. swap each medoid and each non-medoid
 - 3b. compute sum of distances or dissimilarities for each swap
 4. choose swap with smallest sum
 5. repeat 2.-4. until clusters do not change
- **Fuzzy c -means (soft) clustering**
 - ▶ each observation belongs to degree (c_1, \dots, c_k) to each cluster
 - ▶ use a weighted euclidean distance to determine new cluster centers

 Listing 16.2: example_kmedian.R

```

1 zfaithful <- apply(faithful, 2, scale)
2 # k-median
3 library("flexclust")
4 cl1 <- kcca(zfaithful, 2, family=kccaFamily('kmedians'))
5 plot(zfaithful, col=cl1@second)
6 cl1@centers

```

 flexclust::kcca(x, k, family=kccaFamily("kmedians"), weights=NULL)

 Listing 16.3: example_kmedoid.R

```

1 zfaithful <- apply(faithful, 2, scale)
2 # k-medoid
3 library("cluster")
4 cl2 <- pam(zfaithful, 2)
5 plot(zfaithful, col=cl2$clustering)
6 cl2$medoids

```

 cluster::pam(x, k, diss=inherits(x, "dist"), metric="euclidean",
 stand=FALSE)

 Listing 16.4: example_kmeans3.R

```

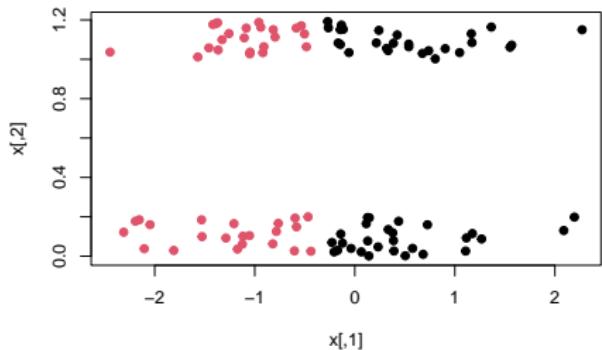
1 zfaithful <- apply(faithful, 2, scale)
2 mcolor      <- colorRamp(c("red", "blue"))
3 # fuzzy c-means 1
4 library("e1071")
5 cl1 <- cmeans(zfaithful, 2)
6 col <- rgb(mcolor(cl1$membership[,1]), max=255)
7 plot(zfaithful, pch=19, col=col)
8 # fuzzy c-means 2
9 library("cluster")
10 cl2 <- fanny(zfaithful, 2)
11 col <- rgb(mcolor(cl2$membership[,1]), max=255)
12 plot(zfaithful, pch=19, col=col)
13 # compare membership values
14 plot(cl1$membership[,1], cl2$membership[,2])

```

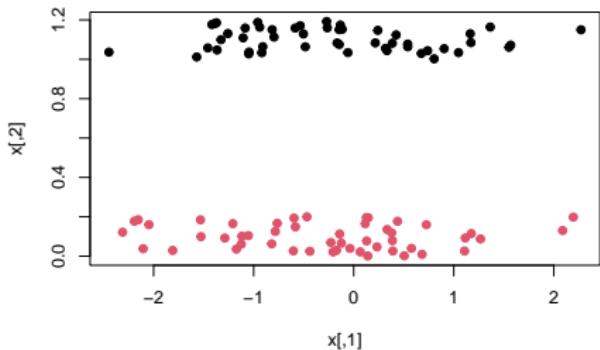
☞ e1071::cmeans(x, centers, iter.max=100, dist="euclidean",
method="cmeans", m=2, weights=1)

☞ cluster::fanny(x, k, dissinherits(x, "dist"), metric="euclidean",
stand=FALSE)

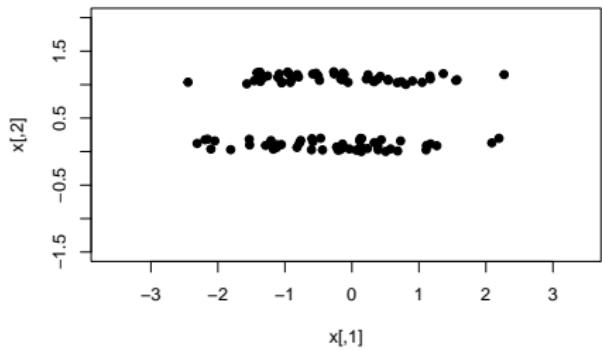
k-means with k=2



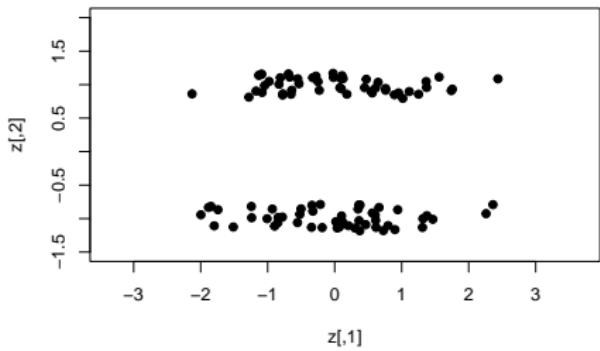
k-means with k=2



Unstandardized data



Standardized data



Hierarchical clustering

- Divisive
 - ▶ start with one cluster which contains all observations
 - ▶ divide one cluster into two subclusters
 - ▶ repeat until each clusters contains only one observation
 - ▶ in practice not often used because of too much division possibilities
- Agglomerative
 - ▶ start with n clusters which contain only one observation
 - ▶ merge two (nearest) cluster into one cluster
 - ▶ repeat until all observations in one cluster
- Questions
 1. how to measure distance?
 2. how to merge clusters?
 3. how to determine the number of clusters?

Diana

Kaufmann und Rousseeuw (1990) describe the following divisive algorithm:

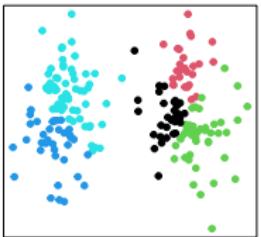
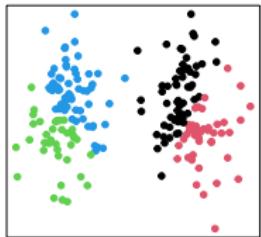
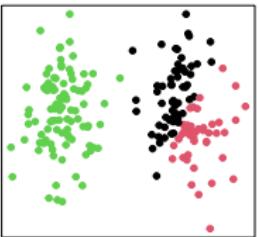
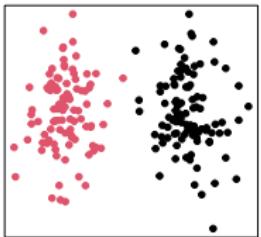
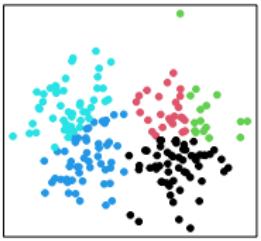
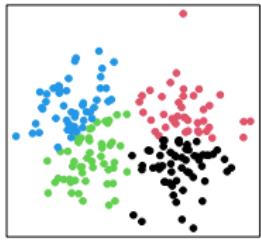
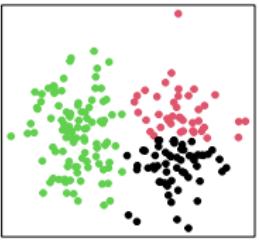
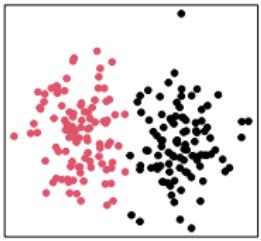
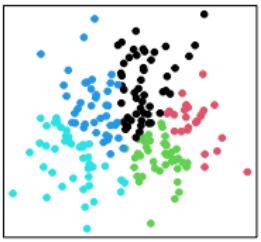
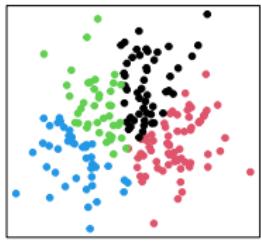
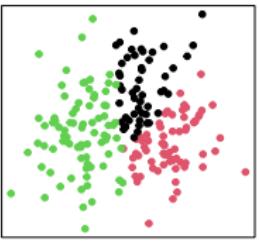
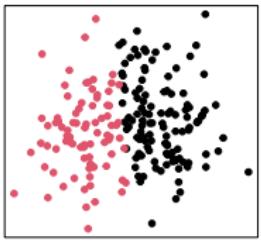
1. Start with a one cluster which contains all observations
2. Compute for each cluster C a diameter: $d_C = \max_{i,j \in C} d(i,j)$
3. Decompose the cluster with the largest diameter in two subclusters
4. Find in the cluster the observation for which holds

$$k = \arg \max_i \frac{1}{n_C} \sum_j d(i,j)$$

5. Make the observation k part of the "splinter group" S
6. Assign observation I to S if

$$\min_{i \in S} d(I, i) < \min_{j \in C \setminus \{S \cup I\}} d(I, j)$$

7. Repeat step 2.-6. until each cluster contains only one observation.





Listing 16.5: example_diana.R

```
1 library("cluster")
2 library("mlbench") # for Boston Housing data
3 data(BostonHousing2)
4 # prepare data
5 x <- BostonHousing2[,-c(1:5,10,15)]
6 x <- scale(x)
7 #
8 cl <- diana(x)
9 hcl <- cutree(as.hclust(cl), k = 2)
10 par(mfrow=c(1,1))
11 plot(BostonHousing2$lon, BostonHousing2$lat, col=hcl,
12 pch=19, cex=0.5)
```



```
cluster::diana(x, diss=inherits(x, "dist"), metric="euclidean",
stand=FALSE)
```

Distance

- Distances (metric variables)
 1. $d(i, j) = d(j, i)$
 2. $d(i, j) \geq 0$ and $d(i, j) = 0 \Leftrightarrow x_i = x_j$
 3. sometimes: $d(i, k) \leq d(i, j) + d(j, k)$ (triangle inequality)
- L_q distances, Minkowski distance

$$d(i, j) = \sqrt[q]{\sum_{k=1}^p |x_{ik} - x_{jk}|^q}$$

- ▶ $q = 1$: Manhattan or City-block distance
- ▶ $q = 2$: euclidean distance

- Pearson

$$d(i, j) = \sqrt{\sum_{k=1}^p \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2}$$

- Gower (metric)

$$d(i,j) = \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{\max_I x_{Ik} - \min_I x_{Ik}}$$

- Mahalanobis

$$d(i,j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

- ordinal or categorical variables
 - ▶ transform to a set of binary variables
 - ▶ weight the result by the number of binary variables
- Gower distance (binary or categorical)

$$d(i, j) = \sum_{k=1}^p I(x_{ik} = x_{jk}) = n_{00} + n_{11}$$

- ▶ since Gower measures variablewise it can be used on mixed data
- ▶ in Gower weights can be assigned to variable
- ▶ Gower can handle missing values

R Listing 16.6: example_distance.R

```
1 library("cluster")
2 zfaithful <- apply(faithful, 2, scale)
3 # euclidean distance
4 heatmap(as.matrix(dist(zfaithful)))
5 heatmap(as.matrix(daisy(zfaithful)))
6 # manhattan
7 heatmap(as.matrix(dist(zfaithful, 'manhattan')))
8 heatmap(as.matrix(daisy(zfaithful, 'manhattan')))
9 # gower
10 heatmap(as.matrix(daisy(zfaithful, 'gower')))
```

- R `dist(x, method="euclidean", diag=FALSE, upper=FALSE, p=2)`
- R `proxy::dist(x, method=NULL, diag=FALSE, upper=FALSE, by_rows=TRUE)`
- R `mahalanobis(x, center, cov, inverted=FALSE)`
- R `cluster::daisy(x, metric=c("euclidean", "manhattan", "gower"),
stand=FALSE, type=list(), weights=rep.int(1, p))`

Similarity

- Similarity (binary variables)
 1. $s(i, j) = s(j, i)$
 2. $s(i, j) \leq s(i, i)$
 3. sometimes: $s(i, j) \geq 0$ and $s(i, i) = 1$
- Distances from similarities

$$d(i, j) = \sqrt{s(i, i) + s(j, j) - 2s(i, j)}$$

- Count pairs with
 - ▶ n_{00} = number of pairs with $x_{i,k} = 0$ and $x_{j,k} = 0$ ($k = 1, \dots, p$)
 - ▶ n_{01} = number of pairs with $x_{i,k} = 0$ and $x_{j,k} = 1$ ($k = 1, \dots, p$)
 - ▶ n_{10} = number of pairs with $x_{i,k} = 1$ and $x_{j,k} = 0$ ($k = 1, \dots, p$)
 - ▶ n_{11} = number of pairs with $x_{i,k} = 1$ and $x_{j,k} = 1$ ($k = 1, \dots, p$)
 - ▶ it holds $n_{00} + n_{01} + n_{10} + n_{11} = p$



Listing 16.7: example_dist2simil.R

```
1 library("proxy")
2 d <- as.matrix(dist(faithful))
3 heatmap(d)
4 # distances to similarities
5 s <- pr_dist2simil(d)
6 heatmap(s)
7 # distance and similarity measures in proxy
8 summary(pr_DB)
9 pr_DB$get_entry('Jaccard')
```

② proxy::pr_dist2simil(x)
② proxy::pr_simil2dist(x)

Braun	$\frac{n_{11}}{\max(n_{11}+n_{01}, n_{11}+n_{10})}$	Phi	$\frac{n_{11}n_{00}-n_{10}n_{01}}{\sqrt{(n_{11}+n_{01})(n_{11}+n_{10})(n_{00}+n_{01})(n_{00}+n_{10})}}$
Dice	$\frac{2n_{11}}{n_{01}+n_{10}+2n_{11}}$	Russel Rao	$\frac{n_{11}}{p}$
Hamann	$\frac{(n_{00}+n_{11})-(n_{01}+n_{10})}{p}$	Simple Matching	$\frac{n_{00}+n_{11}}{p}$
Jaccard	$\frac{n_{11}}{n_{01}+n_{10}+n_{11}}$	Simpson	$\frac{n_{11}}{\min(n_{11}+n_{01}, n_{11}+n_{10})}$
Kappa	$\frac{1}{1+\frac{p(n_{01}+n_{10})}{2(n_{00}n_{11}-n_{01}n_{10})}}$	Sneath	$\frac{n_{11}}{n_{11}+2n_{01}+2n_{10}}$
Kulczynski	$\frac{n_{11}}{n_{01}+n_{10}}$	Tanimoto/Roger	$\frac{n_{00}+n_{11}}{n_{00}+2(n_{01}+n_{10})+n_{11}}$
Ochiai	$\frac{n_{11}}{\sqrt{(n_{11}+n_{01})(n_{11}+n_{10})}}$	Yule	$\frac{n_{00}n_{11}-n_{01}n_{10}}{n_{00}n_{11}+n_{01}n_{10}}$

- choice of measure depends on purpose
- The combinations 00 and 11 are equal important, e.g. for sex
 - ▶ prefer Simple Matching, Hamann or Tanimoto
- The combinations 00 and 11 are not equal important, e.g. for disease appeared
 - ▶ prefer Dice, Jaccard, Kulczynski, Ochiai, Braun, Simpson or Sneath
 - ▶ It holds
$$s_{Sneath}(i,j) \leq s_{Jaccard}(i,j) \leq s_{Dice}(i,j)$$
$$s_{Braun}(i,j) \leq s_{Dice}(i,j) \leq s_{Ochiai}(i,j) \leq s_{Kulczynski}(i,j) \leq s_{Simpson}(i,j)$$
- Kappa, Phi and Yule can be used always

⚠ Some similarity measures are not symmetric!



Listing 16.8: example_simil.R

```
1 # create binary data matrix
2 wtab <- as.data.frame(Titanic)
3 bdat <- cbind(wtab$Sex=="Male",
4
5
6 # compute similarities
7 library("proxy")
8 d <- as.matrix(simil(bdat, method='Jaccard'))
9 heatmap(d)
```

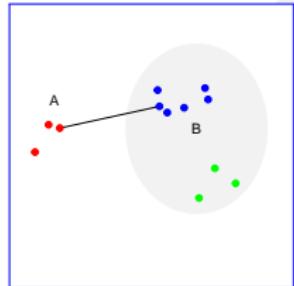
R proxy::simil(x, method=NULL, diag=FALSE, upper=FALSE, by_rows=TRUE)

Agglomeration

1. compute a distance matrix $D^{(0)}$ for n clusters ($n \times n$ matrix)
 - ▶ distance and similarity measures define a distance between two points
 - ▶ linkage methods define how distances (and similarities) are aggregated
2. merge the two nearest cluster
3. update the distance matrix $D^{(k)}$ ($(n - k) \times (n - k)$ matrix) with Lance -Williams formula
$$d(i \cup j, k) = \alpha_i d(i, k) + \alpha_j d(j, k) + \beta d(i, j) + \gamma |d(i, k) - d(j, k)|$$
4. repeat 2. and 3. until only one cluster is left

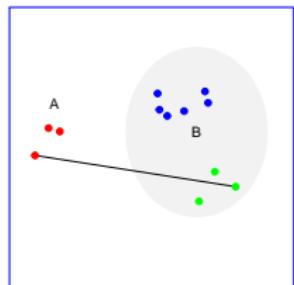
Lance, G. N. and Williams, W. T. (Feb. 1, 1967). "A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems". In: *The Computer Journal* 9.4, pp. 373–380. issn: 0010-4620, 1460-2067. doi: 10.1093/comjnl/9.4.373. url: <http://comjnl.oxfordjournals.org/cgi/doi/10.1093/comjnl/9.4.373> (visited on 12/06/2016).

Single linkage



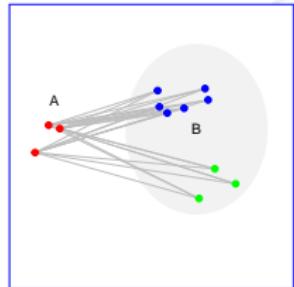
- the distance is the minimal distance between observation pairs
- tends to produce chains of clusters
- fast, performs well on non-globular data, but it performs poorly in the presence of noise
- $\alpha_i = 1/2, \alpha_j = 1/2, \beta = 0, \gamma = -1/2$

Complete linkage



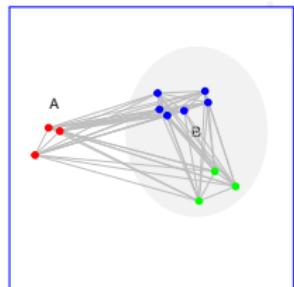
- the distance is the maximal distance between observation pairs
- tends to produce compact clusters
- perform well on cleanly separated globular clusters, but have mixed results otherwise
- $\alpha_i = 1/2, \alpha_j = 1/2, \beta = 0, \gamma = 1/2$

Average linkage between groups (Sokal and Michener 1958)



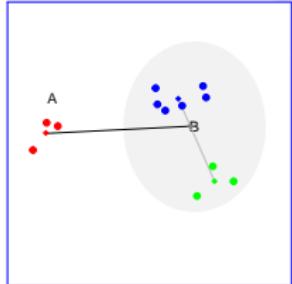
- the distance is the average distance between all observation pairs between two clusters
- $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, $\beta = 0$, $\gamma = 0$

Average linkage within groups (Sokal and Michener 1958)



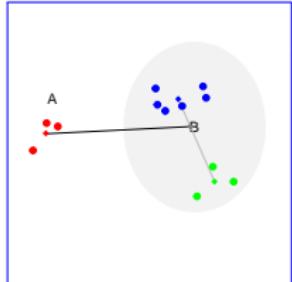
- the distance is the average distance between all observation pairs in the joined cluster
- $\alpha_i = 1/2$, $\alpha_j = 1/2$, $\beta = 0$, $\gamma = 0$

Centroid linkage



- distance between their unweighted centroids
- only for euclidean distance
- $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, $\beta = -\frac{n_i n_j}{(n_i+n_j)^2}$, $\gamma = 0$

Median linkage



- distance between their weighted centroids
- only for euclidean distance
- $\alpha_i = 1/2$, $\alpha_j = 1/2$, $\beta = -1/4$, $\gamma = 0$

Ward's minimum variance

- the distance is the variance increase

$d(i, j) = \text{variance of merged clusters} - \text{sum of weighted variances of clusters}$

- only for euclidean distance
- tends toward equal-sized clusters
- effective method for noisy data
- $\alpha_i = \frac{n_i+n_k}{n_i+n_j+n_k}$, $\alpha_j = \frac{n_j+n_k}{n_i+n_j+n_k}$, $\beta = \frac{n_k}{n_i+n_j+n_k}$, $\gamma = 0$

Lance-Williams flexible-beta

- $\alpha_i = \frac{1-\beta}{2}$, $\alpha_j = \frac{1-\beta}{2}$, $\gamma = 0$
- usually with $\beta = -1/4$

Density linkage

perform another agglomeration technique with d^* distances (mostly single linkage)

$$d^*(i,j) = \begin{cases} \frac{1}{2} \left(\frac{1}{\hat{f}(x_i)} + \frac{1}{\hat{f}(x_j)} \right) & \text{if } d(x_i, x_j) \leq m \\ \infty & \text{otherwise} \end{cases}$$

- Uniform kernel
 - ▶ choose a radius r
 - ▶ compute $\hat{f}(x_i)$ as percentage of observations found within the radius r
 - ▶ set $m := r$
- k th-Nearest-Neighbour
 - ▶ choose the number of neighbours k
 - ▶ compute the distance $r_k(x_i)$ to k th nearest neighbour
 - ▶ compute $\hat{f}(x_i)$ as percentage of observations found within the radius $r_k(x_i)$ divided by the volume of the sphere with radius $r_k(x_i)$
 - ▶ set $m := \max(r_k(x_i), r_k(x_j))$

Single linkage: Nearest-neighbor method, Minimum method, Hierarchical analysis, Space-contracting method, Elementary linkage analysis, Connectedness method

Sneath, P. H. A. (Aug. 1, 1957). "The Application of Computers to Taxonomy". In: *Microbiology* 17.1, pp. 201–226. issn: 1350-0872, 1465-2080. doi: 10.1099/00221287-17-1-201. url: <http://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-17-1-201> (visited on 12/06/2016).

Sibson, R. (Jan. 1, 1973). "SLINK: An optimally efficient algorithm for the single-link cluster method". In: *The Computer Journal* 16.1, pp. 30–34. issn: 0010-4620, 1460-2067. doi: 10.1093/comjnl/16.1.30. url: <http://comjnl.oupjournals.org/cgi/doi/10.1093/comjnl/16.1.30> (visited on 12/09/2016).

Complete linkage: Furthest-neighbor method, Maximum method, Compact method, Space-distorting method, Space-dilating method, Rank-order typal analysis, Diameter analysis

Macnaughton-Smith, P.N.M. (1965). *Some Statistical and Other Numerical Techniques for Classifying Individuals*. Research Unit Report 6. London: Her Majesty's Stationery Office.

Average linkage between groups: Average linkage, Arithmetic-average clustering, Unweighted pair-group method using arithmetic averages (UPGMA), Unweighted clustering, Group-average method, Unweighted group mean, Unweighted pair-group method

Average linkage within groups: Weighted-average linkage, McQuitty's method, Weighted pair-group method using arithmetic averages (WPGMA), Weighted group-average method

Centroid linkage: Unweighted centroid method, Unweighted pair-group centroid method (UPGMC), Nearest-centroid sorting

Sokal, R. R. and Michener, C. D. (1958). "A statistical method for evaluating systematic relationships". In: *University of Kansas Science Bulletin* 38, pp. 1409–1438.

Median linkage: Gower's method, Weighted centroid method, Weighted pair-group centroid method (WPGMC), Weighted pair method, Weighted group method

Gower, J. C. (Dec. 1967). "A Comparison of Some Methods of Cluster Analysis". In: *Biometrics* 23.4, p. 623. issn: 0006341X. doi: 10.2307/2528417. url: <http://www.jstor.org/stable/2528417?origin=crossref> (visited on 12/07/2016).

Ward's minimum variance: Minimum-variance

method, Error-sum-of-squares method, Hierarchical grouping to minimize $tr(W)$, HGROUP

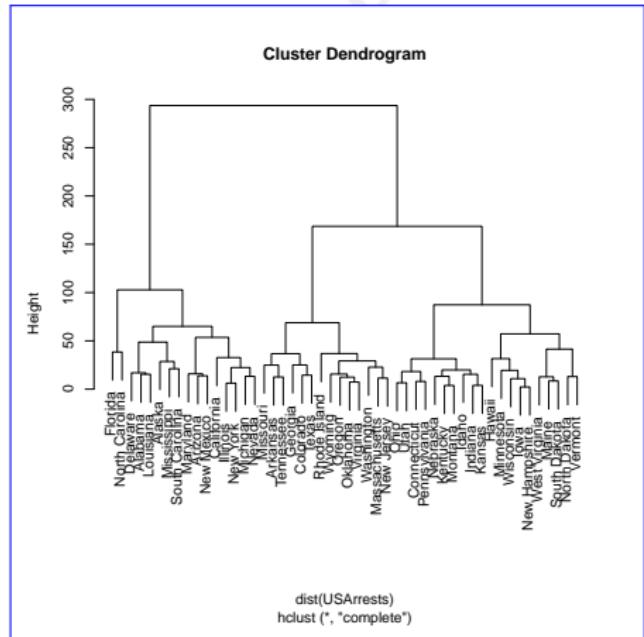
Ward, Joe H. (Mar. 1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1963.10500845. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845> (visited on 12/08/2016).

 Listing 16.9: example_agglom.R

```
1 zfaithful <- scale(faithful)
2 d <- dist(zfaithful)
3 # hclust
4 cl1 <- hclust(d)
5 memb <- cutree(cl1, 3)
6 plot(zfaithful, col=memb)
7 # agnes
8 library("cluster")
9 cl2 <- agnes(d)
10 memb <- cutree(cl2, 3)
11 plot(zfaithful, col=memb)
```

- ② hclust(d, method="complete")
- ② cutree(hclust, k=NULL, h=NULL)
- ② fastcluster::hclust(d, method="complete")
- ② cluster::agnes(x, diss=inherits(x, "dist"), metric="euclidean", stand=FALSE, method="average")

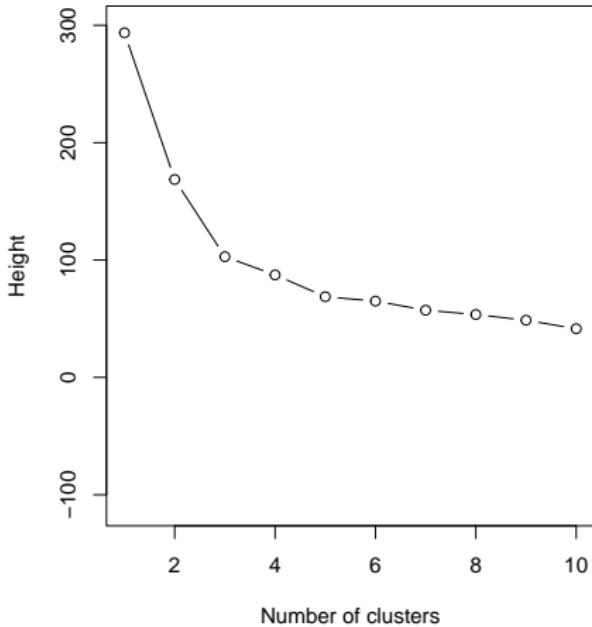
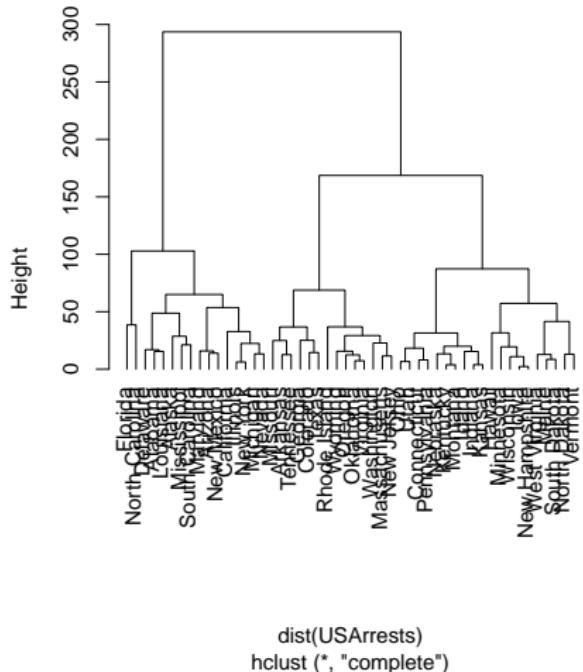
Dendrogram



- visualize the agglomeration process in a scatterplot
- on x-axis put the observations
- on y-axis the distance or dissimilarity measure (or a proxy)
- to choose for the number of clusters look for jumps in y direction

Dendrogram II

Cluster Dendrogram with $\text{hang}=-1$





Listing 16.10: example_dendro.R

```
1 zfaithful <- scale(faithful)
2 d <- dist(zfaithful)
3 # hclust
4 cl1 <- hclust(d)
5 plot(cl1)
6 # agnes
7 library("cluster")
8 cl2 <- agnes(d)
9 plot(as.dendrogram(cl2))
```

```
R plot(hclust, hang=0.1)
R dendrapply(dendr, func, ,...)
```

EM-Clustering

- Model-based clustering
- Assumptions:
 - ▶ K clusters
 - ▶ each cluster follows p dimensional multivariate normal distribution with parameters μ_j and Σ_j
 - ▶ general density function of a multivariate normal distribution

$$f(x, \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)}{(2\pi)^{p/2} |\Sigma|^{1/2}}$$

- Density function for a mixture of K gaussians

$$f(x) = \sum_k \pi_k f(x, \mu_k, \Sigma_k)$$

- ▶ with $\sum_k \pi_k = 1$

- Choose start values $\hat{\pi}_j^{(0)}$, $\hat{\mu}_j^{(0)}$ and $\hat{\Sigma}_j^{(0)}$
 - ▶ randomly
 - ▶ from k-means or hierarchical cluster analysis
- Bayes-Theorem: probability that observation x belongs to cluster j

$$P(C_j|x) = \frac{P(x|C_j)P(C_j)}{\sum_k P(x|C_k)P(C_k)}$$

- E-step: probability that observation x_i belongs to cluster j

$$P^{(t+1)}(C_j|x_i) = \frac{\hat{\pi}_j^{(t)} f(x_i, \hat{\mu}_j^{(t)}, \hat{\Sigma}_j^{(t)})}{\sum_k \hat{\pi}_k^{(t)} f(x_i, \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)})}$$

- M-step: recompute the estimates for μ_j and Σ_j

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n x_i P^{(t+1)}(C_j|x_i)}{\sum_{i=1}^n P^{(t+1)}(C_j|x_i)}$$

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^n P^{(t+1)}(C_j|x_i)(x_i - \hat{\mu}_j^{(t+1)})^T(x_i - \hat{\mu}_j^{(t+1)})}{\sum_{i=1}^n P^{(t+1)}(C_j|x_i)}$$

$$\hat{\pi}_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n P^{(t+1)}(C_j|x_i)$$

- Stop the iteration if either maxit iterations reached or the differences between $|\hat{\mu}_j^{(t+1)} - \hat{\mu}_j^{(t)}|$ etc. become too small

 Listing 16.11: example_em.R

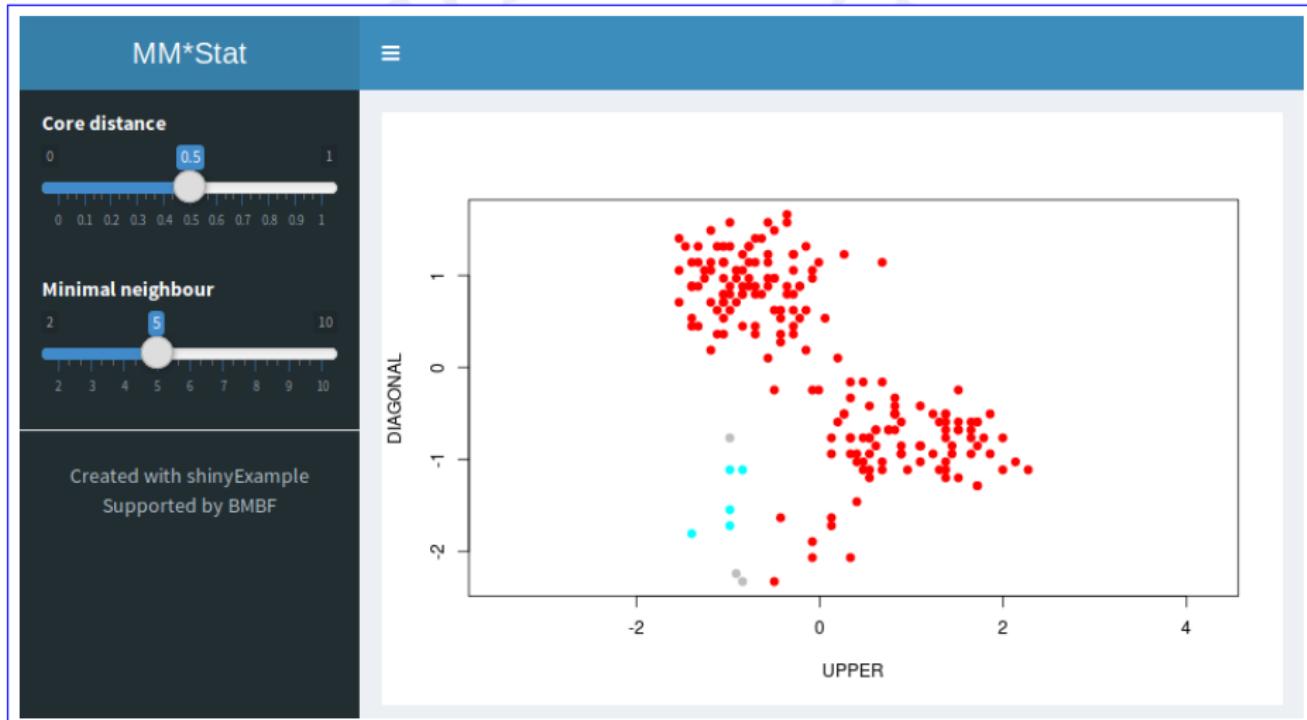
```
1 library("mclust")
2 # since multivariate normal densities are assumed
3 # likelihood theory can be applied, e.g. BIC for
4 # optimal cluster choice
5 cl <- Mclust(faithful)
6 print(cl)
7 summary(cl)
8 par(mfrow=c(2,2))
9 plot(cl, "BIC")
10 plot(cl, "classification")
11 plot(cl, "uncertainty")
12 plot(cl, "density")
13 # model names
14 ?mclustModelNames
```

 mclust::Mclust(data, G=1:9)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

- define a core distance ϵ and minimal neighbour number n_{min}
- for *core observations* holds $\sum_j I(d(i,j) < \epsilon) > n_{min}$
- all core observations with $d(i,j) < \epsilon$ belong to the same core cluster
- all non-core observations with $d(non_core, core) < \epsilon$ belong to the core cluster
 - ▶ a non-core observation could belong to several core clusters
 - ⇒ make a random assignment
- all observations which do not belong to a cluster are *noise observations*
- DBSCAN is the base of a lot of further cluster algorithms

Ester, Martin et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proc. of 2nd International Conference on Knowledge Discovery and Data Mining. Portland, OR, pp. 226–231.



 Listing 16.12: example_dbSCAN.R

```
1 library("fpc")
2 cl <- dbSCAN(faithful, 0.05, scale=T)
3 col <- c('grey', rainbow(max(cl$cluster)))
4 plot(faithful, pch=19, col=col[1+cl$cluster])
5 #
6 cl <- dbSCAN(faithful, 0.1, scale=T)
7 col <- c('grey', rainbow(max(cl$cluster)))
8 plot(faithful, pch=19, col=col[1+cl$cluster])
9 #
10 cl <- dbSCAN(faithful, 0.15, scale=T)
11 col <- c('grey', rainbow(max(cl$cluster)))
12 plot(faithful, pch=19, col=col[1+cl$cluster])
```

- ¶ fpc::dbSCAN(data, eps, MinPts=5, scale=FALSE)
- ¶ dbSCAN::dbSCAN(data, eps, MinPts=5, weights=NULL,
borderPoints=TRUE)

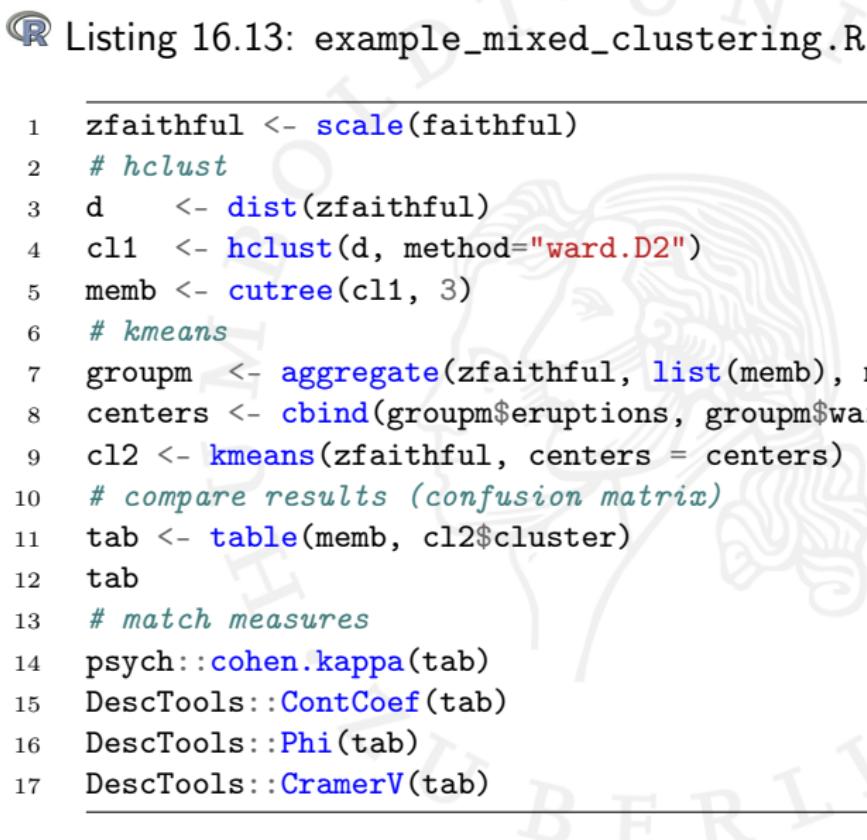
Mixed clustering

- combine advantages of two methods
 - ▶ run first a hierarchical cluster analysis with Ward's method
 - ▶ determine the number of clusters from the dendrogram
 - ▶ run a k -means cluster analysis to improve the solution
- compare two clustering results with a matching/confusion matrix
 - ▶ make a crosstable of cluster memberships

	1	2	3
1	60	0	5
2	0	97	0
3	8	0	102

rows: ward, cols: kmeans

- ▶ if compared with the true membership then the crosstable is called “confusion matrix”
- ▶ to judge the match quality use for example Cohen's κ or an association coefficient

R Listing 16.13: example_mixed_clustering.R

```
1 zfaithful <- scale(faithful)
2 # hclust
3 d      <- dist(zfaithful)
4 cl1   <- hclust(d, method="ward.D2")
5 memb <- cutree(cl1, 3)
6 # kmeans
7 groupm <- aggregate(zfaithful, list(memb), mean)
8 centers <- cbind(groupm$eruptions, groupm$waiting)
9 cl2 <- kmeans(zfaithful, centers = centers)
10 # compare results (confusion matrix)
11 tab <- table(memb, cl2$cluster)
12 tab
13 # match measures
14 psych::cohen.kappa(tab)
15 DescTools::ContCoef(tab)
16 DescTools::Phi(tab)
17 DescTools::CramerV(tab)
```

Measures to compare two classifications

If C is a r cluster solution (rows) and C' is a c cluster solution (columns) then define

- Maximum-Match-Measure (or Meilă-Heckerman-Measure)
 1. find the largest entry $m_{ii'}$ and match cluster i and i'
 2. delete row i and column i' from the matrix and repeat until the matrix has size zero
 3. compute $MM(C, C') = \frac{1}{n} \sum_{i=1}^{\min(r,c)} m_{ii'}$ from all steps and n the number of observations
- Van Dongen-Measure (distance)
 - ▶ for all rows and columns compute the corresponding maximal entries $m_{rr'}$ and $m_{cc'}$

$$D(C, C') = 2n - \sum_r m_{rr'} - \sum_c m_{cc'}$$

- both measures consider only the matching clusters

- Mutual information
- how much we can reduce the uncertainty about the cluster of a random element when knowing its cluster in another clustering of the same set of elements?

► let $P(i) = \frac{\text{obs in cluster } i}{n}$

► let $P(i,j) = \frac{\# \text{ of obs in cluster } i \text{ of } C \text{ and cluster } j \text{ of } C'}{n}$

► entropy of a k cluster solution:

$$H(C) = - \sum_{i=1}^r P(i) \log_2(P(i))$$

► $H(C)$ is a measure for the uncertainty about the cluster of a randomly picked element

$$MI(C, C') = \sum_{i=1}^r \sum_{j=1}^c P(i,j) \log_2 \left(\frac{P(i,j)}{P(i)P(j)} \right)$$

- Dongen, Stijn (2000). *Performance Criteria for Graph Clustering and Markov Cluster Experiments*. Tech. rep. NLD.
- Meilă, Marina and Heckerman, David (2001). "An Experimental Comparison of Model-Based Clustering Methods". In: *Machine Learning* 42.1/2, pp. 9–29. issn: 08856125. doi: 10.1023/A:1007648401407. url: <http://link.springer.com/10.1023/A:1007648401407> (visited on 08/19/2021).
- Meilă, Marina (May 2007). "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5, pp. 873–895. issn: 0047259X. doi: 10.1016/j.jmva.2006.11.013. url: <https://linkinghub.elsevier.com/retrieve/pii/S0047259X06002016> (visited on 08/19/2021).

- Wong's hybrid clustering

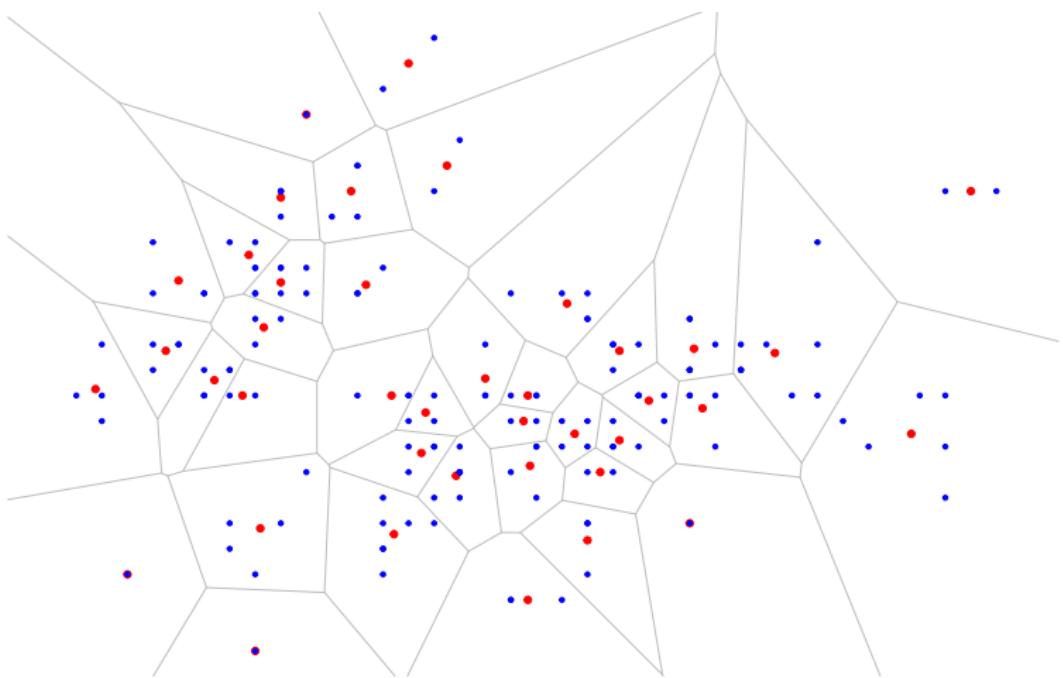
- ▶ precluster with k -means and compute cluster centers \bar{x}_i
- ▶ apply agglomerative hierarchical clustering with single-linkage

$$d^*(i, j) = \begin{cases} \frac{(w_i + w_j + 0.25(n_i + n_j)d(\bar{x}_i, \bar{x}_j))^{p/2}}{(n_i + n_j)^{1+p/2}} & \text{if } i \text{ and } j \text{ are adjacent} \\ \infty & \text{otherwise} \end{cases}$$

- ▶ compute the total variances w_i for each pre-cluster
- ▶ two clusters are adjacent if $d^2(\bar{x}_i, \bar{x}_j) \leq d^2(\bar{x}_i, \bar{x}_m) + d^2(\bar{x}_m, \bar{x}_j)$

- Problem: how to choose the (smoothing) parameters p , r and k ?

Wong, M. Anthony (Dec. 1982). "A Hybrid Clustering Method for Identifying High-Density Clusters". In: *Journal of the American Statistical Association* 77.380, pp. 841–847. issn: 0162-1459, 1537-274X. doi: 10.1080/01621459.1982.10477896. url: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477896> (visited on 12/06/2016).



Two-Step-Clustering in SPSS

- Aim: cluster huge amount of observations
 - ▶ only implemented in SPSS
- Idea
 - ▶ precluster the data
 - ▶ apply hierarchical clustering on (the cluster centers of) the preclusters

- only metric variables: euclidean distance

- general: decrease of log-likelihood function

$$dist(A, B) = \log(L(A)) + \log(L(B)) - \log(L(A + B))$$

- assumptions:

- ▶ all variables are independent \Rightarrow "likelihood" is a product
- ▶ metric variables follow a normal distribution
- ▶ categorical variables follow a multinomial distribution

$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = s^2, \hat{\pi}_i = \frac{n_i}{n} \text{ for category } i$$

- Continuous variable

$$L(x_1, \dots, x_n) = \prod_{i=1}^{n_\bullet} \frac{1}{s_\bullet \sqrt{2\pi}} \exp \left(-0.5 \frac{(x_i - \bar{x}_\bullet)^2}{s_\bullet^2} \right)$$

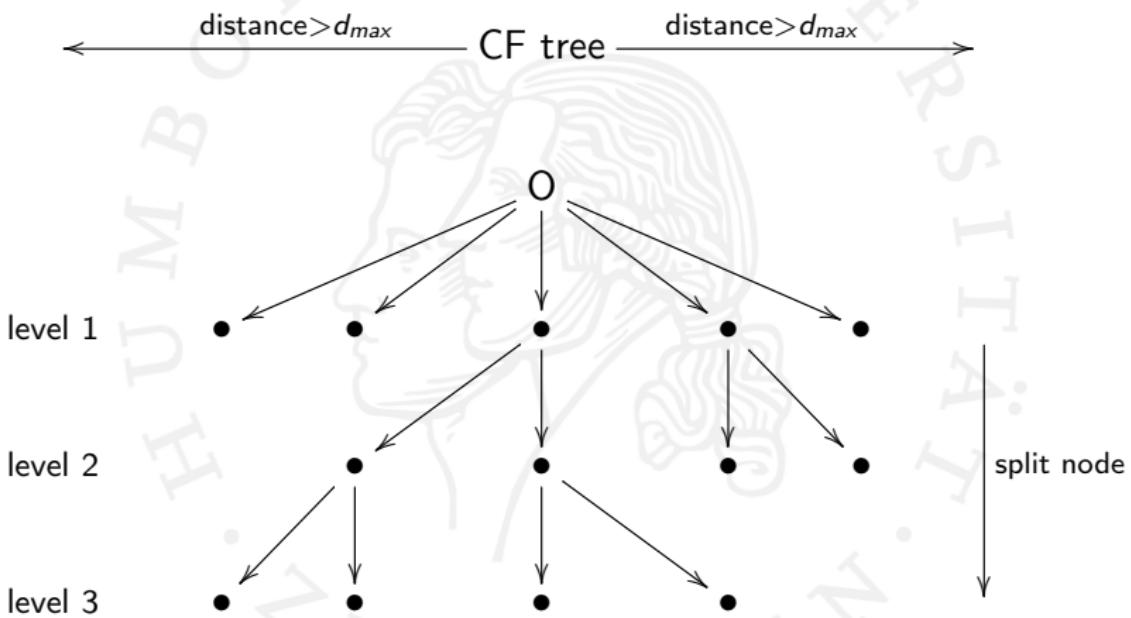
- ▶ \bar{x}_\bullet mean in cluster •
- ▶ s_\bullet^2 variance in cluster •
- ▶ n_\bullet number of observations in cluster •

- Categorical variable

$$L(n_{1,\bullet}, \dots, n_{J,\bullet}) = \frac{n_\bullet!}{\prod_{j=1}^J n_{j,\bullet}!} \prod_{j=1}^J \hat{\pi}_{j,\bullet}^{n_{j,\bullet}}$$

- ▶ $n_{j,\bullet}$ number of observations in category j in cluster •
- ▶ $\hat{\pi}_{j,\bullet} = n_{j,\bullet}/n_\bullet$

- Create a cluster feature tree (CF tree)
 - ▶ for each continuous variable hold mean and variance
 - ▶ for each categorical variable hold category counts
 - ▶ if a new observation is assigned to a cluster then the mean, variance and category counts can be recalculated without accessing previous observations
- Define a maximal distance d_{max} and a maximal number of observations n_{max} per cluster
- If $d(x, \bullet) \leq d_{max}$ then assign observation x to nearest precluster
 - ▶ if $n_\bullet > n_{max}$ then split cluster \bullet into two subpreclusters
- If $d(x, \bullet) > d_{max}$ then create a new precluster
- If the number of preclusters > 512 or CF tree level > 3 then redefine d_{max} and reclassify
- Note: preclustering depends on the order of observations
- Cluster the preclusters with agglomerative hierarchical clustering



Autoclustering

- balance explanatory power and complexity (minimum)

$$\begin{aligned} BIC &= -2 \log(L(C_1, \dots, C_K)) + \log(n)k \\ AIC &= -2 \log(L(C_1, \dots, C_K)) + 2k \\ &\qquad\qquad\qquad \text{explanatory power} \qquad\qquad\qquad \text{complexity} \end{aligned}$$

- ▶ K number of cluster
- ▶ k number of parameters (depends on K)
- Compute BIC for $1, \dots, K_{max}$ cluster(s)
- Rearrange “nearest” clusters in hierarchical clustering to achieve better BIC

Number of clusters

- $K_{opt} \approx \sqrt{n}$ or $K_{opt} \approx \sqrt{n}/2$
- Total variance explained (elbow criterion)

$$\frac{CSS}{TSS}$$

- ▶ Total sum of squares ($\approx n \times$ total variance)

$$TSS = \sum_{i=1}^n d^2(x_i, \bar{x}) = \sum_{i=1}^n d^2(x_i, \bar{x}_{k(i)}) + \sum_{i=1}^n d^2(\bar{x}_{k(i)}, \bar{x})$$

- ▶ Center sum of squares (replace obs. by its cluster center)

$$CSS = \sum_{i=1}^n d^2(\bar{x}_{k(i)}, \bar{x})$$

- Crossvalidation

- ▶ decompose the data randomly into c parts
- ▶ compute cluster solutions with $c - 1$ parts
- ▶ sum the distance between the observations in the last part and the nearest cluster (center)
- ▶ cycle over all parts and average the distances
- ▶ choose K_{opt} as the cluster number with the minimum of averaged distances

- Calinski-Harabasz index or Pseudo F -value (look for maximum)

$$F = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{n-k}}$$

- ▶ Within-cluster dispersion and pooled within-cluster sum of squares

$$WGSS_k = \sum_{x_i \in C_k} d^2(x_i, \bar{x}_k) \text{ and } WGSS = \sum_k WGSS_k$$

- ▶ Between-group dispersion

$$BGSS = \sum_k d^2(\bar{x}_k, \bar{x})$$

Charrad, Malika et al. (2014). “NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set”. In: *Journal of Statistical Software* 61.6. issn: 1548-7660. doi: 10.18637/jss.v061.i06. url: <http://www.jstatsoft.org/v61/i06/> (visited on 08/29/2016).

 Listing 16.14: example_cluster_number.R

```
1 library("NbClust")
2 zfaithful <- scale(faithful)
3 # Total variance explained
4 tve <- rep(NA, 15)
5 for (k in 2:15) {
6     clk <- kmeans(zfaithful, k)
7     tve[k] <- 1-clk$tot.withinss/clk$totss
8 }
9 plot(tve, type="b")
10 # Calinski and Harabasz
11 NbClust(zfaithful, method="ward.D2", index="ch")
12 # All indices
13 NbClust(zfaithful, method="ward.D2")
14 # All indices (takes a long time)
15 #NbClust(zfaithful, method="ward.D2", index="alllong")
```

 NbClust::NbClust(data, method, index="all")

Silhouette

- Distance between observation x and cluster C

$$d(x, C) = \frac{1}{n_C} \sum_{c \in C} d(c, x)$$

- Let $x \in A$ and B the nearest cluster to x (except A)

$$d(x, B) = \min_{C \neq A} d(x, C)$$

- The silhouette $S(x)$ is defined as

$$S(x) = \begin{cases} 0 & \text{if } d(x, A) = 0 \\ \frac{d(x, B) - d(x, A)}{\max(d(x, B), d(x, A))} & \text{otherwise} \end{cases}$$

- If $0 < d(x, A) \leq d(x, B)$ then

$$S(x) = 1 - \frac{d(x, A)}{d(x, B)} \leq 1$$

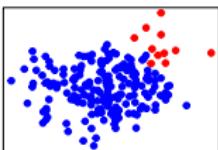
- Interpretation of $S(x)$
 - ▶ $S(x) < 0$: clustering can be improved
 - ▶ $S(x) \approx 0$: x lies between two clusters
 - ▶ $0 < S(X) \leq 0.25$: *no structure*
 - ▶ $0.25 < S(X) \leq 0.5$: *weak structure*
 - ▶ $0.5 < S(X) \leq 0.75$: *medium structure*
 - ▶ $0.75 < S(X) \leq 1$: *strong structure*
- Silhouette coefficient of a data set or cluster

$$s = \frac{1}{n} \sum S(x)$$

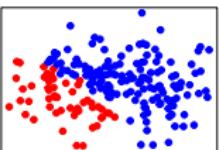
- Silhouette plot
 - ▶ for each cluster plot as- or descending silhouettes

Rousseeuw, Peter J. (Nov. 1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. issn: 03770427. doi: 10.1016/0377-0427(87)90125-7. url: <http://linkinghub.elsevier.com/retrieve/pii/0377042787901257> (visited on 08/29/2016).

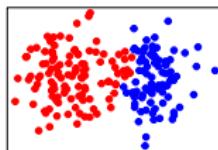
Data



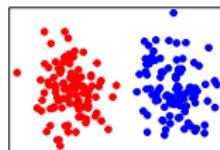
Data



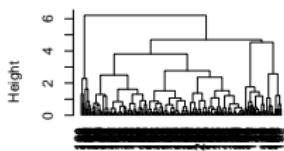
Data



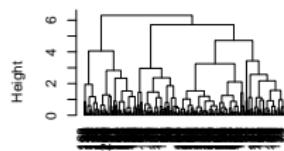
Data



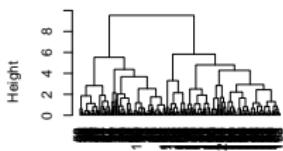
Cluster Dendrogram



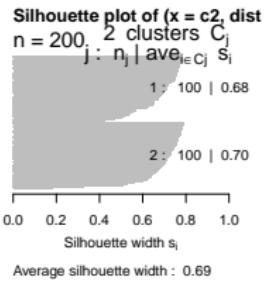
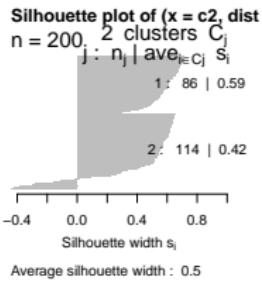
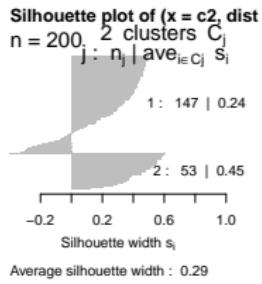
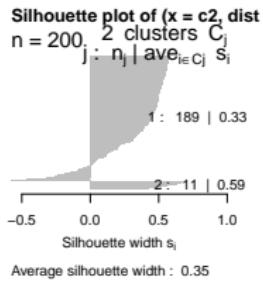
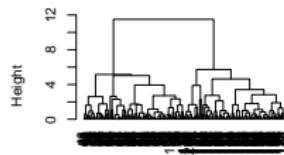
Cluster Dendrogram



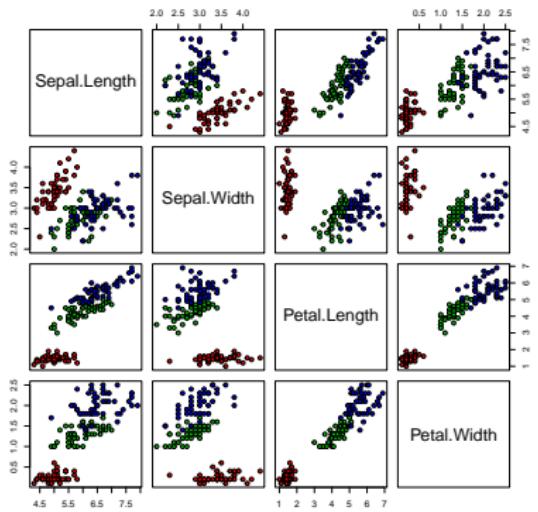
Cluster Dendrogram



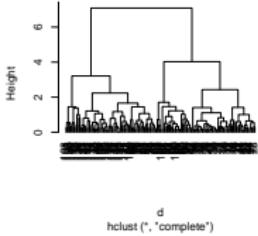
Cluster Dendrogram



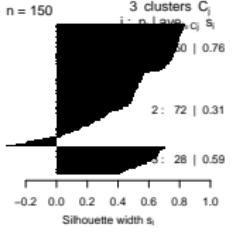
Iris Data (red=setosa,green=versicolor,blue=virginica)



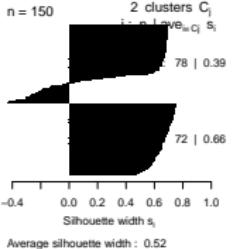
Cluster Dendrogram



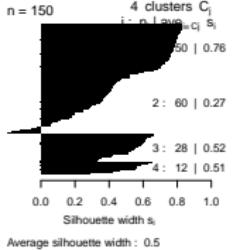
Silhouette plot of ($x = c_2$, dist = d)



Silhouette plot of ($x = c_2$, dist = d)



Silhouette plot of ($x = c_4$, dist = d)



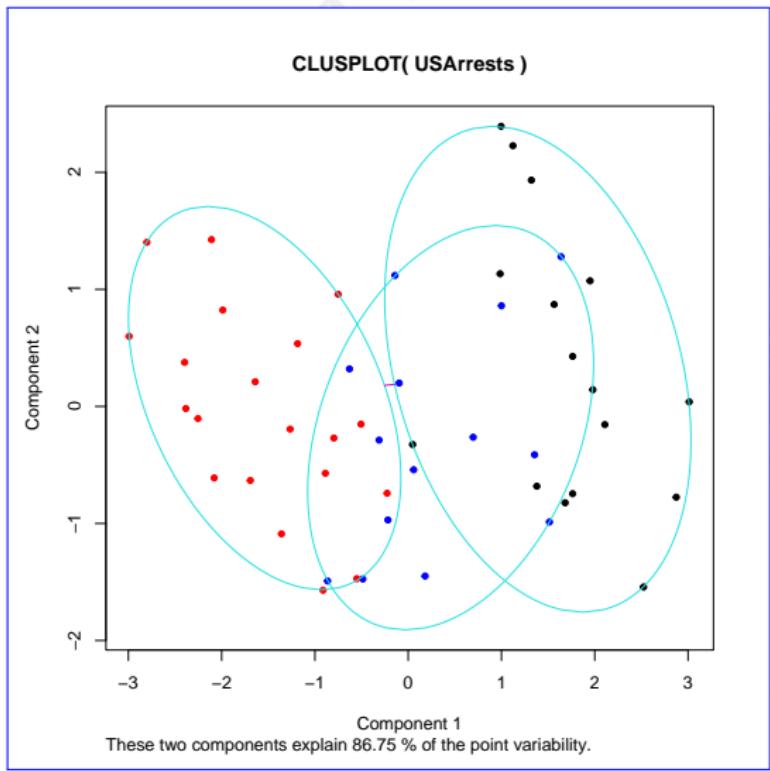


Listing 16.15: example_silhouette.R

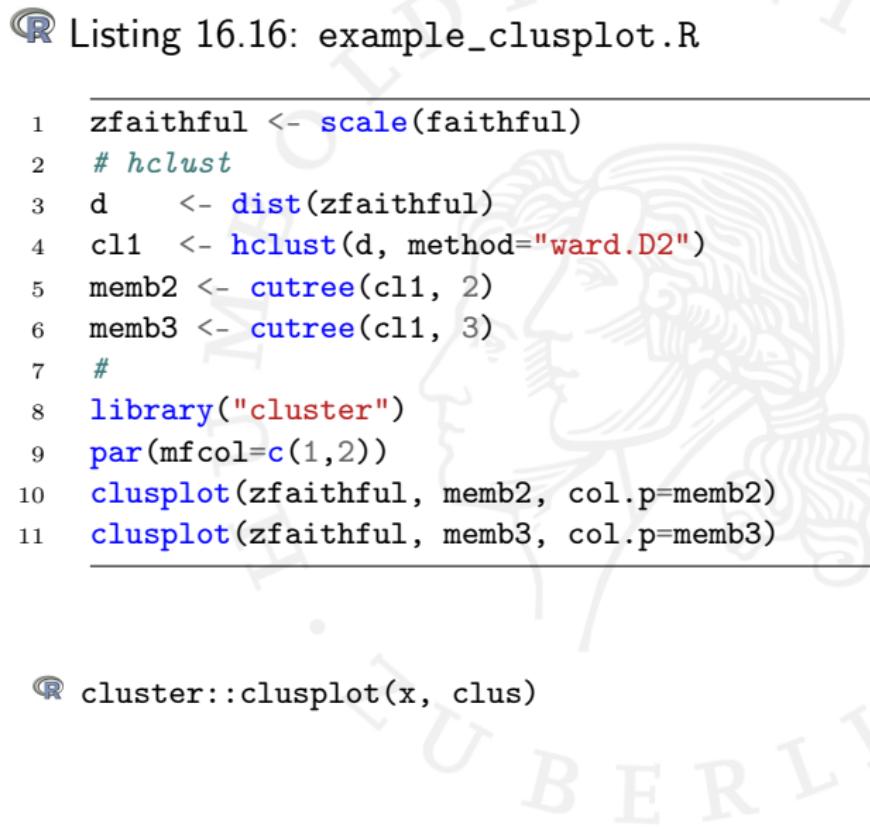
```
1 zfaithful <- scale(faithful)
2 # hclust
3 d      <- dist(zfaithful)
4 cl1   <- hclust(d, method="ward.D2")
5 memb2 <- cutree(cl1, 2)
6 memb3 <- cutree(cl1, 3)
7 #
8 library("cluster")
9 par(mfcol=c(2,2))
10 plot(zfaithful, col=memb2)
11 s2 <- silhouette(memb2, d)
12 plot(s2, col=1:2, border=NA)
13 plot(zfaithful, col=memb3)
14 s3 <- silhouette(memb3, d)
15 plot(s3, col=1:3, border=NA)
```

R cluster::silhouette(tree, dist)

Cluster visualization



- visualize the clustering result via a scatterplot
- on x-axis put the first principal component
- on y-axis the the second principal component
- Idea: the first two components (PCA or MDS) capture most of the variance

A large, faint watermark of a stylized head profile, facing right, composed of concentric curved lines.

R Listing 16.16: example_clusplot.R

```
1 zfaithful <- scale(faithful)
2 # hclust
3 d      <- dist(zfaithful)
4 cl1   <- hclust(d, method="ward.D2")
5 memb2 <- cutree(cl1, 2)
6 memb3 <- cutree(cl1, 3)
7 #
8 library("cluster")
9 par(mfcol=c(1,2))
10 clusplot(zfaithful, memb2, col.p=memb2)
11 clusplot(zfaithful, memb3, col.p=memb3)
```

R cluster::clusplot(x, clus)

Regression analysis

November 3, 2022

Correlation, causality and regression • Regression model • Steps in regression analysis

Correlation, causality and regression

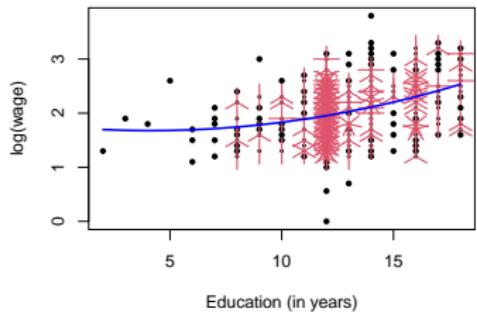
- Correlation: indication of a relationship between two or more variables
 - Correlation between A and B may come
 - ▶ $A \Rightarrow B$: often wrongly assumed
 - ▶ $B \Rightarrow A$: Wind is caused by the rotation of windmills.
 - ▶ $A \Leftarrow C \Rightarrow B$: Stork brings the babies.
 - ▶ By chance: Does the moon control gold and silver prices?
 - Regression: model of a relationship between two or more variables, requires usually theoretical knowledge
 - Neither a regression model nor correlation imply any causality
- ⚠ Statistics alone CAN NOT prove causality!**

Lucey, Brian M. (June 2010). "Lunar seasonality in precious metal returns?" en. In: *Applied Economics Letters* 17.9, pp. 835–838. issn: 1350-4851, 1466-4291. doi: 10.1080/17446540802516188. url: <http://www.tandfonline.com/doi/abs/10.1080/17446540802516188> (visited on 08/09/2015).

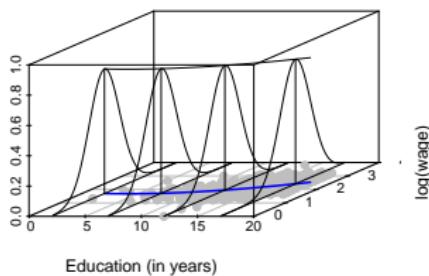
Regression model

- Basic concepts
 - ▶ tendency of a response variable Y to vary systematically with a predictor variable X
 - ▶ a scattering of points around a curve
- Regression model postulate
 - ▶ for each level of X is a probability distribution of Y
 - ▶ the means of the probability distribution of Y varies systematically with X
- Construction of regression models
 - ▶ selection of predictor variables
 - ★ contribution to the variation of Y
 - ★ importance for causal relationship
 - ▶ functional form of regression relation (linear, quadratic or non-parametric)
 - ▶ scope of model (range of predictor variables)

Sunflower plot of CPS 1985 data



Regression model



$$Y = m(X) + U$$

$$Y_i = m(x_{i1}, \dots, x_{ip}) + U_i \quad (i = 1, \dots, n)$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + U_i \quad (\text{linear model})$$

- All regression models are wrong, because they are a simplification
 - ▶ the imposed dependencies might be wrong
 - ▶ not all variables can be included in the model
 - ▶ the model assumptions might be violated
- George Box (1979)
All models are wrong but some are useful.
- Aim of regression model
 - ▶ prediction
 - ▶ interpretation

Box, G.E.P. (1979). "Robustness in the Strategy of Scientific Model Building". In: *Robustness in Statistics*. Elsevier, pp. 201–236. isbn: 978-0-12-438150-6. url:
<http://linkinghub.elsevier.com/retrieve/pii/B9780124381506500182> (visited on 11/25/2016).

Steps in regression analysis

- Model development
 1. exploratory data and regression analysis
 2. develop one or more tentative regression models
 3. is the regression model suitable for the data at hand?
 4. if not then revise model or develop a new one and go to step 3
 5. identify most suitable model
 6. make inference on basis of the chosen regression model
- For all models do
 1. model parameters: estimate, interpret and test (confidence interval)
 2. assess the model quality
 3. analysis of residuals
 4. identification of outliers and influential observations
 5. model improvement

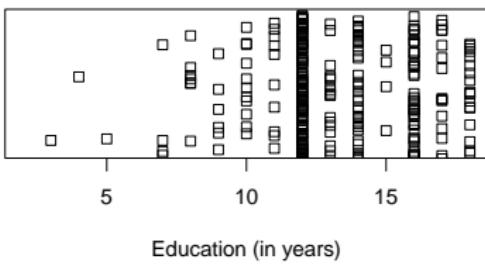
Linear regression

November 3, 2022

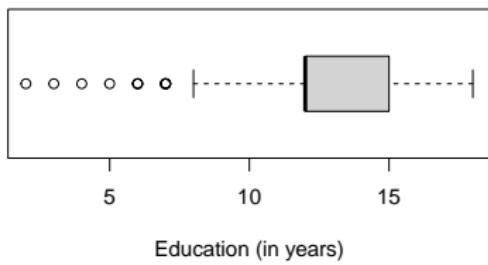
- Predictor plots
- Model assumptions
- Regression coefficients
- Tests and confidence intervals
- Bootstrap testing
- Standardized regression coefficients
- Coefficient of determination
- Test on R^2
- Multicollinearity
- Residuals
- Tests on residuals
- White test
- Breusch-Pagan test
- Durbin-Watson test
- Leverage
- Cook's distance
- Differences
- Handling of unusual observations
- Stepwise models
- Other model selection criteria
- Disadvantage of stepwise modeling
- Model error
- Shrinkage methods
- Comparing linear regression & LASSO/elastic net

Predictor plots

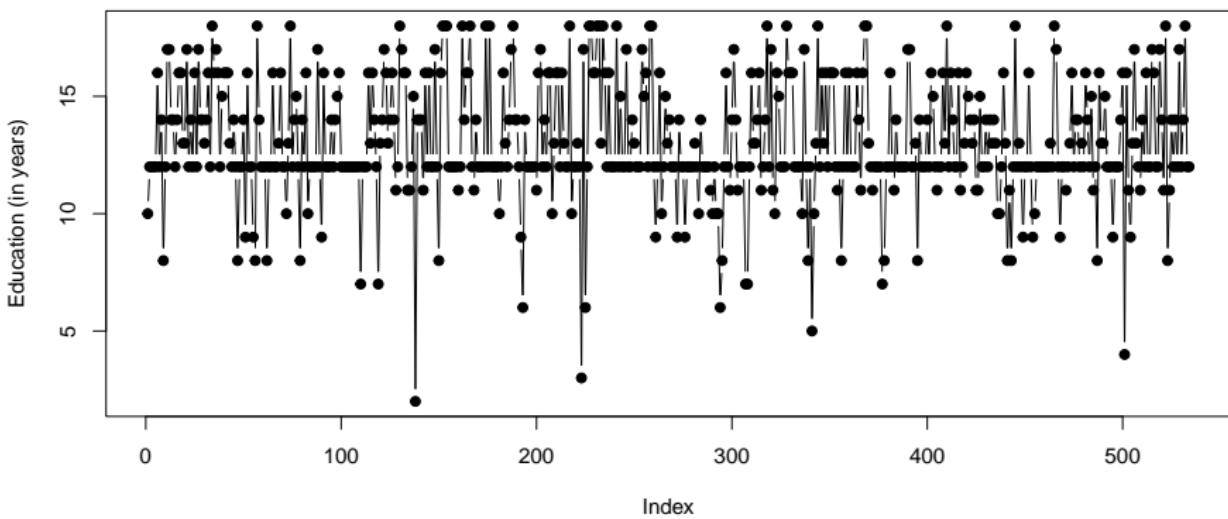
Strip plot of education



Boxplot of education



Sequence plot of education



- if a sequence is suspected in the data (time series, spatial data)

$$\bar{x} = 9$$

$$\bar{y} = 7,5$$

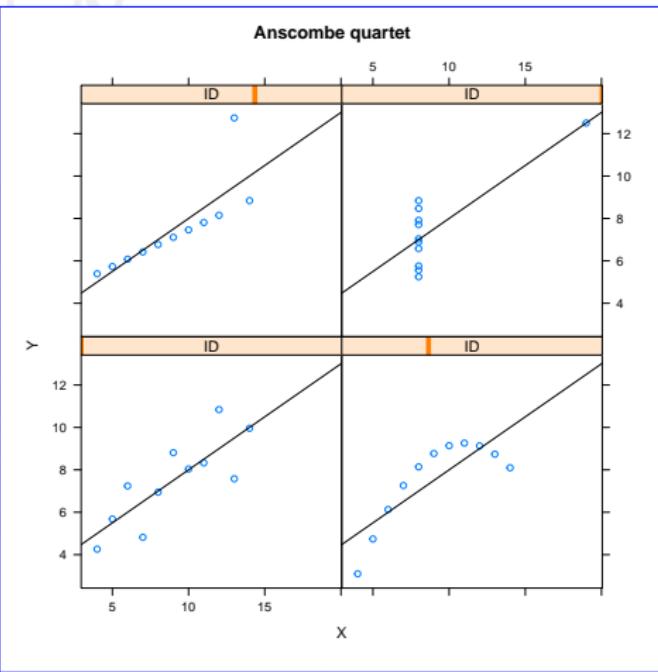
$$s_x^2 = 10$$

$$s_y^2 = 3,75$$

$$r_{xy} = 0,816$$

$$\hat{y} = 3 + 0,5x$$

- Transformation?
- Clustering?
- Outliers?



Anscombe, F. J. (1973). "Graphs in Statistical Analysis". In: *The American Statistician* 27.1, pp. 17–21. doi: 10.1080/00031305.1973.10478966. url:
<https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966>.

Model assumptions

1. X_i are non-random
2. U_i describe only random effects, all systematic effects are captured by the X_i
3. between the X_i are no functional relationships (no multicollinearity)
4. the regression function is linear in the parameters, not necessarily in the variables
5. $E(U_i) = 0$
6. $\text{Var}(U_i) = \sigma_u$ (homoscedasticity)
7. $\text{Cov}(U_{i_1}, U_{i_2}) = 0$
- 8a. $U_i \sim N(0, \sigma_u)$ or
- 8b. $U_i \sim (0, \sigma_u)$ and n large enough (such that CLT holds)

1. X_i are non-random

- *non-random* means we can do a measurement at any x_i of *our* choice
→ design of experiments
- results of linear regression hold for non-random X if:
 - ▶ conditional distribution of Y_i , given X_i , is normal and independent with conditional mean and variance as in the fixed model
 - ▶ X_i are independent and do not depend on β_i or $\text{var}(U_i)$
- otherwise use, e.g. error-in-variable models

2. U_i describe only random effects, ...

- difficult to check
- a small (adjusted) R^2 might be a hint
- analysis of the residuals might give a hint
- implies, e.g. $\text{Cov}(X_i, U_j) = 0$

3. between the X_i are no functional relationships

- Problem 1: estimation problem $X^T X$ might be singular
- Problem 2: identification problem

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

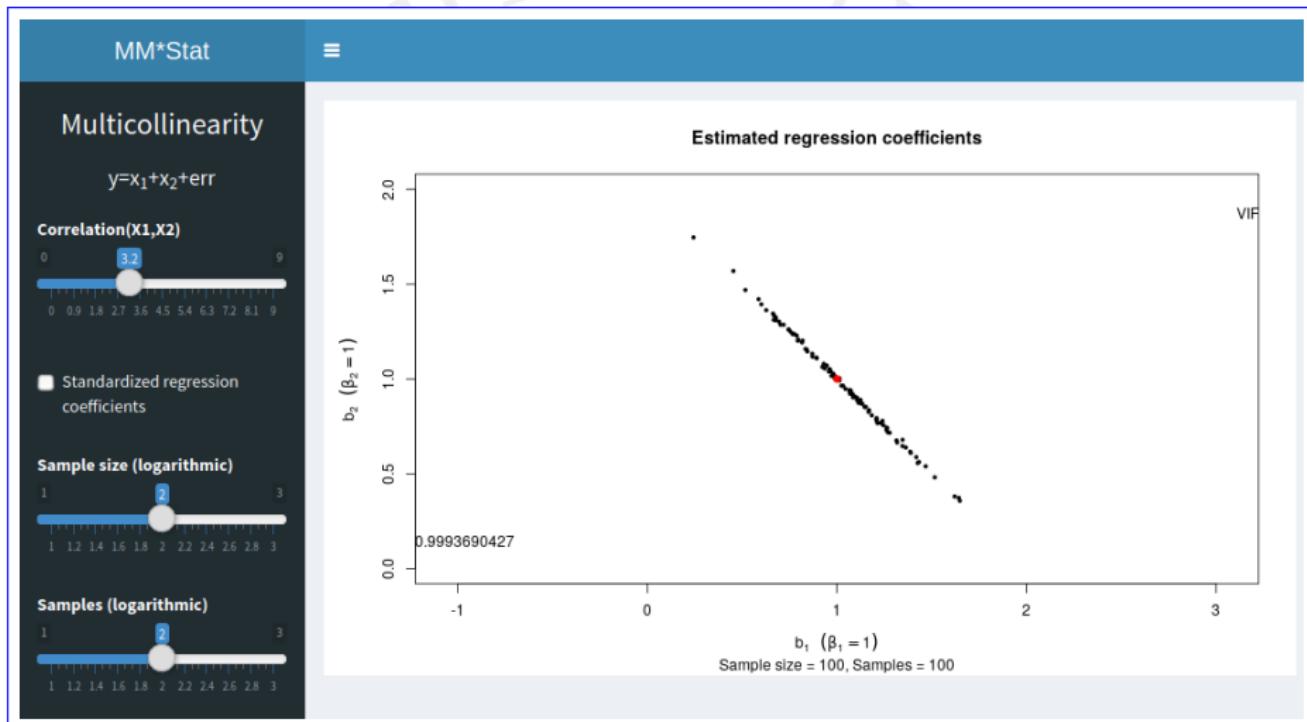
$$X_2 = \gamma_0 + \gamma_1 X_1$$

$$X_1 = \frac{1}{\gamma_1} X_2 - \frac{\gamma_0}{\gamma_1}$$

\Rightarrow

$$Y = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X_1 + U$$

$$Y = \left(\beta_0 - \frac{\gamma_0 \beta_1}{\gamma_1} \right) + \left(\beta_1 + \frac{\beta_1}{\gamma_1} \right) X_2 + U$$



4. the regression function is linear in the parameters

- the true parameters β_j are constant for all given tuples (x_{i1}, \dots, x_{ip})
- β_j dependent leads, e.g. to Hierarchical Linear Models (HLM)
- polynomial regression $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + U$ is incorporated in the linear model

5. $E(U_i) = 0$

- can not be checked, inherits from LS- and ML-estimation
- if $E(U_i) \neq 0$ then we have a systematic effect

6. $\text{Var}(U_i) = \sigma_u$

- Non-constant variance might hint to additional structure/different models
- $\text{Var}(U_i)$ not being constant can be handled, e.g. by transformations or using weighted least squares (WLS)

7. $\text{Cov}(U_{i_1}, U_{i_2}) = 0$

- $\text{Cov}(U_{i_1}, U_{i_2}) \neq 0$ can be incorporated into a more complex model, e.g. see in time series models

8. $U_i \sim N(0, \sigma_u)$

- Not necessary for estimation, but necessary for testing the coefficients
 - ▶ the Gauss-Markov theorem ensures that OLS gives the “best” results when $U_i \sim (0, \sigma_u)$ (white noise)
- Justification for normal error terms
 - ▶ the error might be composed of a lot sources (not dependent on X), therefore the CLT leads to normal distribution
 - ▶ estimation and testing is based on t distribution, which is not sensitive to moderate departure from normality (skewness!)

Regression coefficients

- Multivariate linear regression equations

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + u_i \quad (i = 1, \dots, n)$$

- Define

$$\mathfrak{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \mathfrak{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \mathfrak{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- Rewrite regression equations with matrices

$$\mathfrak{y} = \mathfrak{X}\boldsymbol{\beta} + \mathfrak{u}$$

- Least-Squares (LS) method

$$\sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2 \rightarrow \min.$$

- and Maximum-Likelihood (ML) method

$$(Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma_u))$$

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi s_u^2}} \exp\left(-\frac{(y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2}{2s_u^2}\right) \rightarrow \max.$$

- lead to the same set of equations for $\mathbf{b} = (b_0, \dots, b_p)$

Ordinary least squares

$$\mathfrak{b} = (\mathfrak{X}^T \mathfrak{X})^{-1} \mathfrak{X}^T \mathfrak{y}$$

Weighted least squares

$$\mathfrak{b}_{WLS} = (\mathfrak{X}^T \mathfrak{W} \mathfrak{X})^{-1} \mathfrak{X}^T \mathfrak{W} \mathfrak{y}$$

$$\mathfrak{X}^T \mathfrak{y} = \begin{pmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \vdots \\ \sum x_{ip} y_i \end{pmatrix}, \quad \mathfrak{W} = \begin{pmatrix} w_{11} & 0 & 0 & \dots \\ 0 & w_{22} & 0 & \dots \\ 0 & 0 & w_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\mathfrak{X}^T \mathfrak{X} = \begin{pmatrix} n & \sum x_{i1} & \dots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1} x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \sum x_{ip} x_{i1} & \dots & \sum x_{ip}^2 \end{pmatrix}$$

Call:

```
lm(formula = lwage ~ educ, data = x)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.98098	-0.37158	0.03392	0.34980	1.66100

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.059858	0.107430	9.866	<2e-16 ***
educ	0.076760	0.008091	9.488	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4885 on 532 degrees of freedom

Multiple R-squared: 0.1447, Adjusted R-squared: 0.1431

F-statistic: 90.01 on 1 and 532 DF, p-value: < 2.2e-16

 Listing 18.1: regression_simple.R

```
1 x <- read.csv2("cps78_85.csv")
2 lm <- lm(lwage~educ, data=x, subset=(year==85))
3 summary(lm)
```

 Listing 18.2: regression_plot.R

```
1 x <- read.csv2("cps78_85.csv")
2 lm <- lm(lwage~educ, data=x, subset=(year==85))
3 plot(x$educ, x$lwage)
4 abline(lm)
```

- ② lm(formula, data, subset, weights)
- ② summary(lm)
- ② abline(lm)

Examples:

$$\text{y} \sim x \quad y = \beta_0 + \beta_1 x$$

$$\text{y} \sim x+1$$

$$\text{y} \sim x-1 \quad y = \beta_1 x$$

$$\text{y} \sim x+z \quad y = \beta_0 + \beta_1 x + \beta_2 z$$

$$\text{y} \sim x+x:z \quad y = \beta_0 + \beta_1 x + \beta_2 x \cdot z$$

$$\text{y} \sim x*z-z$$

$$\text{y} \sim x*z \quad y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x \cdot z$$

$$\begin{aligned} \text{y} \sim (x+z+a)^3 \quad y &= \beta_0 + \beta_1 x + \beta_2 z + \beta_3 a + \beta_4 x \cdot z \\ &\quad + \beta_5 x \cdot a + \beta_6 a \cdot z + \beta_7 x \cdot z \cdot a \end{aligned}$$

$$\text{y} \sim I(x*z) \quad y = \beta_0 + \beta_1 x \cdot z$$

$$\text{y} \sim I(x^2) \quad y = \beta_0 + \beta_1 x^2$$

$$\text{y} \sim \text{poly}(x, 3) \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Tests and confidence intervals

$$\mathbf{b} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$$

$$B_j = \sum_{i=1}^n w_{ij}(x_1, \dots, x_n) Y_i$$

- From CLT follows normality of B_j when Y_i or U_i are not normally distributed (see assumption 8b.)

$$\frac{B_j - \beta_j}{\sigma_{B_j}} \approx N(0; 1) \Rightarrow \frac{B_j - \beta_j}{s_{B_j}} \approx t_{n-p}$$

- Covariance matrix of $B \approx N(\beta, \Sigma_B)$

$$\text{Cov}(B) = \sigma_U^2 (\mathbf{x}^T \mathbf{x})^{-1}$$

- $E(S_U^2) = \sigma_U^2$ with $s_U^2 = \frac{1}{n-p-1} \sum_{i=1}^n \hat{u}_i^2$
- $1 - \alpha$ confidence interval: $[b_j - t_{n-p;1-\alpha/2} s_{B_j}; b_j + t_{n-p;1-\alpha/2} s_{B_j}]$

Assumption(s): $U_i \sim N(0, \sigma_u^2) \Rightarrow B_j \sim N(\beta_j, s.e.(\beta_j)^2)$
 or $U_i \sim (0, \sigma_u^2) \xrightarrow{\text{CLT}} B_j \approx N(\beta_j, s.e.(\beta_j)^2)$

Hypotheses: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$

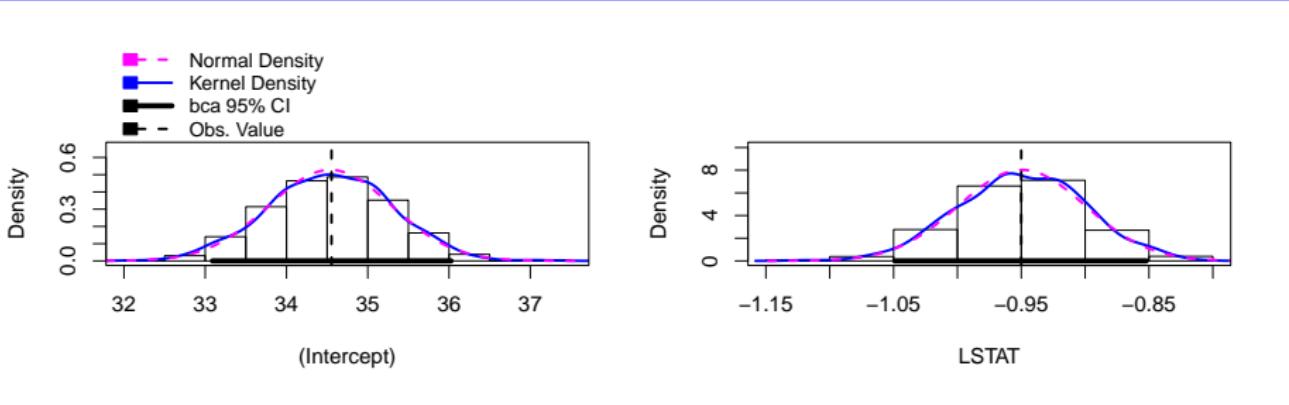
Test statistics: $T = \frac{B_j}{\widehat{s.e.}(B_j)} \sim t_{n-p-1} \approx N(0; 1)$

Reject H_0 : $|t| > t_{n-p-1;1-\alpha/2}$

Note: robust against violation of normal error distribution

Bootstrap testing

- 999 times sampled with replacement
- B_{LSTAT} is not normally distributed
- case resampling: draw with replacement from the observations
 - ▶ ok, if sample size is large enough
- smooth bootstrap: add noise (usually $N(0; \sigma_B)$) to each observation
 - ▶ resampling residuals: draw from the (studentized) residuals and add to the observation
- wild bootstrap: multiply each residual with ± 1



 Listing 18.3: example_bootcase.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 library("car")
4 lm <- lm (lwage~educ, data=cps78_85, subset=(year==85))
5 lmboot <- Boot(lm)
6 summary(lmboot)
7 confint(lmboot)
8 hist(lmboot)
```

```
② car::Boot(lm, R=999)
② car::summary.boot(object)
② car::confint.boot(object, level=0.95, type=c("bca", "norm",
      "basic", "perc"))
② car::hist.boot(object)
```

Standardized regression coefficients

- Standardized coefficients β_j^*

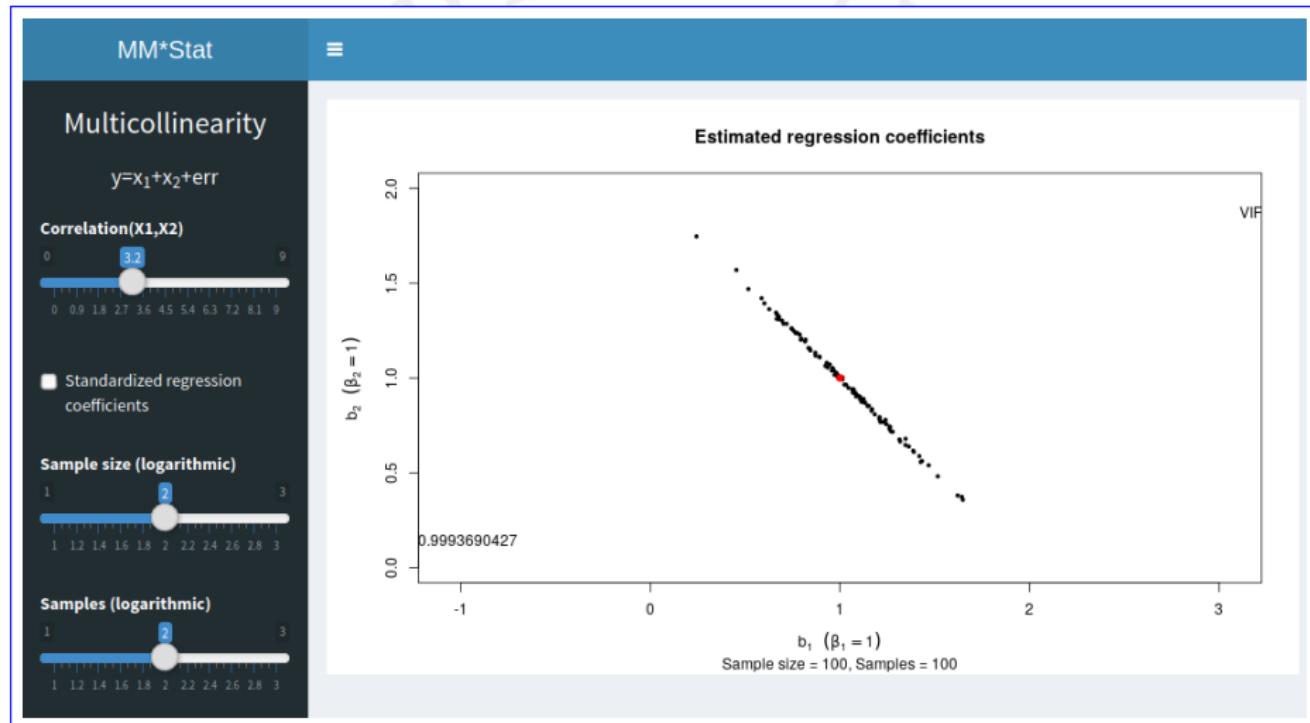
$$Z(Y) = \beta_0^* + \beta_1^* Z(X_1) + \dots + \beta_p^* Z(X_p) + V_i$$

with $Z(\bullet)$ are the standardized variables

- If no multicollinearity is present then it holds

$$1 = (\beta_1^*)^2 + \dots + (\beta_p^*)^2 + Var(V_i)$$

- ▶ $(\beta_j^*)^2$ tells how much variance is explained by an explanatory variable
- ▶ $|\beta_j^*| > 1$ hints to multicollinearity



 Listing 18.4: example_beta.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85, c("lwage", "educ"))
4 lms <- lm (scale(lwage)~scale(educ), data=xs)
5 summary(lms)
6 #
7 library("QuantPsyc")
8 lm <- lm (lwage~educ, data=xs)
9 lm.beta(lm)
```

 QuantPsyc::lm.beta(lm)

Coefficient of determination

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\underbrace{y_i - \hat{y}_i}_{=\hat{u}_i})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ R^2 &= 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ R_{\text{adj.}}^2 &= R^2 - \frac{p(1 - R^2)}{n - p - 1}\end{aligned}$$

- If $p = 1$ then holds $R^2 = r_{xy}^2$
- In linear regression holds $0 \leq R_{\text{adj.}}^2 \leq R^2 \leq 1$
- A *good* R^2 depends on area

- Typical misunderstandings
 - ▶ a large value indicates a useful prediction, but it is a relative error measure
 - ▶ a value near zero indicates no relationship, but the relationship can be non-linear
- It does not indicate whether ([Source: Wikipedia](#))
 - ▶ the independent variables are a cause of the changes in the dependent variable
 - ▶ an omitted-variable bias exists
 - ▶ the correct regression was used
 - ▶ the most appropriate set of independent variables has been chosen
 - ▶ there is collinearity present in the data on the explanatory variables
 - ▶ the model might be improved by using transformed versions of the existing set of independent variables
 - ▶ there are enough data points to make a solid conclusion.

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,862(a)	,743	,736	4.72619

a Einflußvariablen : (Konstante), LSTAT, CHAS, B, PTRATIO, ZN, CRIM, RM, INDUS, AGE, RAD, DIS, NOXSQ, TAX

ANOVA(b)

Modell	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	31635,596	13	2433,507	,000(a)
	Residuen	10967,419	491	22,337	
	Gesamt	42603,015	504		

a Einflußvariablen : (Konstante), LSTAT, CHAS, B, PTRATIO, ZN, CRIM, RM, INDUS, AGE, RAD, DIS, NOXSQ, TAX

b Abhängige Variable: MEDV

Test on R^2

Assumption(s): $U_i \sim N(0, \sigma_u^2)$

Hypotheses: $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$

$(H_0 : \beta_1 = \dots = \beta_p = 0$ vs. $H_1 : \text{at least } \beta_j \neq 0)$

Test statistics: $F = \frac{R^2(n - p - 1)}{p(1 - R^2)} \sim F_{p; n-p-1}$

Reject H_0 : $|f| > F_{p; n-p-1; 1-\alpha}$

Note: $R^2_{\text{adj.}} = R^2(1 - 1/F)$

Remarks:

- robust against violation of normal error distribution
- in fact it is an ANOVA (see regression model assumptions)

$$H_0 : \hat{y}_1 = \dots = \hat{y}_n \quad (= \bar{y})$$

$H_1 : \text{at least one } \hat{y}_i \text{ differs}$

Multicollinearity

- may increase the variance of the coefficient estimates
- tolerance: $T_j = 1 - R_{(j)}^2$
- variance inflation factor: $V_j = \frac{1}{1-R_{(j)}^2}$
 - ▶ factor how much the variance of b_j is inflated
 - ▶ $V_j > 10$ severe multicollinearity
 - ▶ $V_j > 4$ mild multicollinearity (also 2.5, 5)
- condition index: $\eta_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}$ with λ_j the eigenvalues of $X^T X$
 - ▶ $\eta_j > 30$ severe multicollinearity
 - ▶ $\eta_j > 10$ mild multicollinearity
 - ▶ e.g. for Boston Housing data $\eta_{\min} = 87, 3$
- variance proportions: the variance of β_j can be decomposed related to the eigenvalues, large values in different variables hint to highly correlated variables

Kollinearitätsdiagnose(a)

Model	Dimension	Eigenwert	Konditionsindex (Konstante)	Varianzanteile												
				CRIM	ZN	INDUS	CHAS	NOXSQ	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
1	1	10,09	1,0	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
	2	1,5	2,5	,00	,07	,08	,00	,00	,00	,00	,01	,00	,00	,00	,00	,00
	3	,96	3,2	,00	,02	,02	,00	,80	,00	,00	,00	,00	,00	,00	,00	,00
	4	,66	3,9	,00	,31	,17	,00	,12	,00	,00	,00	,00	,00	,00	,00	,00
	5	,24	6,5	,00	,51	,18	,02	,00	,00	,00	,02	,08	,01	,00	,01	,00
	6	,17	7,8	,00	,00	,25	,03	,00	,00	,02	,03	,09	,00	,00	,00	,16
	7	,11	9,7	,00	,05	,07	,06	,03	,00	,00	,11	,02	,00	,00	,02	,36
	8	,07	11,6	,00	,03	,02	,48	,00	,00	,16	,09	,03	,00	,00	,00	,01
	9	,04	15,6	,00	,00	,01	,00	,00	,00	,19	,16	,01	,00	,00	,69	,08
	10	,03	19,8	,01	,00	,00	,07	,00	,04	,56	,31	,03	,00	,01	,18	,06
	11	,01	27,7	,00	,00	,01	,23	,02	,02	,00	,01	,64	,90	,01	,00	,01
	12	,01	28,9	,00	,00	,08	,04	,01	,37	,01	,02	,16	,01	,03	,27	,03
	13	,01	37,4	,00	,00	,11	,05	,00	,26	,39	,00	,00	,00	,05	,32	,01
	14	,00	87,3	,99	,00	,00	,00	,00	,31	,53	,02	,10	,08	,01	,39	,06

a Abhängige Variable: MEDV

- colored variables may responsible for the multicollinearity

- analysis steps
 1. Compute condition index
 2. Analyze tolerances
- multicollinearity does not matter if x and $f(x)$ considered
 - ▶ you always can consider $y = ax + b$ and $f(y)$ instead, choose a and b such that $\text{cor}(y, f(y)) = 0$
 - ▶ multicollinearity aims at relationships between independent varying variables, but you will never vary x and $f(x)$ independently
- multicollinearity does not matter if you only want to make predictions
- solutions
 - ▶ exclude one or more variables
 - ▶ do a principal component analysis of the explanatory variables and use the PCs
 - ▶ restrict the range of the regression coefficients (ridge regression)

R Listing 18.5: example_multicollinearity.R

```
1 data(Boston, package="MASS")
2 x <- Boston[,-c(4,9)]
3 lm <- lm (medv~., data=x)
4 #
5 library("car")
6 vif(lm)
7 #
8 library("perturb")
9 colldiag(lm)
```

R car::vif(x)

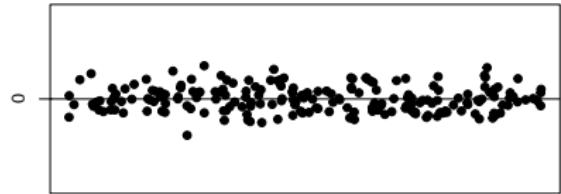
R perturb::colldiag(x)

Residuals

- Check model assumptions
 - ▶ nonconstancy of the error variance
 - ▶ nonindependency of error terms
 - ▶ nonlinearity of the regression function
 - ★ residuals vs. predictors
 - ★ residuals vs. fitted values
 - ★ predictor vs. dependent can be less informative
 - ▶ presence of outliers
 - ★ standardized residuals vs. predictors
 - ★ standardized residuals vs. fitted values
 - ★ boxplot etc.
 - ▶ nonnormality of error terms
 - ★ Q-Q plot of standardized residuals
 - ▶ omission of important predictor variables
 - ★ residuals vs. omitted/unused variables
- use graphics, coefficients and tests

Appropriate model

residuals



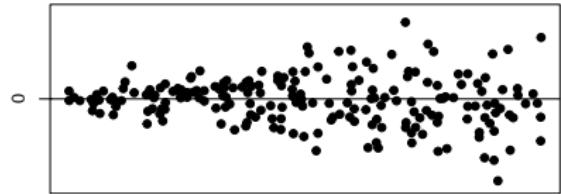
Nonlinearity of regression function

residuals



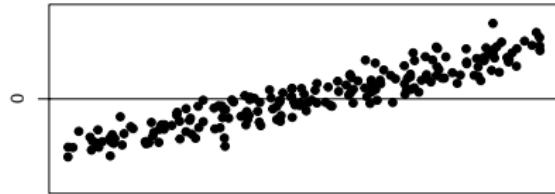
Nonconstancy of error

residuals



Nonindependence of error term

residuals



- residuals:

$$\hat{u}_i = y_i - \hat{y}_i$$

- ▶ U_i are not independent
- ▶ $\sum_i u_i = 0$ and $\sum_i x_i u_i = 0$
- ▶ if n large then dependence can be ignored

- standardized/semi-studentized:

$$r_i = \frac{\hat{u}_i}{s_u}$$

- ▶ $U_i \sim N(0, \sigma_U^2) \Rightarrow R_i \sim N(0, 1)$
- ▶ $s_u^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$ an estimator for σ_U^2

- studentized:

$$r_{i,(s)} = \frac{\hat{u}_i}{s_u \sqrt{1 - h_i}}$$

- ▶ Leverage $h_i = x_i^T (\mathfrak{X}^T \mathfrak{X})^{-1} x_i$ from $\hat{\eta} = \mathfrak{X} (\mathfrak{X}^T \mathfrak{X})^{-1} \mathfrak{X}^T \eta$
- ▶ $s_u^2(1 - h_i)$ an estimator for $\text{Var}(U_i)$

- deleted:

$$r_{i,(d)} = \frac{\hat{u}_i}{1 - h_i}$$

- studentized deleted:

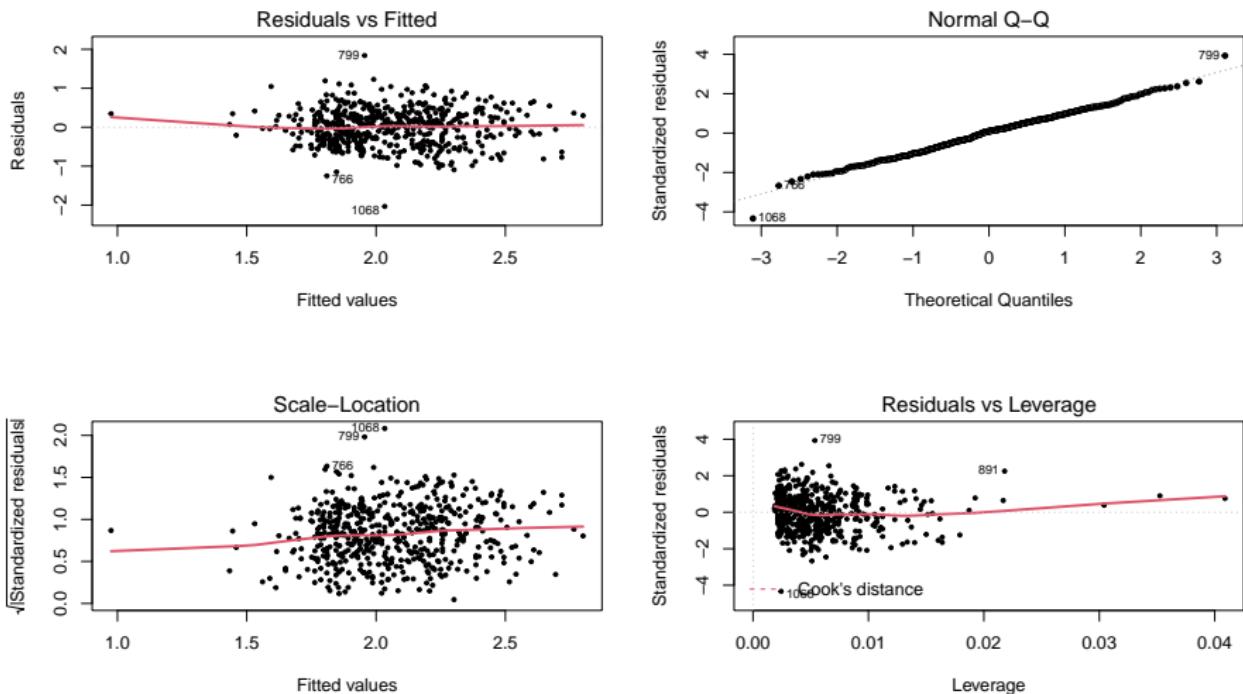
$$r_{i,(sd)} = \frac{\hat{u}_i}{s_{u(i)} \sqrt{1 - h_i}}$$

- ▶ $s_{u(i)}^2$ the estimated variance of the residuals with obs. i deleted

R Listing 18.6: regression_residuals.R

```
1 x <- read.csv2("cps78_85.csv")
2 lm <- lm(lwage~educ+exper, data=x, subset=(year==85))
3 par(mfrow=c(2,2))
4 plot(lm)
```

- R residuals(lm)
- R MASS::stdres(lm)
- R MASS::studres(lm)
- R rstandard(lm)
- R rstudent(lm)
- R plot(lm)



Tests on residuals

- Nonconstancy of error variance
 - ▶ Goldfeld-Quandt-Test: F test with two groups and regressions
 - ▶ Browne-Forsythe test (Levene test using medians rather than means)
 - ▶ White test
 - ▶ Breusch-Pagan test
- Nonnormality of residuals
 - ▶ Tests on normality: Kolmogorov-Smirnow test or others
- Nonindependency of error terms
 - ▶ Durbin-Watson test

White test

- Regress

$$\hat{u}_k^2 = a_0 + \sum_{i=1}^p a_i x_{ik} + \sum_{i \leq j} a_{ij} x_{ik} x_{jk}$$

- ▶ H_0 : all $a_{ij} = 0$ and all $a_i = 0$ ($i > 0$)
- ▶ H_1 : at least one coefficient $\neq 0$
- Test statistic under H_0 : $nR^2 \sim \chi_q^2$ with $q = p(p+3)/2$
(Lagrange multiplier test for normal errors)
- Dummy variables are not included in the cross-product part
(multicollinearity !)
- If n is not much larger than q then the cross-products $x_i x_j$ could be omitted
- Large sample test
- Less sensitive to deviation of normal errors

Breusch-Pagan test

- Idea:

$$\sigma_{U_k}^2 = \sigma^2 f \left(\beta_0 + \sum_j \beta_j \tilde{z}_{jk} \right)$$

- ▶ \tilde{z}_j are variables from which the heteroscedasticity depends
- ▶ \tilde{z}_j may include some (or all) explanatory variables

- Regress

$$\hat{u}_k^2 = a_0 + \sum_{j=1}^J a_j z_{jk}$$

- ▶ $H_0 : \text{all } a_i = 0 \ (i > 0)$
- ▶ $H_1 : \text{at least one } a_i \neq 0$
- Test statistic under $H_0 : nR^2 \sim \chi_J^2$
 (Lagrange multiplier test for normal errors)
- Large sample test
- More sensitive against violations of normal errors

 Listing 18.7: example_bptest.R

```
1 library("rio")
2 data("staedtemietenr", package="mmstat4")
3 x <- staedtemietenr[complete.cases(staedtemietenr),]
4 lm <- lm (Miete~Fläche, data=x)
5 summary(lm)
6 plot(x$Fläche, residuals(lm))
7 abline(h=0, col="red")
8 #
9 library("lmtest")
10 bptest(Miete~Fläche, data=x)
```

☞ lmtest::bptest(lm)

⚠ White test without interactions

☞ lmtest::bptest(formula, varformula=NULL, data=list())

Breusch-Pagan test, varformula must be set

Durbin-Watson test

Assumptions: $U_i = \rho U_{i-1} + \varepsilon_i$ or $\text{corr}(U_i, U_{i-1}) = \rho$

Hypotheses: $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$

Test statistics: $D_n = \frac{\sum_{i=2}^n (U_i - U_{i-1})^2}{\sum_{i=1}^n U_i^2}$, $\lim_{n \rightarrow \infty} E(D_n) = (1 - \rho)^2$

$[0, d_{L,\alpha}]$ positive autocorrelation

$(d_{L,\alpha}, d_{U,\alpha})$ uncertainty about autocorrelation

Reject H_0 : $[d_{U,\alpha}, 4 - d_{U,\alpha}]$ no autocorrelation

$(4 - d_{U,\alpha}, 4 - d_{L,\alpha})$ uncertainty about autocorrelation

$[4 - d_{L,\alpha}, 4]$ negative autocorrelation

exact values for $d_{L,\alpha}$ and $d_{U,\alpha}$ are tabulated

rule-of-thumb: reject if $[0; 1.5]$ and $[2.5; 4]$

Remark: only necessary if observations have an order,
e.g. times series or spatial data

 Listing 18.8: example_dwtest.R

```
1 data("fatalities_statlib", package="mmstat4")
2 # see http://lib.stat.cmu.edu/DASL/Stories/hwfatal.html
3 lm <- lm (US~YR, data=fatalities_statlib)
4 summary(lm)
5 plot(fatalities_statlib$YR, fatalities_statlib$US)
6 abline(lm, col="red")
7 #
8 library("car")
9 durbinWatsonTest(lm)
10 #
11 library("lmtest")
12 dwtest(lm)
13 dwtest(lm, alternative="two.sided")
```

```
④ car::durbin.watson(lm, max.lag=1, alternative=c("two.sided",
                                                 "positive", "negative"))
④ lmtest::dwtest(formula, alternative=c("greater", "two.sided",
                                         "less"))
```

Leverage

- Leverage measures how far away an observation from the center of the data is
- High-leverage points might be influential on the regression
- Hat values h_{ij} (do not depend on Y !)
 - ▶ $\hat{Y}_i = h_{1i} Y_1 + \dots + h_{ni} Y_n$
 - ▶ $h_{ii} = x_i^T (\mathfrak{X}^T \mathfrak{X})^{-1} x_i$ (from $\hat{\eta} = \mathfrak{X}(\mathfrak{X}^T \mathfrak{X})^{-1} \mathfrak{X}^T \eta$)
- Leverage $h_i = h_{ii}$
 - ▶ influence of estimating \hat{y}_i by y_i
 - ▶ $0 \leq h_i \leq 1$
 - ▶ $\sum_{i=1}^n h_i = p + 1 \Rightarrow \bar{h} = \frac{p+1}{n}$
- Rule-of-thumb: $h_i > 3(p + 1)/n$ or $h_i > 2(p + 1)/n$ points should be analyzed in detail

② Listing 18.9: example_leverage.R

```
1 data(Boston, package="MASS")
2 lm <- lm(medv~., data=Boston)
3 plot(hatvalues(lm), pch=19, main="Leverage", cex=0.5)
4 #
5 n <- nrow(Boston)
6 p <- ncol(Boston)
7 abline(h=(1:3)*(p+1)/n, col=c("black", "darkred", "red"))
```

② hatvalues(lm)

② car::leverage.plots(lm, term.name)

Cook's distance

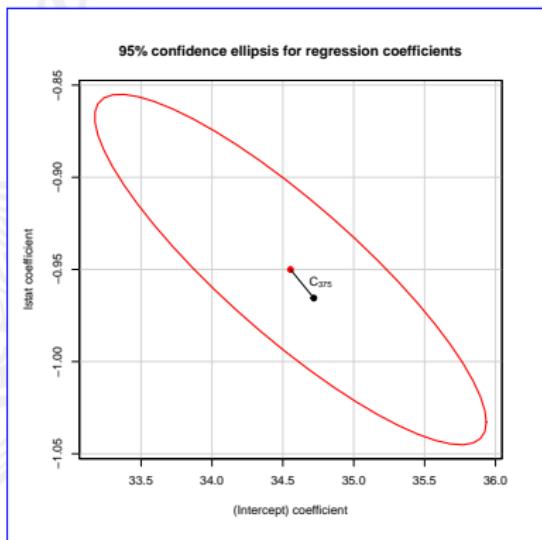
$$C_i = \frac{\sum_{k=1}^n (\hat{y}_{k;(i)} - \hat{y}_k)^2}{(n-p)s_u^2} = \frac{\hat{u}_i^2}{(n-p)s_u^2} \frac{h_i}{(1-h_i)^2}$$

- Influence of the i th observation on the fit of all other observations
- Can be computed without doing all n regressions
- Can be interpreted as the “movement” of regression coefficients when excluding the i th observation
- more important for prediction (changes in fitted values)

Cook, R. Dennis (Feb. 1977). "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19.1, p. 15. issn: 00401706. doi: 10.2307/1268249. url: <http://www.jstor.org/stable/1268249?origin=crossref> (visited on 04/28/2016).

Rule-of-thumbs

- $C_i > \frac{4}{n}$
- $C_i > F_{p;n-p;1-\alpha}$ with $\alpha = 0.5$
- $C_i > 1$ since $F_{p;n-p;0.5} \approx 1$ for large n
- values of C_i that are substantially larger than the rest



Cook, R. Dennis and Weisberg, Sanford (1982). *Residuals and influence in regression*.

Monographs on statistics and applied probability. New York: Chapman and Hall. 230 pp.
isbn: 978-0-412-24280-9.

Fox, John and Long, J. Scott, eds. (1990). *Modern methods of data analysis*. Newbury Park, Calif: Sage Publications. 446 pp. isbn: 978-0-8039-3366-8.



Listing 18.10: example_cook.R

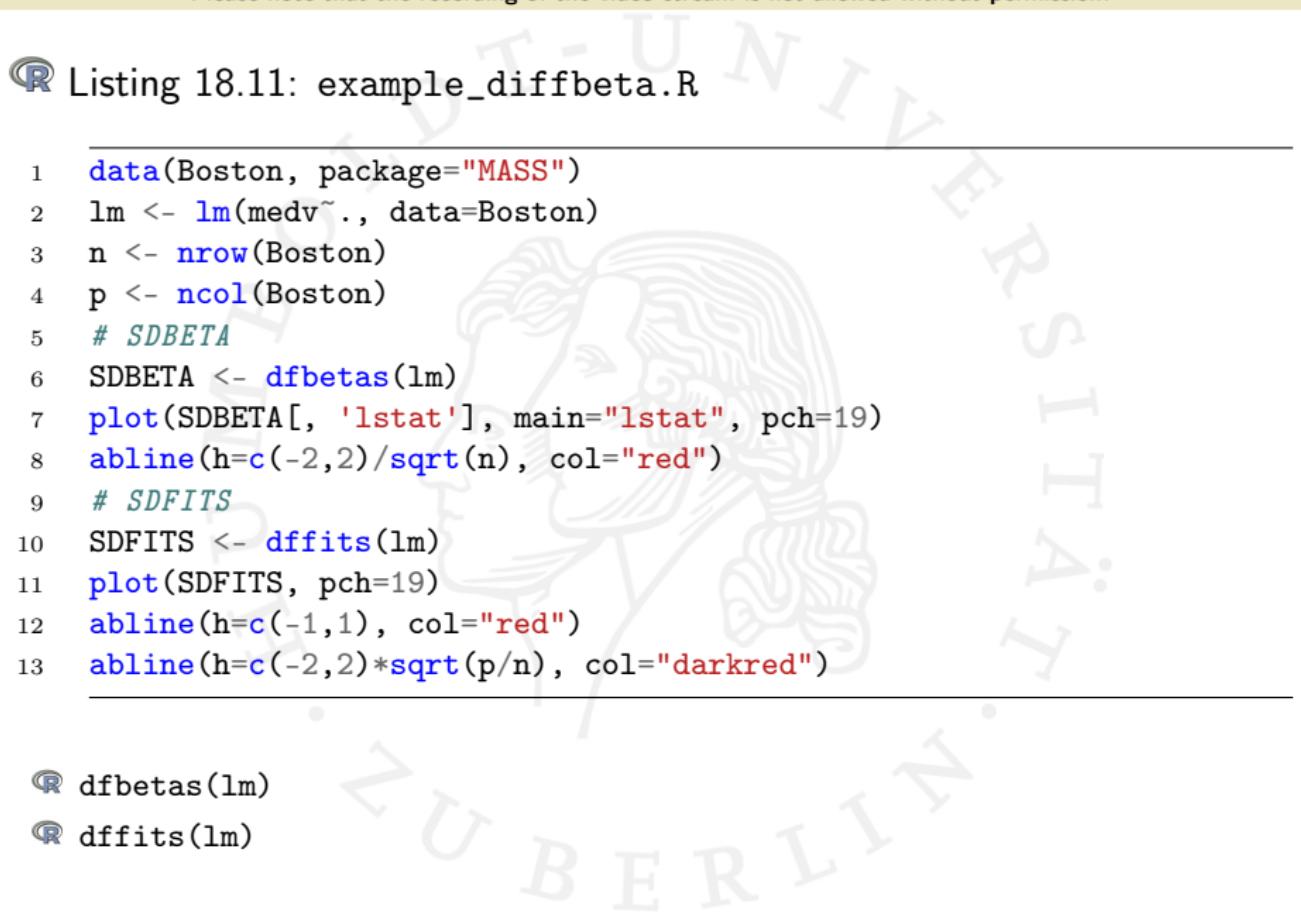
```
1 data(Boston, package="MASS")
2 lm <- lm(medv~., data=Boston)
3 plot(cooks.distance(lm), pch=19, main="Cook's distances",
4                   cex=0.5)
5 #
6 n <- nrow(Boston)
7 p <- ncol(Boston)
8 abline(h=4/n, col="red")
```



```
cooks.distance(lm)
```

Differences

- DFBETA: $\Delta\beta_{ij} = \hat{\beta}_j - \hat{\beta}_{j;(i)}$
- SDBETA: $\delta\beta_{ij} = \frac{\Delta\beta_{ij}}{s_{u;(i)}^2 \sqrt{a_{jj}}}$
 - ▶ with a_{jj} the diagonal element of $(X^T X)^{-1}$
 - ▶ Rule-of-thumb: $|\delta\beta_{ij}| > 2/\sqrt{n}$ should be analyzed
- DFBETA and DBETA are more important for explanation changes in parameter estimates
- DFFITS : $\Delta y_i = \hat{y}_i - \hat{y}_{i;(i)}$
- SDFITS: $\delta y_i = \frac{\Delta y_i}{s_{u;(i)} \sqrt{h_i}}$
- Rule-of-thumb: observations should be analyzed for
 - ▶ small & medium datasets: $\delta y_i > 1$
 - ▶ large datasets: $\delta y_i > 2\sqrt{p/n}$
- DFFITS and SDFITS are more important for prediction changes in fitted values

A faint watermark of a classical bust of a man, possibly a Greek or Roman figure, is visible in the background of the slide.

R Listing 18.11: example_diffbeta.R

```
1 data(Boston, package="MASS")
2 lm <- lm(medv~., data=Boston)
3 n <- nrow(Boston)
4 p <- ncol(Boston)
5 # SDBETA
6 SDBETA <- dfbetas(lm)
7 plot(SDBETA[, 'lstat'], main="lstat", pch=19)
8 abline(h=c(-2,2)/sqrt(n), col="red")
9 # SDFITS
10 SDFITS <- dffits(lm)
11 plot(SDFITS, pch=19)
12 abline(h=c(-1,1), col="red")
13 abline(h=c(-2,2)*sqrt(p/n), col="darkred")
```

R dfbetas(lm)
R dffits(lm)

Handling of unusual observations

- Neither ignore them, nor throw them out without thinking
- Check for data entry errors (exclusion possible)
- Think of reasons why an observation may be different
- Fit model with and without the observations to see the effect
- Collect more data
- Change the model (model misspecification?)
- Use robust regression

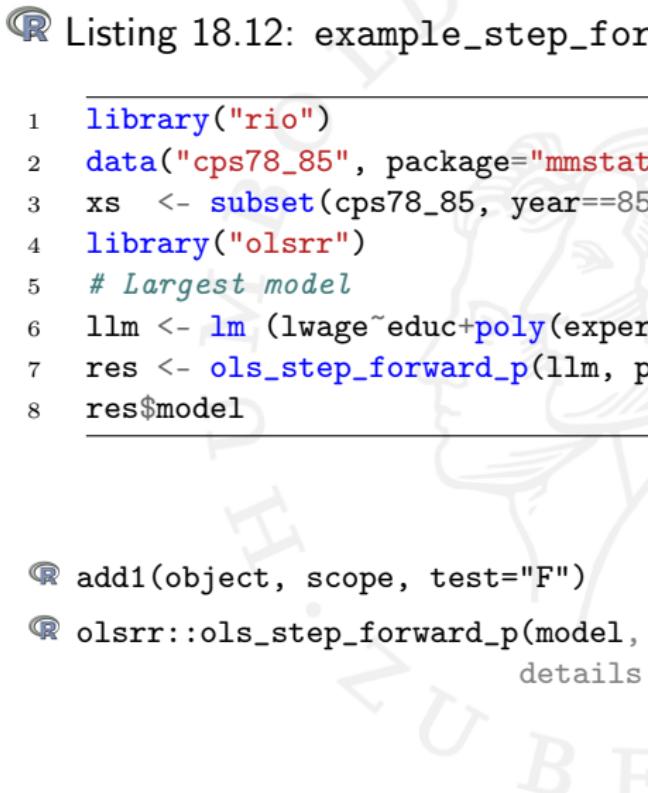
Stepwise models

- Forward regression
 1. start with the variable with the largest absolute correlation
 2. add the variable with largest absolute partial correlation until the model is good enough or no more variables available
- Backward regression
 1. start with all variables
 2. delete the variable with smallest absolute partial correlation until the model is not too bad or all variables are deleted
- Stepwise regression
 1. do a forward step
 2. do a backward step if possible
 3. stop if no variable can be deleted or added or all or no variables in the model

- Partial F-test for stepwise models

$$\frac{\frac{R_2^2 - R_1^2}{p_2 - p_1}}{\frac{1 - R_2^2}{n - p_2 - 1}} \sim F_{p_2 - p_1; n - p_2 - 1}$$

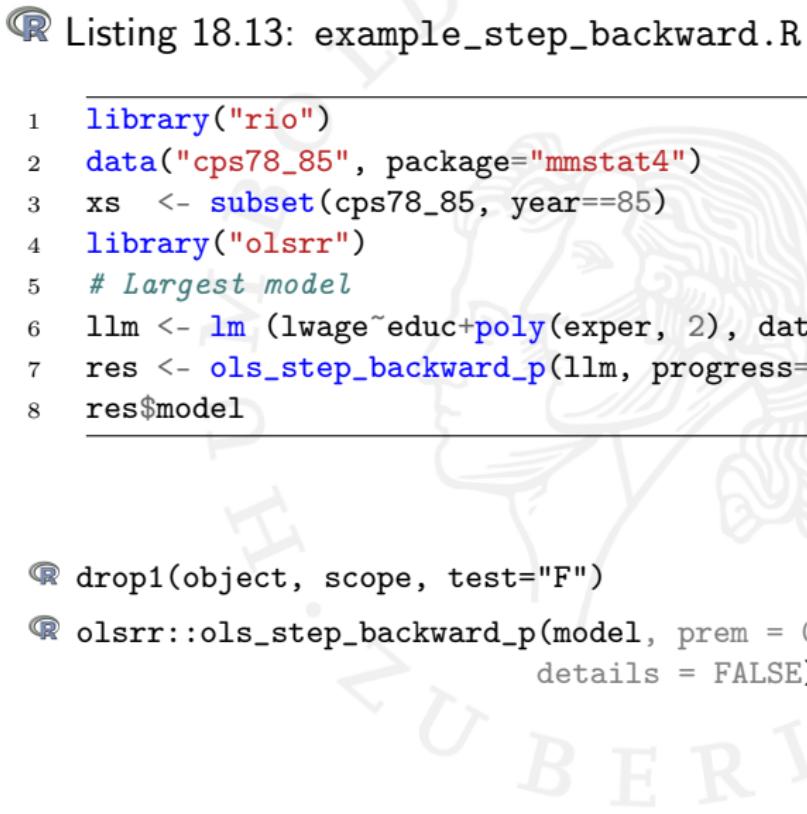
- ▶ model 1 is nested in model 2
- ▶ model i with p_i predictor variables
- ▶ test if the increase of R^2 is different to zero
 $H_0 : R_2^2 - R_1^2 = 0$ vs. $H_1 : R_2^2 - R_1^2 > 0$
- Backward regression is preferred over forward regression
 - ▶ full equation is computed
 - ▶ can handle multicollinearity better

R Listing 18.12: example_step_forward.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 library("olsrr")
5 # Largest model
6 llm <- lm (lwage~educ+poly(exper,2), data=xs)
7 res <- ols_step_forward_p(llm, progress=TRUE)
8 res$model
```

R add1(object, scope, test="F")

R olsrr::ols_step_forward_p(model, penter = 0.3, progress = FALSE,
details = FALSE)

The logo of Humboldt University Berlin is a watermark-style watermark in the background of the slide. It features a profile of a man's head facing left, with the university's name "HUMBOLDT-UNIVERSITÄT BERLIN" written in a stylized, overlapping font across the top and sides of the head.

R Listing 18.13: example_step_backward.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 library("olsrr")
5 # Largest model
6 llm <- lm (lwage~educ+poly(exper, 2), data=xs)
7 res <- ols_step_backward_p(llm, progress=TRUE)
8 res$model
```

```
R drop1(object, scope, test="F")
R olsrr::ols_step_backward_p(model, prem = 0.3, progress = FALSE,
                           details = FALSE)
```



Listing 18.14: example_step_both.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 library("olsrr")
5 # Largest model
6 llm <- lm (lwage~educ+poly(exper,2), data=xs)
7 res <- ols_step_both_p(llm, progress=TRUE)
8 res$model
```

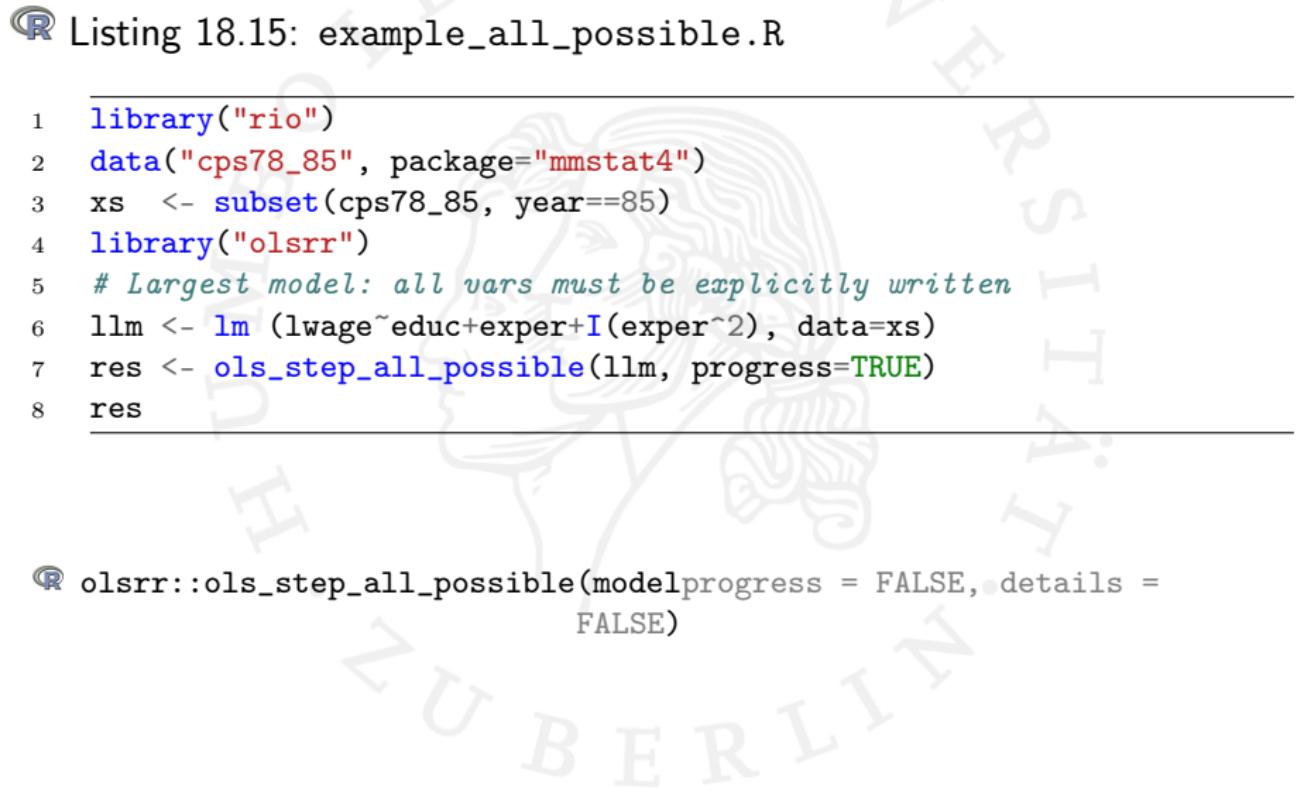
④ `olsrr::ols_step_both_p(model, pent=0.1, prem = 0.3, progress = FALSE, details = FALSE)`

Other model selection criteria

- Compute all possible models and compare
 - ▶ is feasible in small problems today
 - ▶ statistical software is not prepared to do so
- Mallow's C_p analyses the mean squared prediction error in relation to residual variance σ of the full model

$$\frac{1}{\sigma^2} E \left(\sum_i (\hat{Y}_i - \mu_i)^2 \right) \text{ estimated by } \hat{C}_p = \frac{\text{ssr}_p}{\hat{\sigma}} + (2p - n)$$

- ▶ ssr_p sum of squared residuals of model with p predictors
- ▶ $\hat{\sigma}$ variance of residuals in full model
- ▶ $E(C_p) = p + d$ with $d = 0$ if estimation unbiased otherwise $d > 0$
- ▶ create scatter plot with (p, \hat{C}_p)
- ▶ choose a subset that has \hat{C}_p approaching p from above
- ▶ problem: estimation requires a large n

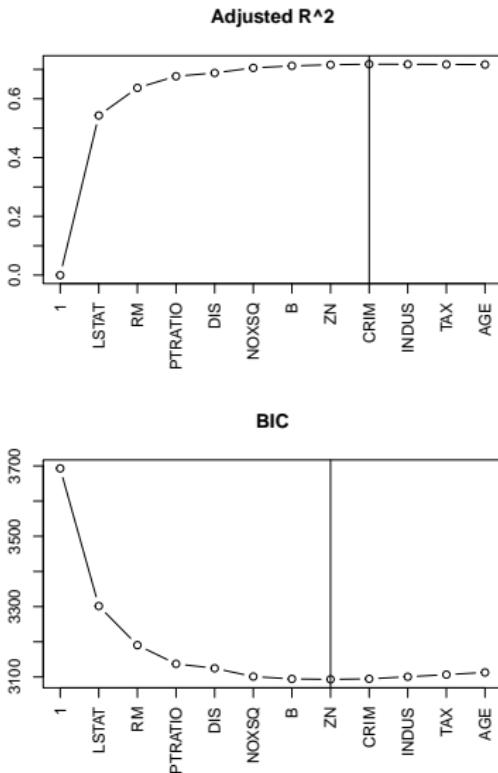
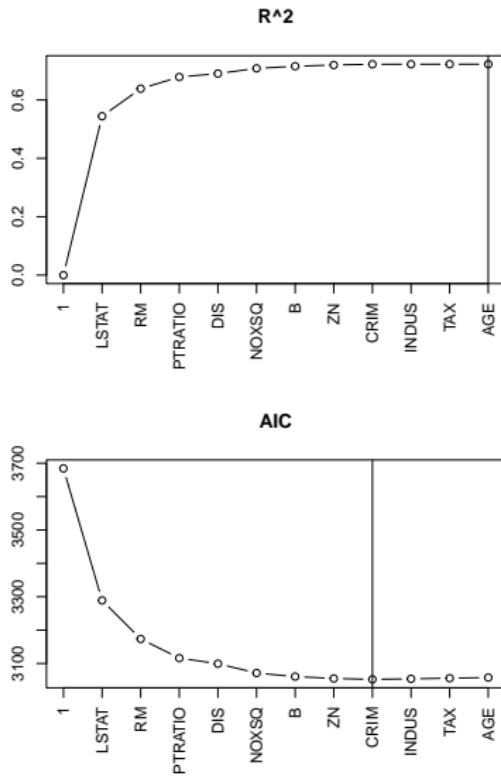


R Listing 18.15: example_all_possible.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 library("olsrr")
5 # Largest model: all vars must be explicitly written
6 llm <- lm (lwage~educ+exper+I(exper^2), data=xs)
7 res <- ols_step_all_possible(llm, progress=TRUE)
8 res
```

R olsrr::ols_step_all_possible(modelprogress = FALSE, details = FALSE)

- balance model complexity vs. model explanatory power
- Occams razor
 - ▶ if you have two theories which explain the same facts then keep the simpler one
- Akaike information criteria (predictive model)
 - ▶ $AIC = 2k - 2 \log(\text{max-likelihood})$
 - ▶ $AIC = 2k + n \log \left(\frac{\text{residual sum of squares}}{n} \right)$
(model errors are normal)
- Bayesian information criterion (true model)
 - ▶ $BIC = \log(n)k - 2 \log(\text{max-likelihood})$
 - ▶ $BIC = \log(n)k + n \log \left(\frac{\text{residual sum of squares}}{n} \right)$
(model errors are normal)
- k number of free parameters, n number of observations



Model criteria for linear regression for Boston Housing data

 Listing 18.16: example_step_ic.R

```
1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 library("olsrr")
5 # Largest model: all vars must be explicitly written
6 llm <- lm (lwage~educ+exper+I(exper^2), data=xs)
7 ols_step_forward_aic(llm)
8 ols_step_backwardward_aic(llm)
9 ols_step_both_aic(llm)
```

- ④ olsrr::ols_step_forward_aic(modelprogress = FALSE, details = FALSE)
- ④ olsrr::ols_step_backward_aic(modelprogress = FALSE, details = FALSE)
- ④ olsrr::ols_step_both_aic(modelprogress = FALSE, details = FALSE)

 Listing 18.17: example_step_aic.R

```

1 library("rio")
2 data("cps78_85", package="mmstat4")
3 xs <- subset(cps78_85, year==85)
4 # Add exper^2 ?
5 lms <- lm(lwage~educ+exper, data=xs)
6 add1(lms, ~.+I(exper^2))
7 lms <- lm(lwage~educ+south+nonwhite+female+married+exper+union+expersq
8 data=xs)
9 drop1(lms) # Drop one variable
10 step(lms) # Automatic backward

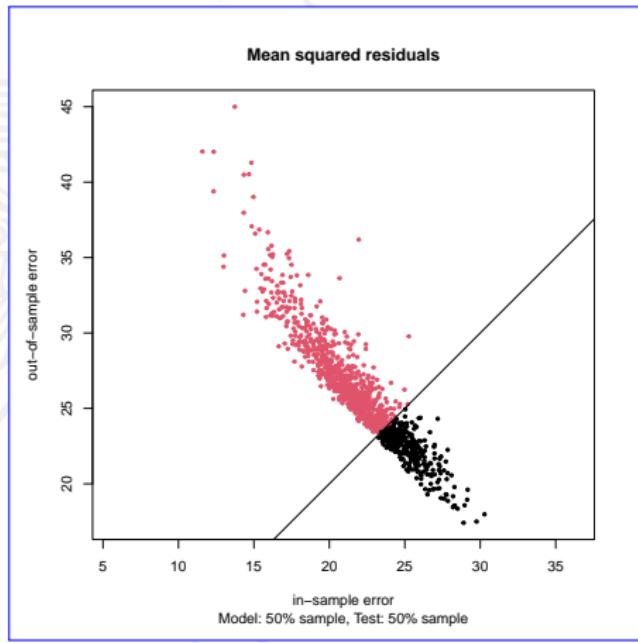
```

-  `add1` (object, scope, k=2)
-  `drop1` (object, scope, k=2)
-  `step` (object, scope, direction = c("both", "backward",
 "forward"), k=2)
-  `MASS:::stepAIC(...)`

Differences between `step` and `stepAIC` are marginal, AIC: $k=2$, BIC: $k=\log(n)$

Disadvantage of stepwise modeling

- Stepwise modeling analyzes a lot of possible candidate models
- By chance uninformative variables may enter the model
- Prone to overfitting
- Compare in-sample and out-of-sample error
 - ▶ above 45 degree line: out-of-sample error > in-sample error
 - ▶ below 45 degree line: out-of-sample error < in-sample error

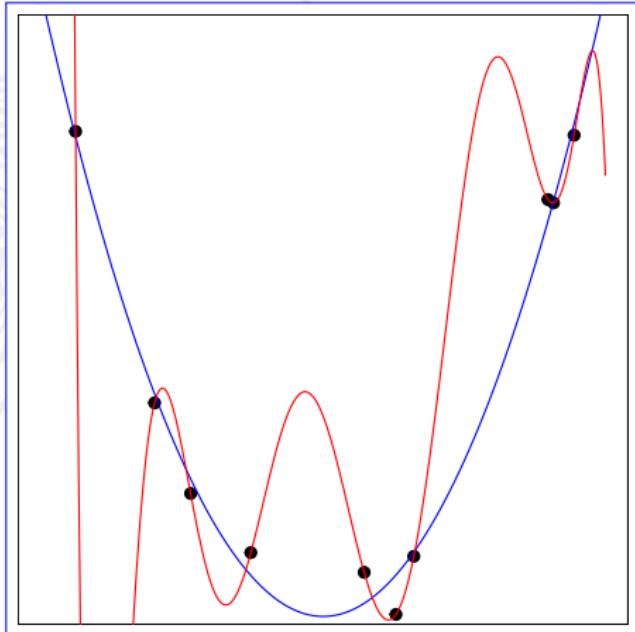


Model error

- Linear regression: model error is based on residuals $y_i - \hat{y}_i$
- R^2 is an in-sample-error: the model is built and assessed with the same data

⚠ The model may not generalize well

- ▶ the red and blue model fit the data points well
 - ▶ the red model generalizes badly
- Consider the out-of-sample error with “fresh” data

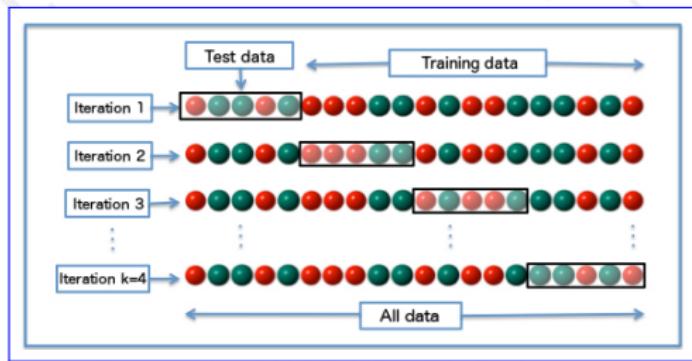


- Blue: degree 2 polynomial
- Red: degree 9 polynomial

- Split data in two parts
 - ▶ one to build the model (training data)
 - ▶ one to estimate the error (test data)
- Pareto principle: 80% training, 20% test
- Rule-of-thumb:

$$\frac{\#test}{\#training} \approx \frac{1}{\sqrt{\text{free adjustable parameters}}}$$

- small data sets \Rightarrow use k -fold crossvalidation



- ▶ split data set into k parts
- ▶ use $k - 1$ parts (training data) to build the model
- ▶ evaluate the model on the test data
- ▶ repeat this for all k parts and average the error
- Rule-of-thumb: $K = 10$
- Special case: Leave-one-out crossvalidation ($k = n$)

Shrinkage methods

- Ridge regression

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \hat{\beta}_j^2 \rightarrow \min$$

- coefficients shrinked towards zero $\hat{\beta}_{j,\lambda}^{ridge} = \frac{\hat{\beta}_j^{OLS}}{1+\lambda}$ (if X_i orthonormal)
- Properties
 - ▶ $\lambda = 0$ one gets the OLS estimates
 - ▶ introduces bias, but reduces variance ($MSE = bias^2 + variance$)
 - ▶ it exists always a λ such that $MSE(\hat{\beta}^{ridge}) < MSE(\hat{\beta}^{OLS})$
 - ▶ is suited to deal with ill-posed problems resulting from highly correlated features (multicollinearity)
 - ▶ intended for prediction, not interpretation

- Choice of λ
 - ▶ ridge traces
 - ▶ k -fold crossvalidation
- Ridge trace
 - ▶ plot $\hat{\beta}_{j,\lambda}^{\text{ridge}}$'s against λ
 - ▶ the faster a coefficient is shrinking the less important

Hoerl, Arthur E. and Kennard, Robert W. (Feb. 1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1, pp. 55–67. issn: 0040-1706, 1537-2723.

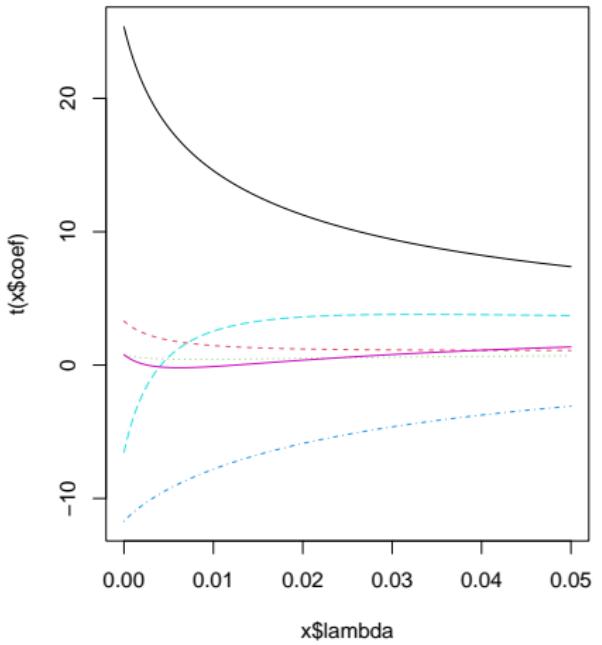
doi: 10.1080/00401706.1970.10488634. url:

<http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634> (visited on 11/30/2016).

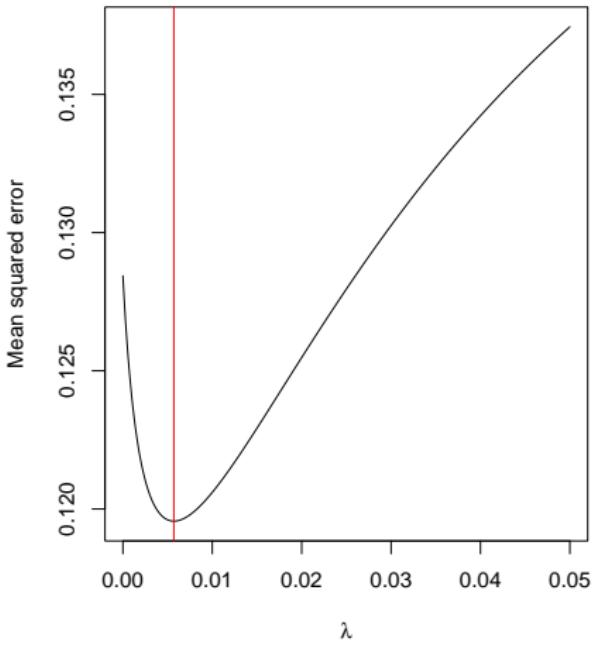
Golub, Gene H., Heath, Michael, and Wahba, Grace (May 1979). "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* 21.2, pp. 215–223. issn: 0040-1706, 1537-2723. doi: 10.1080/00401706.1979.10489751. url:

<http://www.tandfonline.com/doi/abs/10.1080/00401706.1979.10489751> (visited on 11/30/2016).

Ridge traces



Generalized crossvalidation



 Listing 18.18: example_ridge.R

```
1 library("MASS")
2 bhd <- Boston[,-9]
3 # find lambda
4 lmridge <- lm.ridge(medv~., data=bhd, lambda=seq(0, 10, 0.01))
5 select(lmridge)
6 # ridge traces
7 plot(lmridge)
8 # compare coefficients
9 lm(medv~., data=bhd)
10 lm.ridge(medv~., data=bhd, lambda=8.98)
```

② MASS::lm.ridge(formula, data, lambda=0)

② MASS::select(obj)

- Least Absolute Shrinkage and Selection Operator (LASSO)

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j| \rightarrow \min$$

► $\hat{\beta}_j^{lasso} = \hat{\beta}_j^{OLS} \max \left(0, 1 - \frac{\lambda}{|\hat{\beta}_j^{OLS}|} \right)$ (if X_i orthonormal)

- Properties

- $\lambda = 0$ one gets the OLS estimates
- $\lambda > |\hat{\beta}_j^{OLS}| \Rightarrow \beta_j^{lasso} = 0$
- creates sparse solutions (variable selection)
- tends to select the feature most strongly correlated with the outcome and zero out the rest

- Choice of λ by k -fold crossvalidation

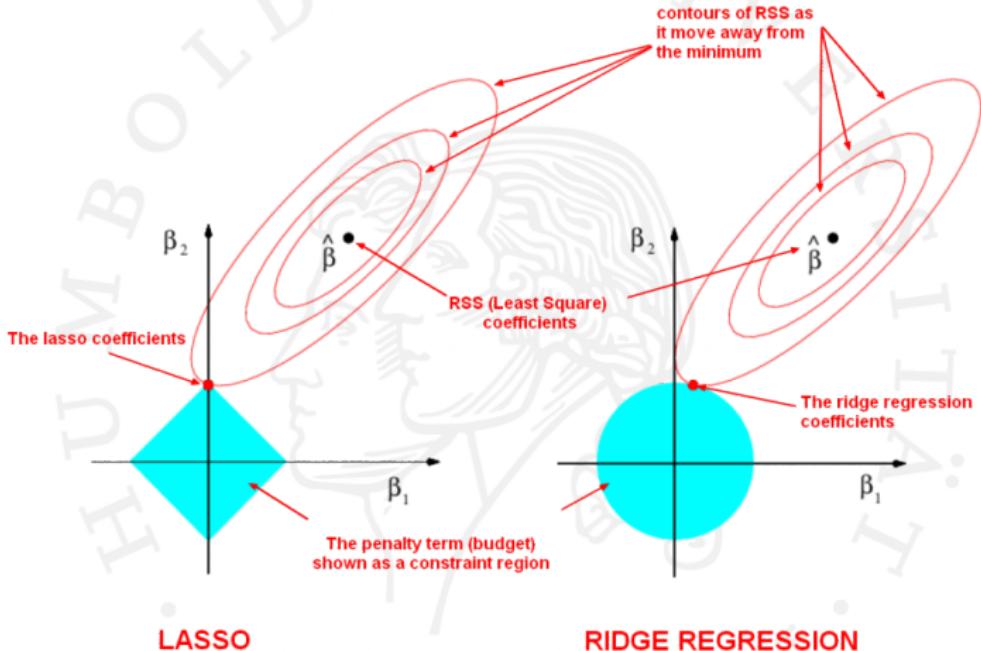
Tibshirani, Robert (1994). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.

 Listing 18.19: example_lasso.R

```
1  data(Boston, package="MASS")
2  library("glmnet")
3  x <- as.matrix(Boston[,-c(9,14)])
4  y <- Boston[,14]
5  # find lambda
6  lmlasso <- cv.glmnet(x, y)
7  plot(lmlasso)
8  #
9  lmlasso <- cv.glmnet(x, y, lambda=seq(0, 1, 0.001))
10 plot(lmlasso)
11 lmlasso$lambda.min
12 lmlasso$lambda.1se
13 # compare coefficients
14 lm(y~x)
15 coef(lmlasso, s="lambda.1se")
```

⑧ `glmnet::cv.glmnet(x,y, lambda, alpha=1)`

⑧ `glmnet::glmnet(x,y, lambda, alpha=1)`



Source: gerardnico.com/wiki/data_mining/lasso

- Elastic net

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum_{j=1}^p |\hat{\beta}_j| + (1-\alpha) \sum_{j=1}^p \hat{\beta}_j^2 \right) \rightarrow \min$$

► $\hat{\beta}_j^{elastic} = \frac{\hat{\beta}_j^{OLS}}{1+\lambda(1-\alpha)} \max \left(0, 1 - \frac{\lambda\alpha}{|\hat{\beta}_j^{OLS}|} \right)$ (if X_i orthonormal)

- combines advantages of Ridge and LASSO regression
- convex combination has better MSE than with LASSO or Ridge
- Choice of parameters
 - λ by k -fold crossvalidation
 - $\alpha \in [0, 1]$ by grid search

Zou, Hui and Hastie, Trevor (Apr. 2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. issn: 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. url: <http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x> (visited on 11/30/2016).

 Listing 18.20: example_elasticnet.R

```

1  data(Boston, package="MASS")
2  library("glmnet")
3  x <- as.matrix(Boston[,-c(9,14)])
4  y <- Boston[,14]
5  # find lambda
6  lmenet<- cv.glmnet(x, y, alpha=0.5)
7  plot(lmenet)
8  lmenet$lambda.min
9  lmenet$lambda.1se
10 # compare coefficients
11 lm(y~x)
12 coef(lmenet, s="lambda.1se")

```

② `glmnet::cv.glmnet(x,y, lambda, alpha)`

② `glmnet::glmnet(x,y, lambda, alpha)`

Minimizes $\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p \beta_j^2 \right)$

Comparing linear regression & LASSO/elastic net

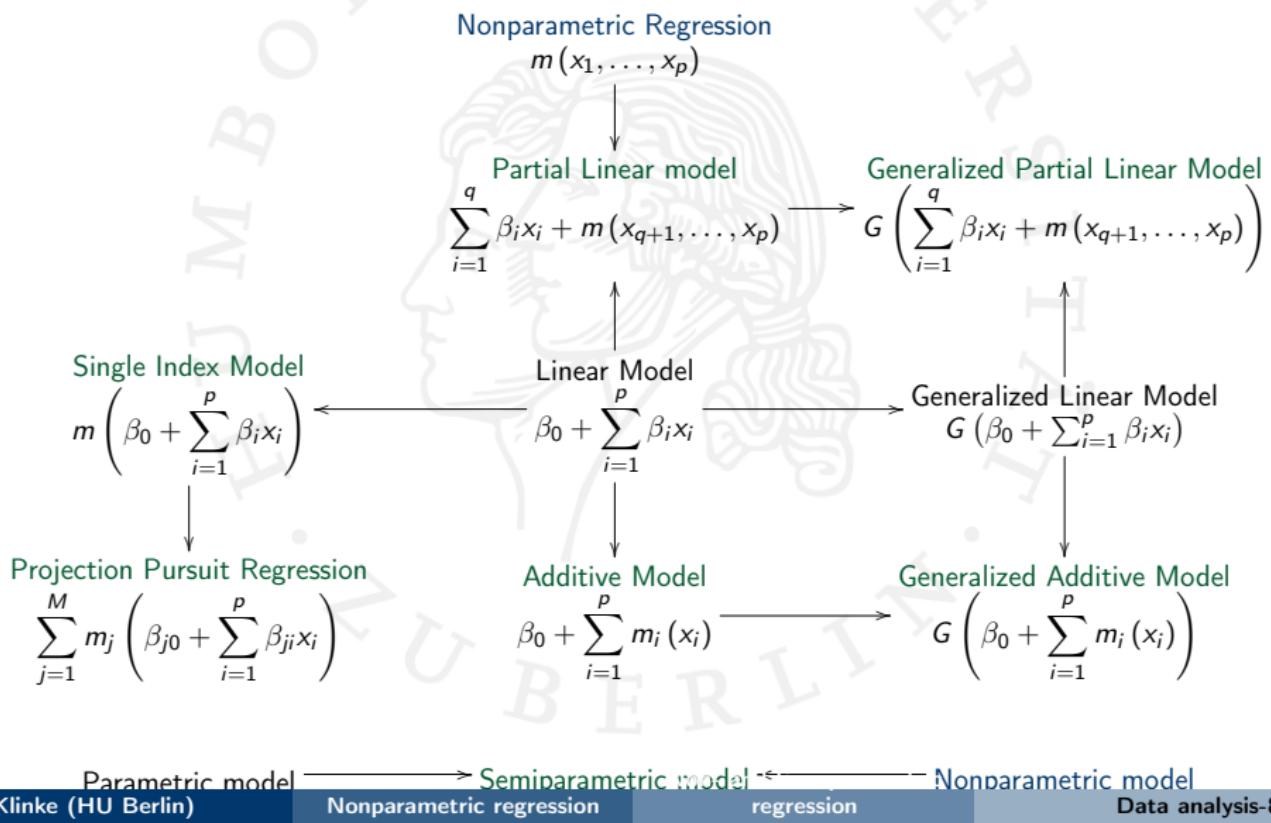
- R^2 is not a good measure to compare results of linear regression and LASSO/elastic net
- R^2 tends to choose models with a larger number of variables
 - ▶ R^2 can be increased by adding nonsense variables
- linear regression model is a LASSO with no penalty
- use BIC or a cross-validated R^2

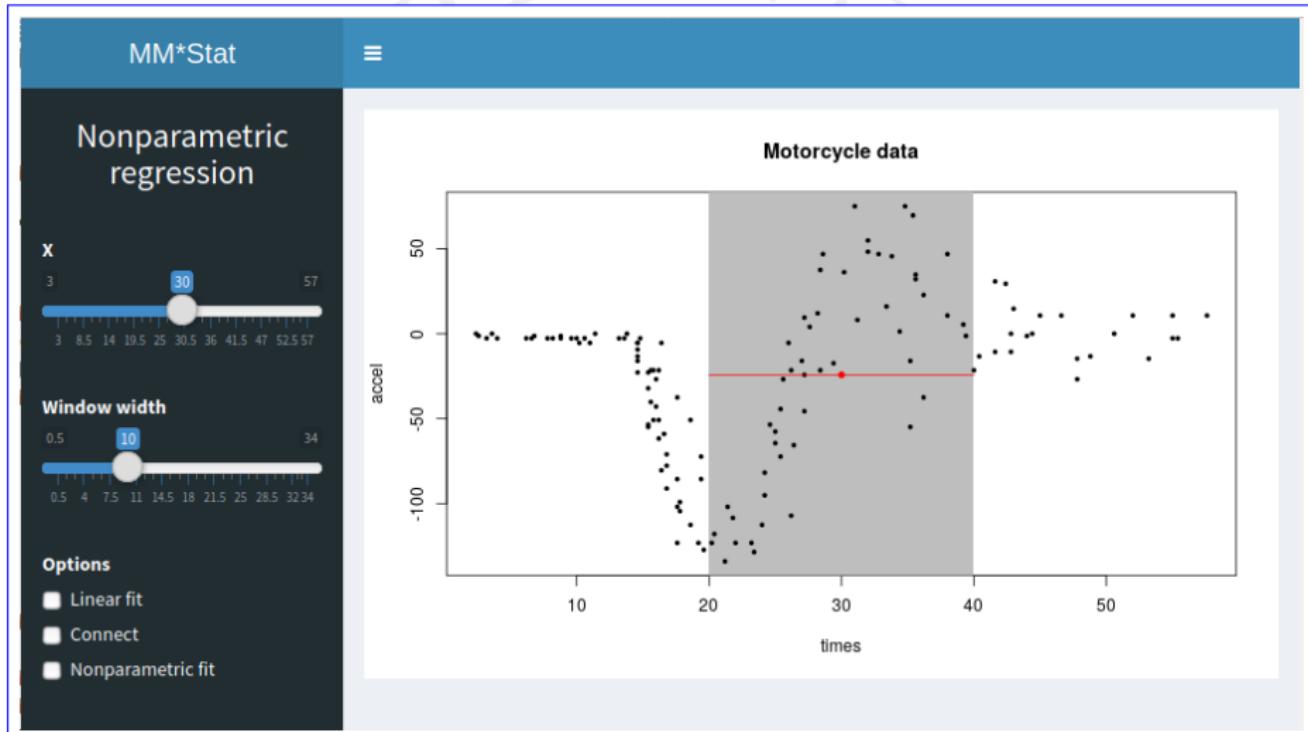
Nonparametric regression

November 3, 2022

- Non- and Semiparametric regression • Kernel density estimator •
- Nadaraya-Watson estimator • Linear regression • Partial linear model •
- Bivariate Nadaraya-Watson estimator • Additive model • Single Index Model •
- Projection Pursuit Regression • Overview

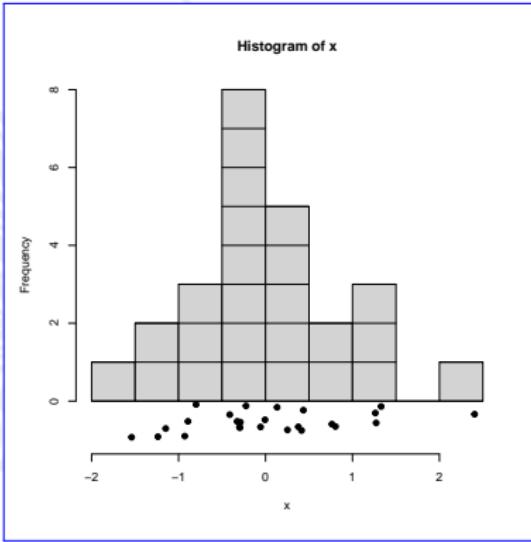
Non- and Semiparametric regression





	regression model		
	parametric	semi-parametric	non-parametric
determined by parameters	completely	to a minor degree	no
prior assumptions about functional form	yes	to a minor degree	no
inclusion of categorical variables	yes	yes	no
curse of dimensionality	unproblematic	unproblematic	problematic
inclusion of further variables	unproblematic	unproblematic	problematic
statistical precision	high	large	low
interpretation	easy	simple	difficult

- Use in estimating
 - ▶ an unknown density function → Kernel density estimator
 - ▶ an unknown regression function → Nadaraya-Watson estimator
- Idea (as in histogram)
 - ▶ replace each observations by a “kernel”
 - ▶ sum up “kernels”
- can easily be extended to the multivariate case

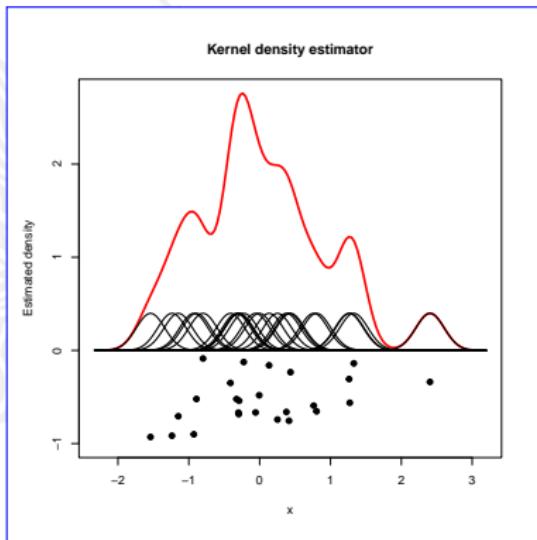


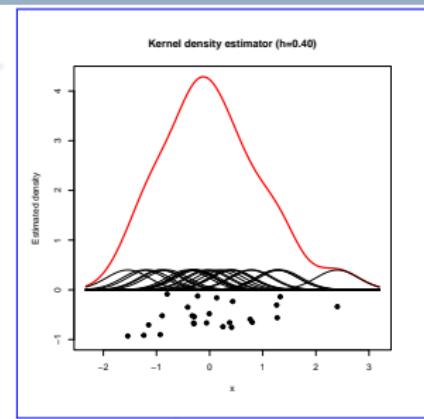
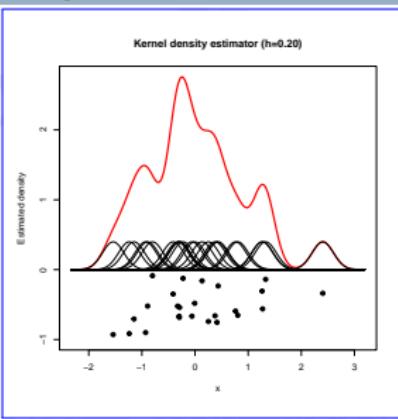
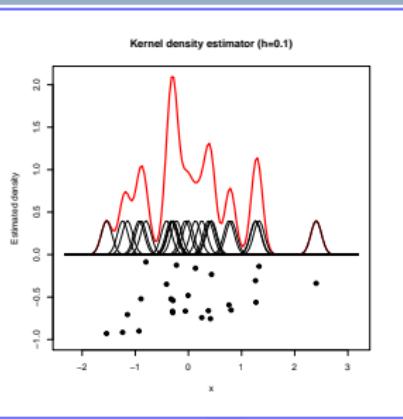
Kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

- replace each observation by a kernel function
- h , the so called *bandwidth*, determines the “width” of the kernel
- kernels must fulfill the properties

1. $K(u) \geq 0$ for all $u \Rightarrow \hat{f}_h(x) \geq 0$
2. $\int_{-\infty}^{\infty} K(u)du = 1 \Rightarrow \int_{-\infty}^{\infty} \hat{f}_h(x) = 1$





$$Bias \left(\hat{f}_h(x) \right) \approx \frac{h^2}{2} f''(x) \underbrace{\int u^2 K(u) du}_{=\mu_2(K)}$$

$$Var \left(\hat{f}_h(x) \right) \approx \frac{1}{nh} \underbrace{\int K^2(u) du}_{=|K|_2^2}$$

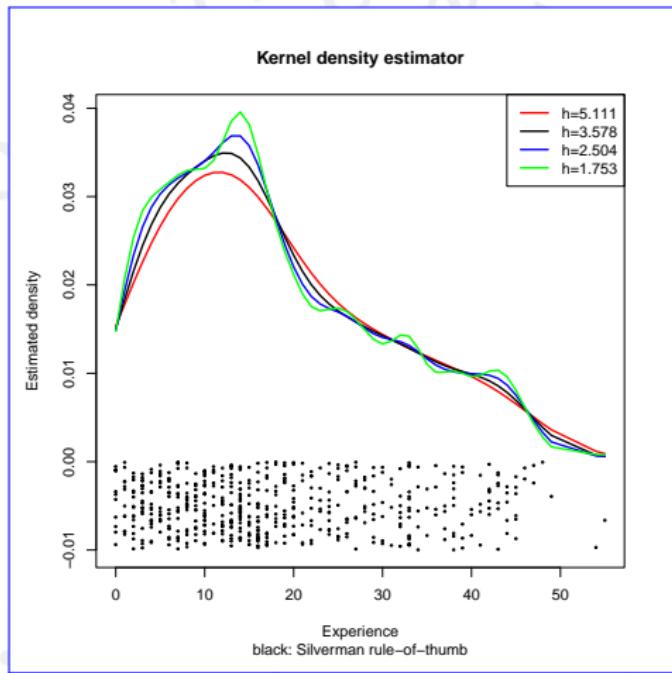
Approximate Mean Integrated Square Error:

$$AMISE\left(\hat{f}_h(x)\right) \approx \int \underbrace{\left(\frac{h^4}{4} f''(x)^2 \mu_2^2(K) + \frac{1}{nh} |K|_2^2 f(x)\right)}_{\approx MSE(\hat{f}_h(x))} dx$$

- AMISE optimal bandwidth: $h_{min} = C(K, f)n^{-1/5}$
- Silverman derived a rule of thumb by choosing
 - ▶ K is the gaussian kernel
 - ▶ f the unknown density is gaussian $N(\mu; \sigma^2)$
- Leave-one-out-crossvalidation
 - ▶ Minimize the integrated squared error with respect to h

$$ISE\left(\hat{f}_h(x)\right) = \int \left(\hat{f}_h(x) - f(x)\right)^2 dx = \int \left(\hat{f}_h^2(x) - 2\hat{f}_h f(x) + f^2(x)\right) dx$$

- ▶ To estimate $\int \hat{f}_h(x) f(x) dx$ use $\hat{f}_{h;-i}(x)$ (i th observation excluded)



```
R density(x, bw, adjust, kernel=c("gaussian",...))
```

```
R np::npudens(x, bws, bwmethod=c("normal-reference",...),  
              ckertype=c("gaussian",...))
```

Common bandwidth choices:

- Silverman: $\hat{h} = 0.9 \min(\hat{\sigma}, IQR/1.34) n^{-1/5}$ (`bw="nrd0"`)
- Scott: $\hat{h} = 1.06\hat{\sigma} n^{-1/5}$ (`bw="nrd"`)
- Unbiased crossvalidation (`bw="ucv"`)
- Biased crossvalidation (`bw="bcv"`)
- Sheather-Jones use pilot estimators for the derivatives (recommended, `bw="SJ"`)

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Sheather, Simon J. and Jones, M. Chris (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, pp. 683–690.

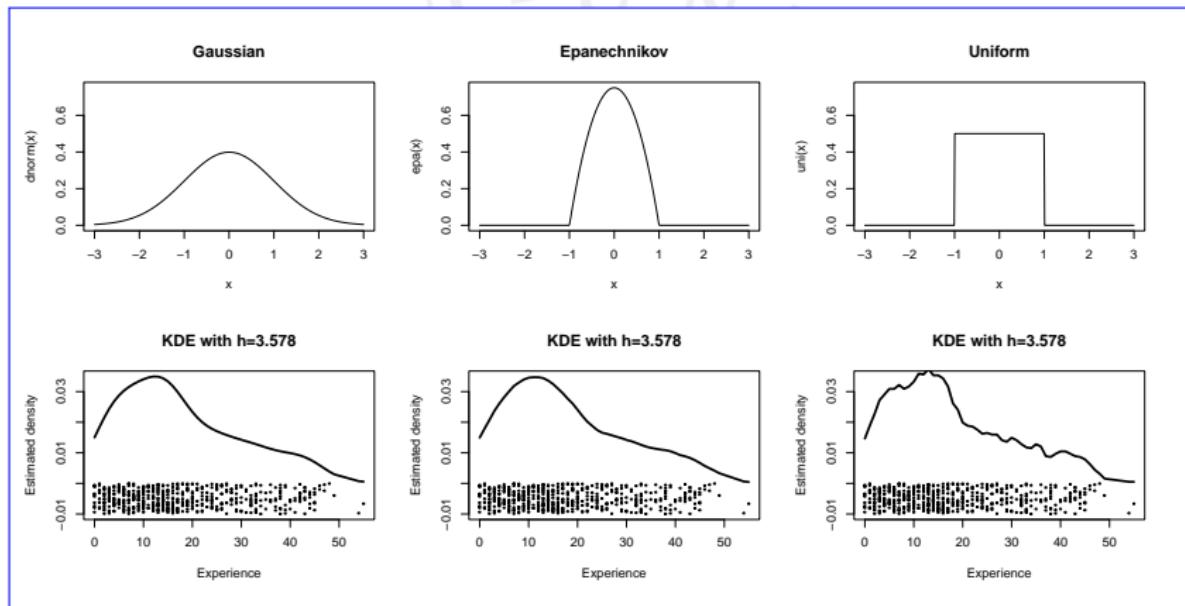
Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Vol. 275. John Wiley & Sons.

 Listing 19.1: example_kernel1.R

```
1 data(Boston, package="MASS")
2 library("np")
3 bw <- npudensbw(~medv, data=Boston)
4 fhat <- npudens(bw)
5 fhat
6 plot(fhat, main=sprintf("%s with h=%.2f", fhat$pckertype, fhat$bw))
7 rug(Boston$medv)
```

 Listing 19.2: example_kernel2.R

```
1 library("MASS") # for Boston housing data
2 library("np")
3 bw <- npudensbw(~medv+lstat, data=Boston,
                  bwmethod="normal-reference")
4 fhat <- npudens(bw)
5 plot(fhat, view="fixed", theta=60, phi=30)
```



- Smoothness properties of kernel are inherited by kernel density estimator
- Choice of kernel can be balanced by bandwidth (\rightarrow equivalent kernels/bandwidths)

- Bivariate kernel density estimator

$$\hat{f}_{h_1, h_2}(x_1, x_2) = \frac{1}{nh_1 h_2} \sum_{i=1}^n K\left(\frac{x_1 - x_{1i}}{h_1}, \frac{x_2 - x_{2i}}{h_2}\right)$$

- ▶ e.g. with $K(x_1, x_2) = K(x_1)K(x_2)$
- Multivariate kernel density estimator

$$\hat{f}_{h_1, \dots, h_p}(x_1, \dots, x_p) = \frac{1}{nh_1 \dots h_p} \sum_{i=1}^n K\left(\frac{x_1 - x_{1i}}{h_1}, \dots, \frac{x_p - x_{pi}}{h_p}\right)$$

- Curse of dimensionality: estimate $\varphi(0, \dots, 0)$ for a given accuracy

Dimension	1	2	3	4	5	6	7	8	9	10
Sample size	4	19	67	223	768	2.790	10.700	43.700	187.000	842.000

Nadaraya-Watson estimator

- Regression function $m(x)$ relates an independent variable X with a dependent variable Y
- It holds $E(Y|X = x) = m(x)$

$$E(Y|X = x) = \int_{-\infty}^{\infty} yf(y|x)dy = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f(x)} dy$$

- ▶ $f(x, y)$ the joint density
- ▶ $f(x)$ the marginal density of X
- Estimate both densities with kernel densities

$$\hat{f}_{h,g}(x, y) = \frac{1}{nhg} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{g}\right)$$

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

- Nadaraya-Watson estimator

$$\hat{m}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}$$

- Linear estimator in Y (like in linear regression)

$$\hat{m}(x) = \sum_{i=1}^n y_i \underbrace{\frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}}_{=W_{hi}(x)}$$

- ▶ with $\sum_i W_{hi}(x) = 1$

- Problem 1: Choice of bandwidth
 - ▶ Small $h \rightarrow$ “wiggly” estimates (undersmoothing/overfitting)
 - ▶ Large $h \rightarrow$ very “smooth” estimates (oversmoothing/underfitting)
 - ▶ Use leave-one-out-crossvalidation to find “optimal” h

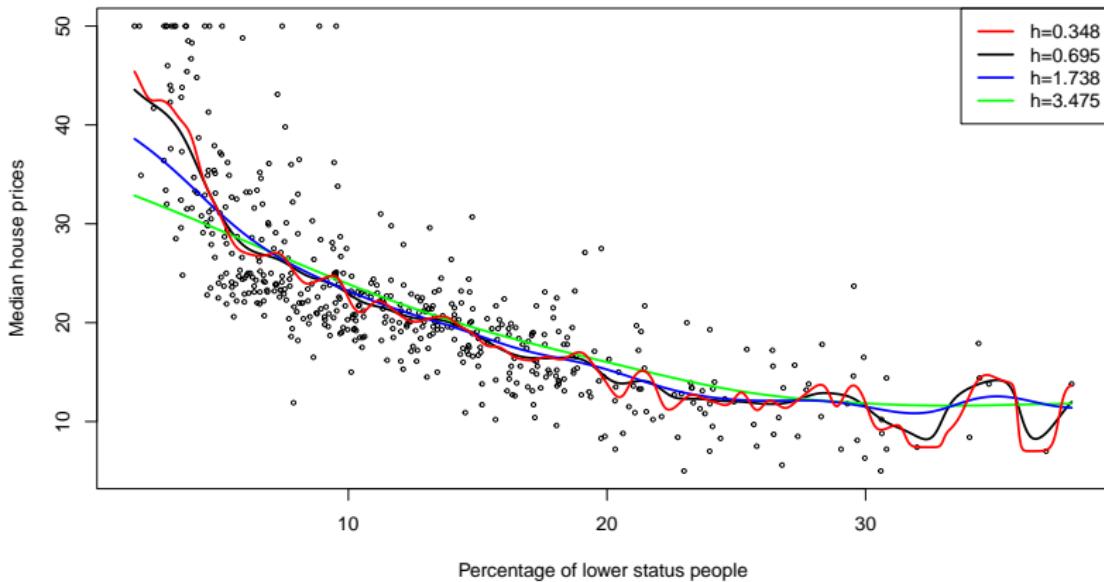
$$CV(h) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{m}_{h,-i}(x_i))^2$$

- Problem 2: Sparse data region
 - ▶ if $\hat{f}(x) = 0$ then $\hat{m}(x)$ is undefined
- Alternative non-parametric regression techniques:
 - ▶ Local polynomial estimation, e.g. lo(w)ess
 - ▶ Smoothing splines
 - ▶ k -nearest-neighbour regression

⌚ `lowess(x, y, f=2/3)`

⌚ `loess(formula, data, span=0.75)`

⌚ `smooth.spline(x, y)`



```
R np::npreg(formula, bws)
```

```
R locfit::locfit(formula)
```

R Listing 19.3: example_nw1.R

```

1 library("MASS")  # for Boston Housing data
2 library("np")
3 bw <- npregbw(medv~lstat, data=Boston)
4 mhat <- npreg(bw)
5 main <- sprintf("%s with h=%.2f", mhat$pckertype, mhat$bw)
6 plot(Boston$lstat, Boston$medv, pch=19, cex=0.3, main=main)
7 ind <- order(Boston$lstat)
8 xs  <- cbind(Boston$lstat, fitted(mhat))[ind,]
9 lines(xs[,1], xs[,2], col="red", lwd=2)
10 rug(Boston$lstat)

```

R Listing 19.4: example_nw2.R

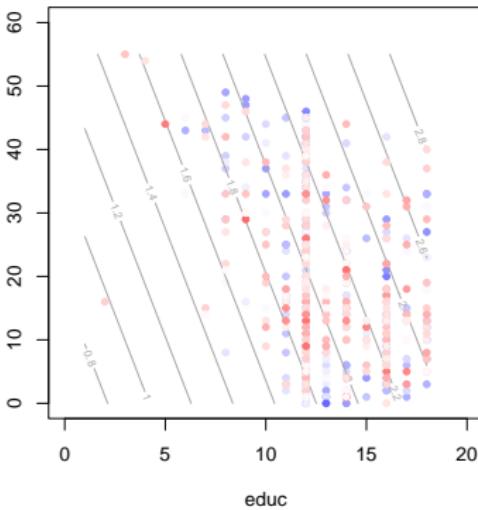
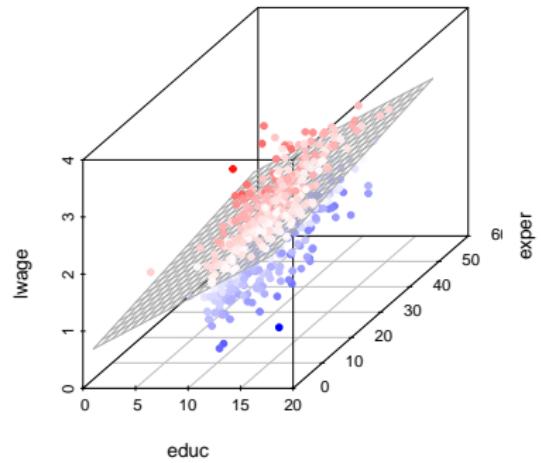
```

1 library("MASS")  # for Boston Housing data
2 library("np")
3 bw <- npregbw(medv~lstat+rm, data=Boston)
4 mhat <- npreg(bw)
5 plot(mhat, view="fixed", phi=30, theta=75)
6 plot(Boston$lstat, Boston$rm, pch=19, cex=0.5)

```

Linear regression

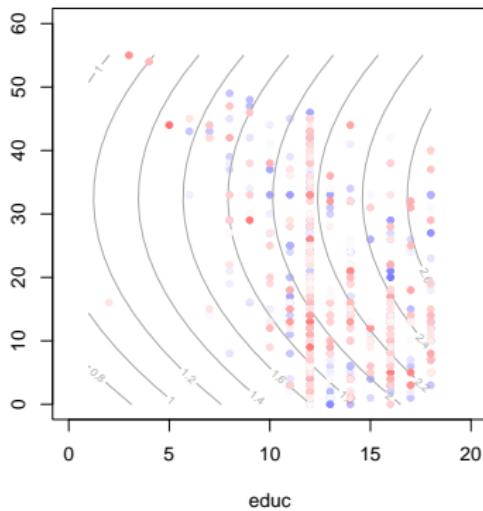
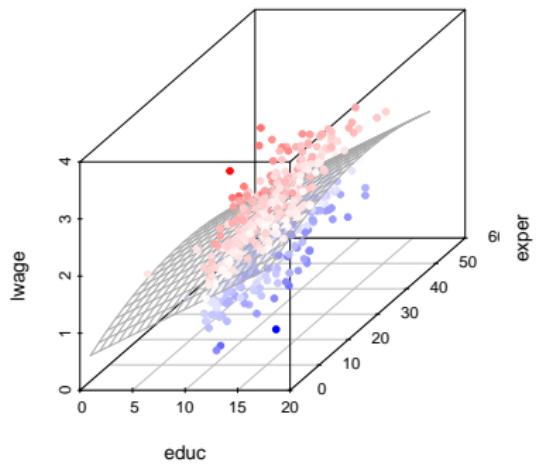
Linear Regression Model ($R^2=0.212$)



`lm(formula = l wage ~ educ + exper, data = x)`

$$\hat{y} = 0,594 + 0,096 \text{ educ} + 0,012 \text{ exper}$$

Quadratic Regression Model ($R^2=0.238$)



lm(formula = lwage ~ educ + exper + I(exper^2), data = x)

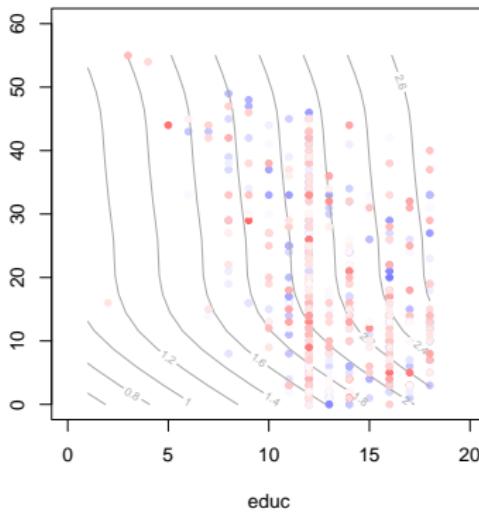
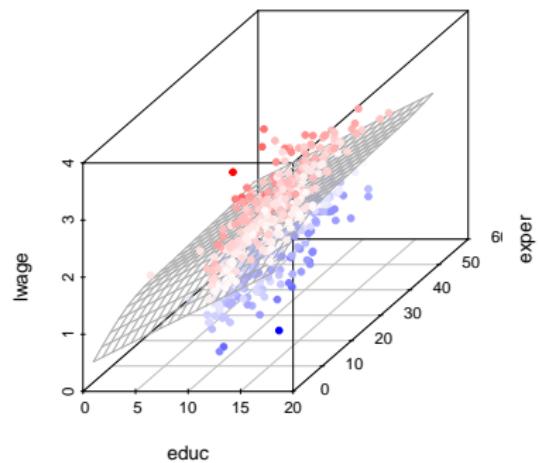
$$\hat{y} = 0,520 + 0,090 \text{ educ} + 0,0349 \text{ exper} - 0,001 \text{ exper}^2$$



Listing 19.5: example_lm2.R

```
1 library("plot3d")
2 data(Boston, package="MASS")
3 model <- lm(medv~lstat+rm, data=Boston)
4 par(mfrow=c(1,1))
5 new3d("s") %>% regression3d(model, Boston) %>% plot3d()
6 par(mfrow=c(2,2))
7 plot(model)
```

Partial linear model

Partial Linear Model ($R^2=0.250$)

```
gam(formula = l wage ~ educ + s(exper), data = x)
```

$$\hat{y} = 0,669 + 0,091 \text{ educ} + 0,011 \hat{m}(\text{exper})$$

 Listing 19.6: example_plm2.R

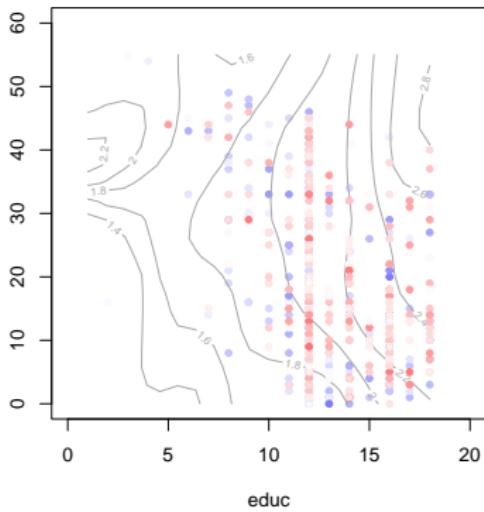
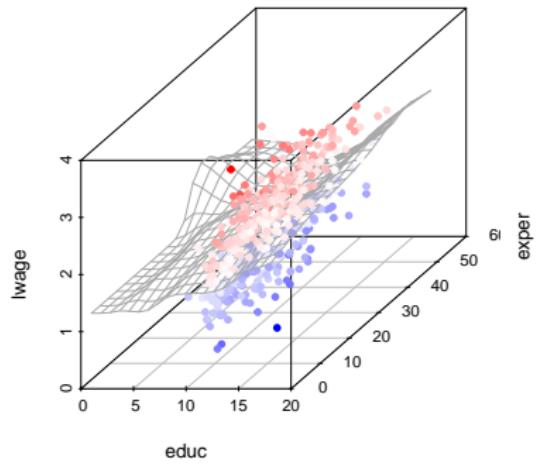
```
1 library("MASS")  # for Boston Housing data
2 library("mgcv")
3 library("plot.3d")
4 model <- gam(medv~lstat+rm, data=Boston)
5 new3d("c") %>% regression3d(model, Boston) %>% plot3d()
6 plot(model)
```

 `mgcv:::gam(formula, family=gaussian)`

 `gam:::gam(formula, family=gaussian)`

 `library mgcv` is preferred

Bivariate Nadaraya-Watson estimator

Bivariate Kernel Regression ($R^2=0.250$)

```
npreg.default(bws = lwage ~ educ + exper, ... = pairlist(data = x))
```

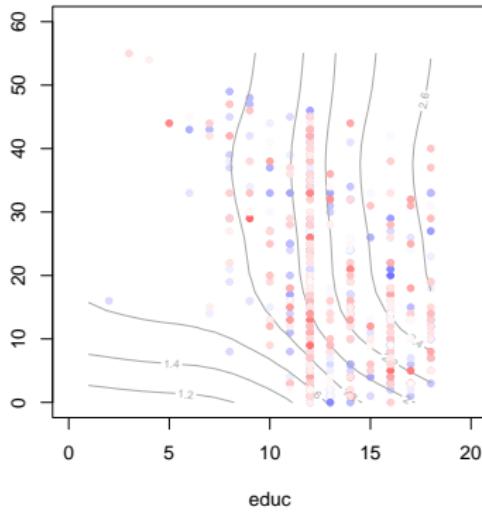
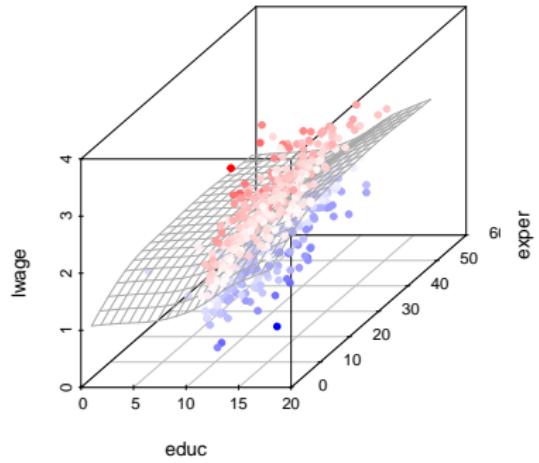
$$\hat{y} = \hat{m}(educ, exper) \text{ with } h_{educ} = 1,381, h_{exper} = 4,796$$

R Listing 19.7: example_bmw2.R

```
1 library("np")
2 library("plot.3d")
3 data(Boston, package="MASS")
4 model <- npreg(medv~lstat+rm, data=Boston)
5 par(mfrow=c(1,1))
6 new3d("s") %>% regression3d(model, Boston) %>% plot3d()
7 plot(model, view="fixed", theta=45)
```

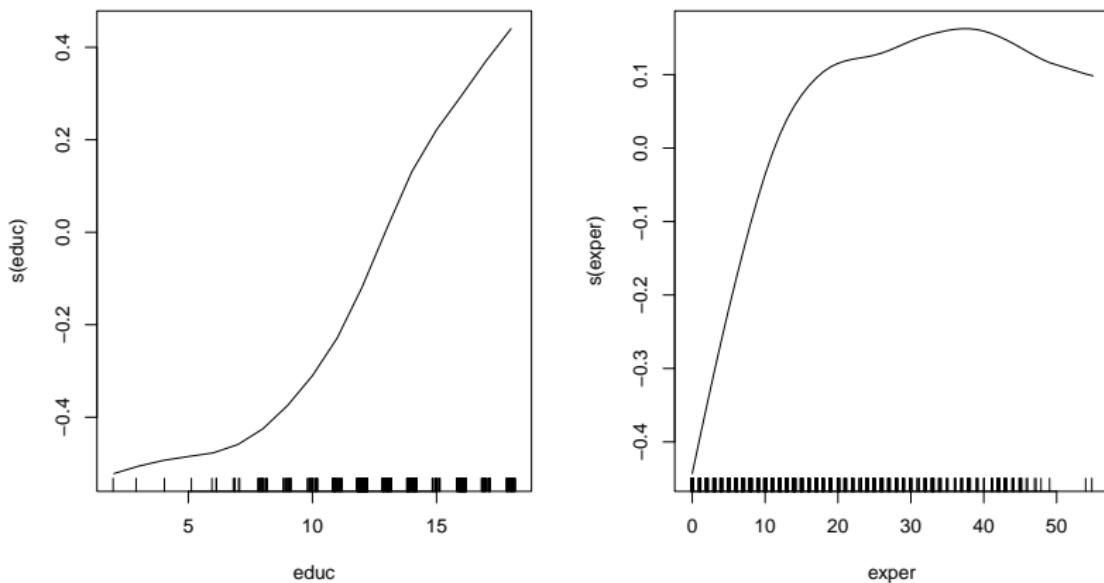
R np::npreg(formula, bws)

Additive model

Additive Model ($R^2=0.260$)

```
gam(formula = l wage ~ s(educ) + s(exper), data = x)
```

$$\hat{y} = 0,696 + 0,090 \hat{m}_1(\text{educ}) + 0,011 \hat{m}_2(\text{exper})$$



- compare fits to linear regression, e.g. via confidence bands



Listing 19.8: example_gam2.R

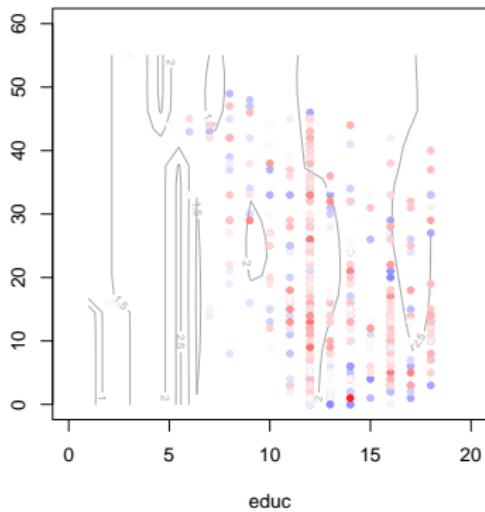
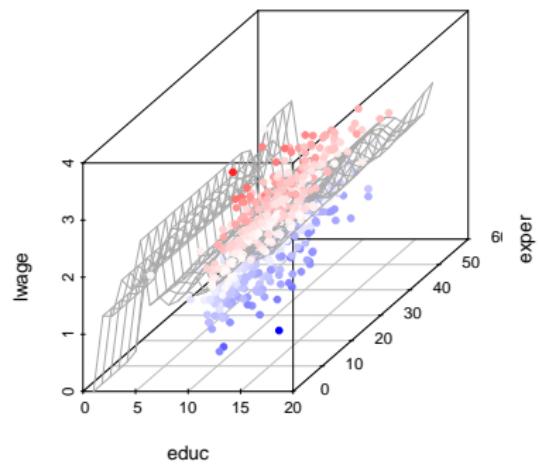
```
1 library("MASS") # for Boston Housing data
2 library("mgcv")
3 library("plot.3d")
4 model <- gam(medv~s(lstat)+s(rm), data=Boston)
5 new3d("s") %>% regression3d(model, Boston) %>% plot3d()
6 par(mfrow=c(1,2))
7 plot(model)
```

☞ `mgcv:::gam(formula, family=gaussian)`

☞ `gam:::gam(formula, family=gaussian)`

⚠ `library mgcv` is preferred

Single Index Model

Single Index Model ($R^2=0.271$)

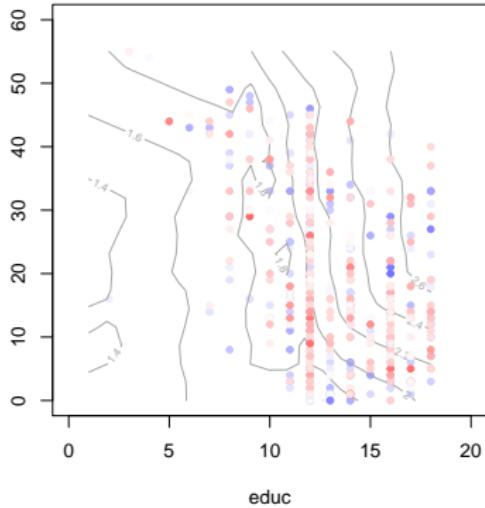
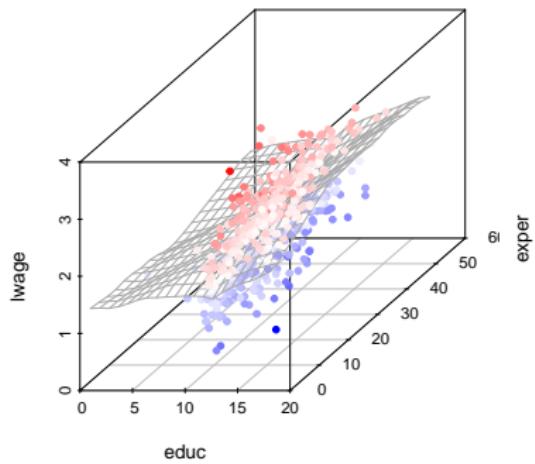
$$\hat{y} = \hat{m}(educ + 0, 111 exper) \text{ with } h = 0, 965$$

R Listing 19.9: example_sim.R

```
1 library("MASS") # for Boston Housing data
2 library("np")
3 model <- npindex(medv~lstat+rm, data=Boston)
4 library("plot.3d")
5 new3d("s") %>% regression3d(model, Boston) %>% plot3d()
6 plot(model)
```

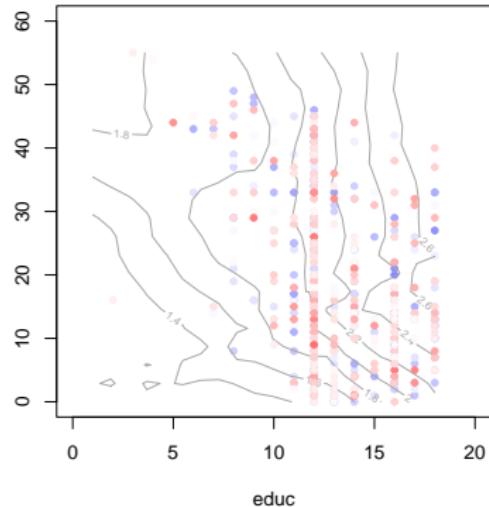
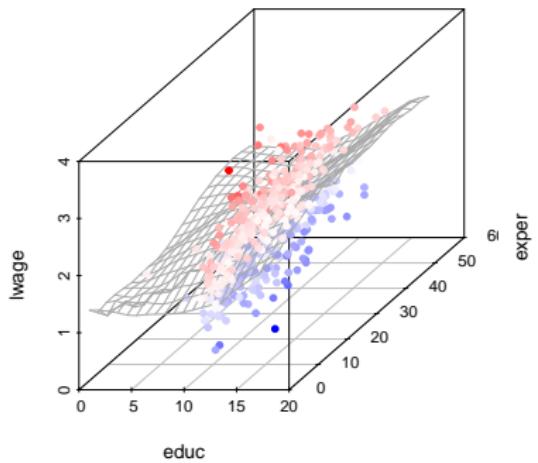
R np:npindex(formula, bws)

Projection Pursuit Regression

Projection Pursuit Regression Model ($R^2=0.261$)

```
ppr(formula = l wage ~ educ + exper, data = x, nterm = 2)
```

$$\hat{y} = 0,271 \hat{m}_1(0,998 \text{ educ} + 0,070 \text{ exper}) + 0,113 \hat{m}_2(-0,868 \text{ educ} + 0,497 \text{ exper})$$

Projection Pursuit Regression Model ($R^2=0.281$)

```
ppr(formula = lwage ~ educ + exper, data = x, nterm = 5)
```

$$\hat{y} = \sum_{i=1}^{5} r_i \hat{m}_i (\beta_{i1} \text{educ} + \beta_{i2} \text{exper})$$

 Listing 19.10: example_ppr.R

```
1 library("MASS") # for Boston Housing data
2 library("plot.3d")
3 model <- ppr(medv~lstat+rm, data=Boston, nterm=2)
4 new3d("c") %>% regression3d(model, Boston) %>% plot3d()
5 par(mfrow=c(1,2))
6 plot(model)
```

 ppr(formula, nterms)

Overview

Model	LM	LM^2	PLM	NW	AM	SIM	PPR2	PPR5
R^2	0,212	0,238	0,250	0,250	0,260	0,214	0,260	0,279
Param.	2	3	3	0	3	1	6	15
Func.	0	0	1	1	2	1	2	5

- Best model: Projection Pursuit Regression with five terms
 - ▶ but 15 parameter and 5 non-parametric functions to estimate
- Second best models: Projection Pursuit Regression with two terms and Additive model
 - ▶ Additive model preferred since less parameters and non-parametric functions to estimate
 - ▶ easier to interpret
- All other models have a smaller $R^2 \Rightarrow$ Additive model

Classification and regression trees

November 3, 2022

- CART • Construction of trees • 1R - Categorical variables • 1R - Numeric variables • Simplicity pays off! • Statistical modeling • Criterions for variable selection • Information gain • Gini index • CHAID • Regression tree • Pruning • Mincer equation • Ensemble methods

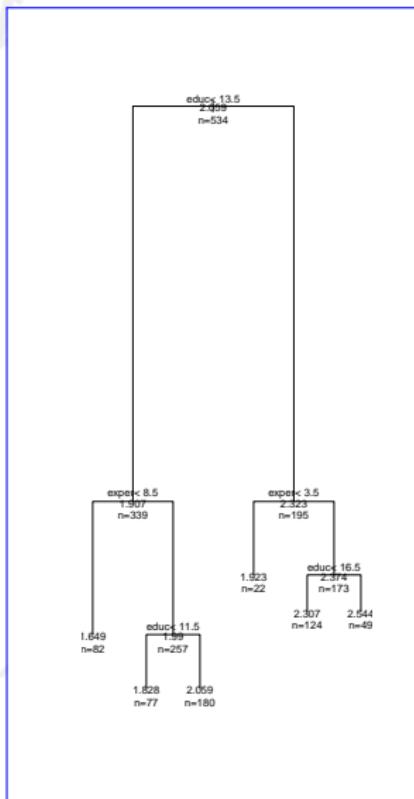
CART

- Setting:
 - ▶ Explanatory variables: X_1, \dots, X_p
 - ▶ Dependent variable: Y
 - ▶ Y numeric \Rightarrow regression models
 - ▶ Y categorical \Rightarrow classification models
- Variety of algorithms:
 - ▶ 1R
 - ▶ CART
 - ▶ C4.5, C5.0
 - ▶ CHAID (Chi-square Automatic Interaction Detectors)
 - ▶ ...

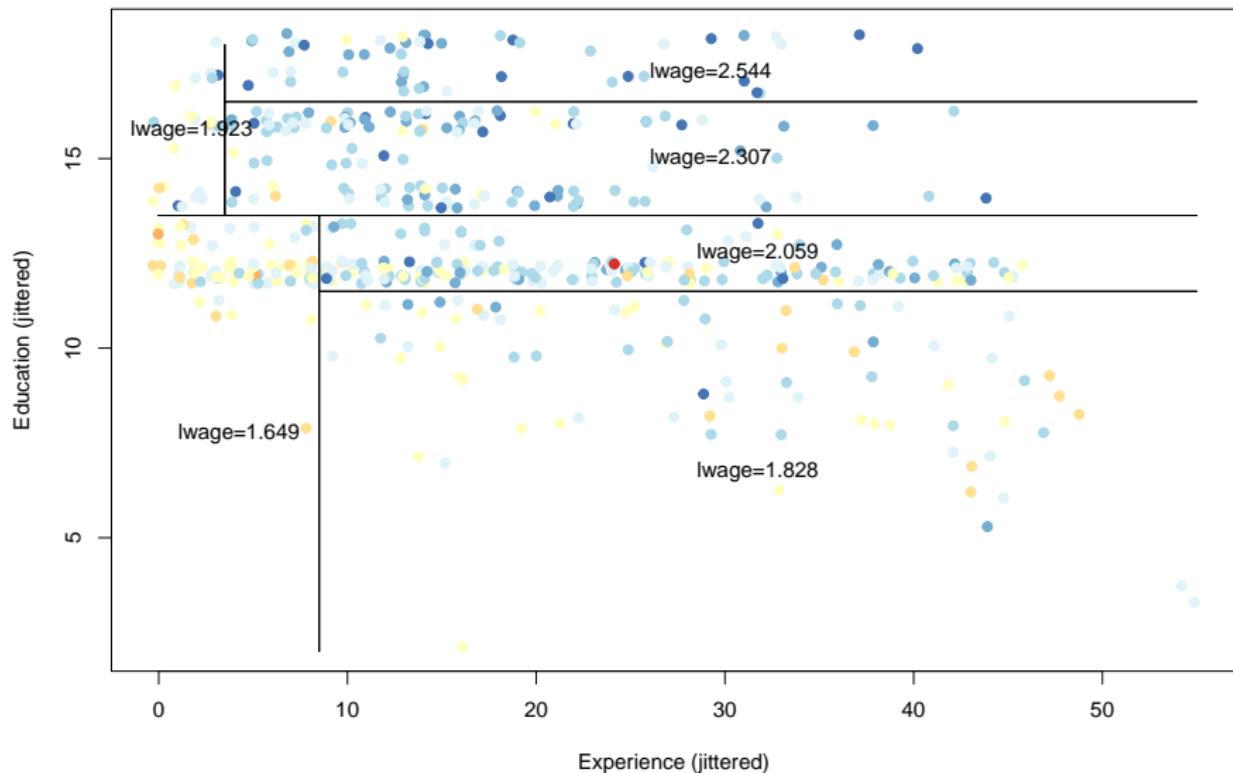
Task View: [MachineLearning](#)→Recursive Partitioning & Random Forests

Construction of trees

- Strategy: top down in a recursive divide-and-conquer fashion
- The whole set of observations constitutes a root leaf
- For the current leaf (= current set of observations) do:
 - ▶ For each variable find split(s) such that the “error” is minimized
 - ▶ Split the observations with the “best” split(s) in two (or more!) leafs
- Repeat with each leaf until some stopping criterion is reached
- Advantage: Interpretability
- Disadvantage: Unstable



CPS85 data with rpart defaults



 Listing 20.1: example_cart.R

```
1 library("rpart")
2 library("MASS") # for Boston Housing data
3 model <- rpart(medv~lstat+rm, data=Boston)
4 summary(model)
5 plot(model)
6 text(model)
```

 Listing 20.2: example_cartplot.R

```
1 library("rpart")
2 library("rpart.plot")
3 library("MASS") # for Boston Housing data
4 model <- rpart(medv~lstat+rm, data=Boston)
5 rpart.plot(model)
```

✉ rpart::rpart(formula, data, method, parms, control)

✉ rpart::rpart.control(minsplit=20, minbucket=round(minsplit/3),
cp=0.01, maxdepth=30)

✉ rpart.plot::rpart.plot(x, type=2)

1R - Categorical variables

- Choose only one “good” variable for prediction

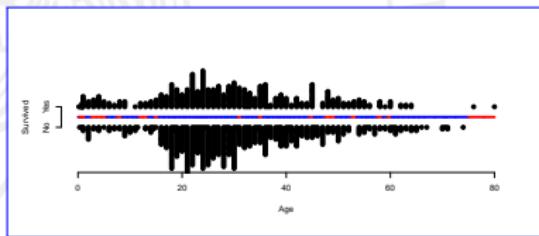
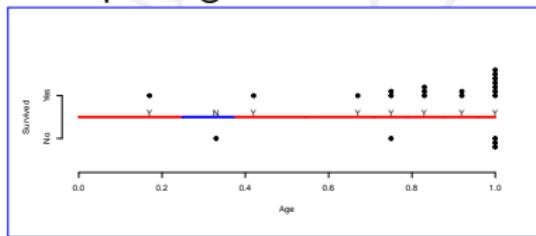
Variable Survived	Sex				Child				
	No	Yes	Pred.	Err	No	Yes	Pred.	Err	
F	127	339	Yes	127	N	761	439	No	439
M	682	161	No	161	Y	48	61	Yes	48
Total Percent					282				
					21%				
Variable Survived	Class								
	No	Yes	Pred.	Err					
1	123	200	Yes	123					
2	158	119	No	119					
3	528	181	No	181					
					423				
					32%				

- Missing values are treated

- as another category (categorical)
- are split up to leafs (numeric)

1R - Numeric variables

- Discretize the variable (*nary split !*)
- Sort observations according to their value
- Place breakpoints where class changes (majority class)
- Example: age from titanic data



- Problem 1: sensitive to noise
- Problem 2: e.g. passenger name would be perfect predictor
- Solution: Require a minimum number n_{\min} of observations for a class (blue and red areas)



Listing 20.3: example_rpart1.R

```
1 library("MASS") # for Boston Housing data
2 library("rpart")
3 ctrl <- rpart.control(maxdepth=1)
4 model <- rpart(medv~., data=Boston, control=ctrl)
5 print(model)
6 plot(model)
7 text(model)
8 summary(model)
```

Simplicity pays off!

- Holte (1993)
 - ▶ Contains an experimental evaluation on 16 datasets
 - ▶ Minimum number of observations was set to $n_{\min} = 6$ after some experimentation
 - ▶ 1R's simple rules performed not much worse than much more complex decision trees
- Another simple approach for multi-class response:
Build a tree for each class
- “Opposite” of 1R: statistical modelling
 - ▶ Use all variables
 - ▶ Variables are (equally) important
 - ▶ Variables are statistically independent (given the class value)
 - ▶ Independence assumption is never correct, but statistical modeling works well in practice

Holte, Robert C. (1993). "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets". In: *Machine Learning* 11.1, pp. 63–90. issn: 1573-0565. doi: [10.1023/A:1022631118932](https://doi.org/10.1023/A:1022631118932). url: <http://dx.doi.org/10.1023/A:1022631118932>.

Statistical modeling

Variable	Survived	
Sex	No	Yes
Female	16%	68%
Male	84%	32%
Child	Survived	
No	94%	88%
Yes	6%	12%
Class	Survived	
1	15%	24%
2	20%	36%
3	65%	38%
Sum	62%	38%

- Compute surviving and non-surviving likelihoods for an observation $P(E|S)$ and $P(E|\bar{S})$
- Example: female, child, first class
 - ▶ $P(E|S) = 0,68 \cdot 0,12 \cdot 0,24 \approx 0,0196$
 - ▶ $P(E|\bar{S}) = 0,16 \cdot 0,06 \cdot 0,15 \approx 0,0014$
- Theorem of Bayes (Normalization):
 - ▶ $P(S|E) = \frac{P(E|S) \cdot P(S)}{P(E)} = \frac{P(E|S) \cdot P(S)}{P(E|S)P(S) + P(E|\bar{S})P(\bar{S})}$
 - ▶ $\frac{0,0196 \cdot 0,38}{0,0196 \cdot 0,38 + 0,0014 \cdot 0,62} = \frac{0,0126}{0,0126 + 0,009} \approx 0,935$
 - ▶ $P(\bar{S}|E) = 1 - P(S|E) \approx 0,065$
 - ▶ Prediction: Survived=Yes
- for numeric variables use the density function
- missing values are not included in frequency and likelihood calculation

Criterions for variable selection

- Which is the best variable?
 - ▶ Want to get the smallest tree
 - ▶ Heuristic: choose the variable that produces the “purest” nodes
- Criterion for classification (impurity measures)
 - ▶ Information gain based on entropy
 - ▶ Gini index reduction
- Criterion for regression
 - ▶ Variance reduction
- Most criterions can be extended to the multiple split case
- Strategy: choose variable that gives the largest reduction
- Consideration of surrogate splits might be helpful

Information gain

- A node is splitted in a left and right node

- ▶ Number of observations $n = n_l + n_r$
 - ▶ Relative frequencies $f_1, \dots, f_k, f'_1, \dots, f'_k, f''_1, \dots, f''_k$

- Entropy for k classes

$$H = - \sum_{i=1}^k f_i \log(f_i), H_l = - \sum_{i=1}^k f'_i \log(f'_i), H_r = - \sum_{i=1}^k f''_i \log(f''_i)$$

- Information gain: information before splitting - expected information after splitting

$$gain = H - \left(\frac{n_l}{n} H_l + \frac{n_r}{n} H_r \right)$$

- Problematic: variables with a large number of values (extreme case: ID code)
- Subsets are more likely to be pure if there is a large number of values
 - ▶ Information gain is biased towards choosing variables with a large number of values
 - ▶ This may result in overfitting (selection of a variable that is non-optimal for prediction)
- Possible solution: gain ratio

$$\frac{gain}{H_{\text{Variable}}}$$

- Another problem: fragmentation

```
R rpart::rpart(formula, data, method="class",
               parms=list(split="information"))
```

Example 20.29 (Titanic)

	Survived	root	left	right
Sex		All	Male	Female
	Yes	711	367	344
	No	1490	1364	126
Entropy		0.629	0.517	0.581
Information gain		0.629-0.406-0.124=0.099		
Child		All	Yes	No
	Yes	711	57	654
	No	1490	52	1438
Entropy		0.629	0.692	0.621
Information gain		0.629-0.034-0.590=0.004		

Gini index

- A node is splitted in a left and right node
 - ▶ Number of observations $n = n_l + n_r$
 - ▶ Relative frequencies $f_1, \dots, f_k, f'_1, \dots, f'_k, f''_1, \dots, f''_k$
- Probability for *not* selecting a element of the same class within two draws

$$G = 1 - \sum_{i=1}^k f_i^2, G_l = 1 - \sum_{i=1}^k (f'_i)^2, G_r = 1 - \sum_{i=1}^k (f''_i)^2$$

- Impurity reduction

$$G - \left(\frac{n_l}{n} G_l + \frac{n_r}{n} G_r \right)$$

Example 20.30 (Titanic)

	Survived	root	left	right
Sex		All	Male	Female
	Yes	711	367	344
	No	1490	1364	126
Gini		0.437	0.339	0.392
Impurity reduction		0.437-0.266-0.084=0.087		
Child		All	Yes	No
	Yes	711	57	654
	No	1490	52	1438
Gini		0.437	0.498	0.430
Impurity reduction		0.437-0.025-0.408=0.004		

 Listing 20.4: example_rpartclass.R

```
1 library("rpart")
2 tit <- as.data.frame(Titanic)
3 ind <- rep(seq(nrow(tit)), tit$Freq)
4 x <- tit[ind,]
5 #
6 fit <- rpart(Survived~Class+Sex+Age, data=x)
7 fit
8 #
9 fit <- rpart(Survived~Class+Sex+Age, data=x,
10               method="class",
11               parms=list(split="information"))
12 fit
13 plot(fit); text(fit)
```

 `rpart::rpart(formula, data, method="class",
 parms=list(split="gini"))`

 Default for a categorical dependent variable

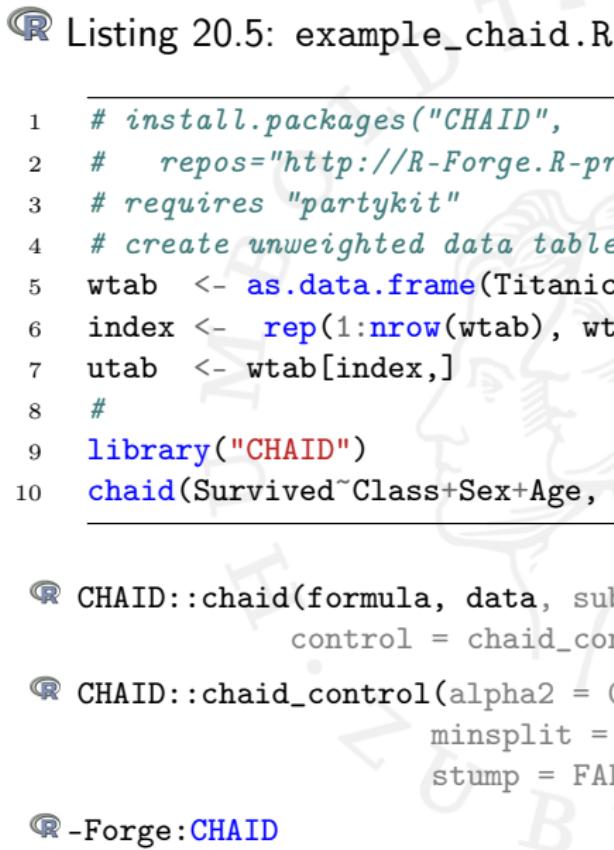
CHAID

- Chi-square Automatic Interaction Detectors
- Only for categorical variables
- Use χ^2 test statistics on nodes and classes (adjusted for multiple testing)
- Always uses multiway splits

Example 20.31 (Titanic)

	Survived	left	right
Sex		Male	Female
	Yes	367	344
	No	1364	126
χ^2		454.500	
Child		Yes	No
	Yes	57	654
	No	52	1438
χ^2		20.005	

→ choose sex as split variable

R Listing 20.5: example_chaid.R

```
1 # install.packages("CHAID",
2 #   repos="http://R-Forge.R-project.org")
3 # requires "partykit"
4 # create unweighted data table
5 wtab <- as.data.frame(Titanic)
6 index <- rep(1:nrow(wtab), wtab$Freq)
7 utab <- wtab[index,]
8 #
9 library("CHAID")
10 chaid(Survived~Class+Sex+Age, data=utab)
```

R CHAID::chaid(formula, data, subset, weights, na.action = na.omit,
control = chaid_control())

R CHAID::chaid_control(alpha2 = 0.05, alpha3 = -1, alpha4 = 0.05,
minsplit = 20, minbucket = 7, minprob = 0.01,
stump = FALSE, maxheight = -1)

R -Forge:CHAID

Regression tree

- CART: estimate for each node a constant $\bar{y} = \frac{1}{n} \sum_{i \in \text{node}} y_i$
 - ▶ Split a node in two subnodes (left and right) by variance reduction, it holds

$$n = n_l + n_r$$

$$\bar{y} = \frac{n_l}{n} \bar{y}_l + \frac{n_r}{n} \bar{y}_r$$

$$V = \frac{n_l}{n} V_l + \frac{n_r}{n} V_r + \frac{n_l}{n} (\bar{y}_l - \bar{y})^2 + \frac{n_r}{n} (\bar{y}_r - \bar{y})^2$$

- model tree: estimate for each node a linear model
 $\hat{y}_i = b_0 + \sum_{j=1}^p b_j x_{ij}$ (with all $i \in \text{node}$)
 - ▶ use the error variances of the linear models as V , V_l and V_r
- Variance reduction

$$V - \left(\frac{n_l}{n} V_l + \frac{n_r}{n} V_r \right)$$

Breiman, L. et al. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

⌚ Listing 20.6: example_rpartanova.R

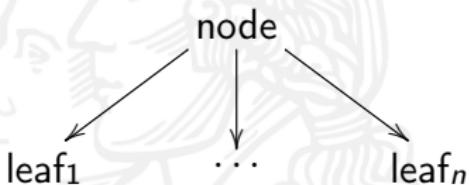
```
1 library("MASS") # for Boston Housing data
2 library("rpart")
3 model <- rpart(medv~., data=Boston)
4 print(model)
5 plot(model); text(model, cex=0.8)
```

⌚ `rpart::rpart(formula, data, method="anova")`

⚠ Default for a continuous dependent variable

Pruning

- Grow largest possible tree (overfitting) and reduce later (pruning)
- For each terminal node compute missclassifications rates (errors) e_n of a node and its leaves $e_L = e_1 + \dots + e_n$



- Reduced error pruning
 - ▶ create training and test data
 - ▶ build largest possible tree with training data
 - ▶ start at leaves (bottom-up)
 - ▶ compare missclassifications rates (errors) on test data
 - ▶ if $e_L \geq e_n$ then reduce the tree to the node

- Pessimistic error pruning
 - ▶ missclassifications in the node 'follow a binomial distribution
 $E_n \sim B(n, p)$
 - ▶ use a normal distribution with continuity correction as approximation
 - ▶ start at root node (top-down)
 - ▶ if $e_L \geq e_n - \sqrt{\text{Var}(E_n)}$ then reduce the tree to the node
- Cost-complexity pruning
 - ▶ For possible trees compute $C_{\alpha;n} = e_T + \alpha \cdot \#Tn$ for
 - ★ $\#Tn$ number of leaves
 - ★ α balances accuracy vs. complexity
 - ★ e_T sum of missclassifications in all terminal leafs
 - ▶ To find α make k -fold crossvalidation
 - ▶ If $C_{\alpha;n} \leq C_{\alpha;i}$ then reduce subtree to the node



Listing 20.7: example_prune.R

```
1 library("MASS") # for Boston Housing data
2 library("rpart")
3 model <- rpart(medv~., data=Boston, cp=0)
4 plot(model); text(model, cex=0.8)
5 printcp(model)
6 plotcp(model)
7 pmodel1 <- prune(model, cp=0.25)
8 pmodel1
9 cp       <- model$cptable[which.min(
10                  model$cptable[, "xerror"]),"CP"]
11 pmodel2 <- prune(model, cp=cp)
12 pmodel2
```



rpart::prune(tree, cp)

Mincer equation

- Mincer (1974) related the wage to “education” and “working experience”

$$\text{lwage} = b_0 + b_1 \text{educ} + b_2 \text{exper}$$

- From Current Population Survey 1985

- ▶ lwage - log of hourly wage
- ▶ educ - number of years of education
- ▶ exper - number of years of work experience

Residuals:

Min	1Q	Median	3Q	Max
-2.0337	-0.3305	0.0423	0.3187	1.8398

Coefficients:

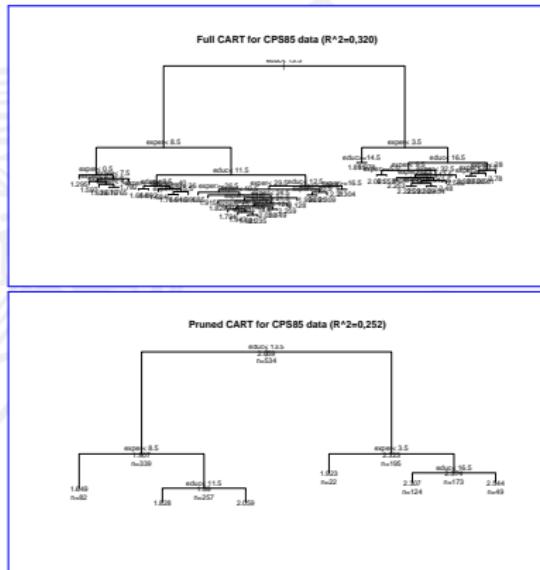
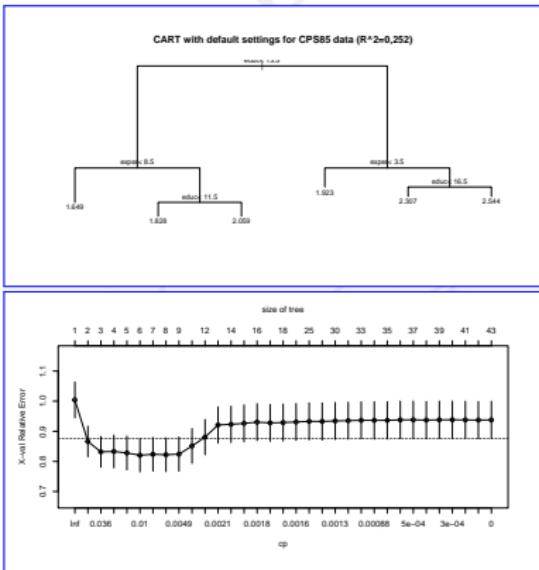
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.594137	0.124425	4.775	2.33e-06 ***
educ	0.096416	0.008309	11.603	< 2e-16 ***
exper	0.011774	0.001755	6.707	5.10e-11 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4695 on 531 degrees of freedom

Multiple R-squared: 0.2115, Adjusted R-squared: 0.2085

F-statistic: 71.22 on 2 and 531 DF, p-value: < 2.2e-16



Ensemble methods

- Basic idea
 - ▶ Build several models and “average” the results
 - ▶ Advantage: often improves predictive performance
 - ▶ Disadvantage: usually produces output that is very hard to analyze
- Bootstrap aggregating (Bagging)
 - ▶ Combining predictions by voting/averaging
 - ▶ Sample several data sets from data (bootstrap)
 - ▶ Build a model for each bootstrap data set
 - ▶ Average over the model results
- Random forest
 - ▶ applies bagging idea to trees
 - ▶ no pruning required

② Listing 20.8: example_randomforest.R

```
1 library("MASS") # for Boston Housing data
2 library("randomForest")
3 model <- randomForest(medv~., data=Boston)
4 summary(model)
5 imp <- importance(model)
6 ind <- order(imp, decreasing=T)
7 imp[ind,]
```

② randomForest::randomForest(formula, data)
② randomForest::importance(forest)

- Boosting
 - ▶ As in bagging
 - ▶ But weight models according to performance
 - ▶ Iterative: new models are influenced by performance of previously built ones
 - ★ Encourage new model to become an “expert” for observations misclassified by earlier models
 - ★ Intuitive justification: models should be experts that complement each other
 - ▶ In practice, boosting sometimes overfits (in contrast to bagging)
- Several variants
 - ▶ AdaBoost.M1: weight observations for next model
 - ▶ LogitBoost: minimizes the logistic loss
 - ▶ ...

⌚ package adabag or ada

⌚ package mboost

Example 20.32

- Bagging: additive model

$$\beta_0 + \sum_{i=1}^p m_i(x_i)$$

- ▶ each m_i becomes an expert on variable X_i
- ▶ a weighted average of the experts is the additive model

- Boosting: projection pursuit model

$$\sum_{j=1}^M m_j \left(\beta_{j0} + \sum_{i=1}^p \beta_{ji} x_i \right)$$

- ▶ estimate on the residuals m_j and direction $\beta_{j\bullet}$
- ▶ repeat until $j = M$

- Classification and regression trees

```
  R tree::tree(formula, data, control=tree.control(nobs,...),  
              split=c("deviance","gini"))  
  R tree::prune(tree)
```

 Package rpart is preferred over tree!

```
  R party::ctree(formula, data)
```

- Random forests

```
  R party::cforest(formula, data)
```

- Ensemble methods and misc.

```
  R ipred::bagging(formula, data, nbagg=25)  
  R mboost::mboost(formula, data, baselearner=c("bbs",...))  
  R bootstrap::crossval(x, y, theta.fit, theta.predict, ngroup=n)  
  R bootstrap::bootstrap(x, nboot, theta)
```

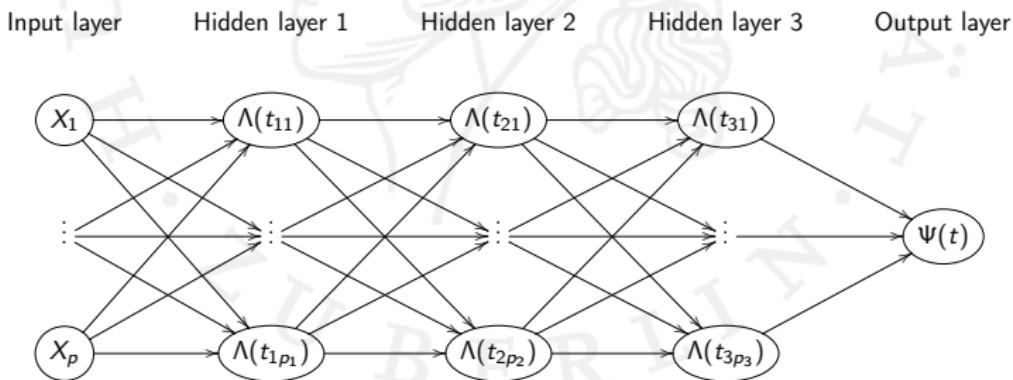
Neural networks

November 3, 2022

- Neural networks • Universal approximation theorem • Activation functions
- Error functions • Backpropagation • Data decomposition • Early stopping • Weight or soft decay • Mincer data with neural networks • Example • Convolutional Neural Networks • Deep learning frameworks

Neural networks

- Human brain:
 - ▶ neurons are connected with each other
 - ▶ electric impulses are send from one neuron to another
 - ▶ if the sum of the incoming electric impulses is above a certain level then the neuron sends out a electric impulse
- Neural networks imitate this process
 - ▶ Example: Three (hidden) layer feed forward neural network



A lot of different network architectures are possible:

- more than one layer
- connections between non-neighbouring (hidden) layers
- connections between different neurons in one hidden layer
- different number of neurons per layer
- different activation functions Λ in each neuron
- other output function Ψ than identity
- BUT theory says that one hidden layer feed forward networks are sufficient, but may be not efficient

Universal approximation theorem

Let Λ be a nonconstant, bounded, and monotonically-increasing continuous function. Let I_p denote the p -dimensional unit hypercube $[0, 1]^p$. The space of continuous functions on I_p is denoted by $C(I_p)$. Then, given any function $f \in C(I_p)$ and $\varepsilon > 0$, there exist an integer N and sets of weights $w_j, w_{ij} \in \mathbb{R}$ such that we may define:

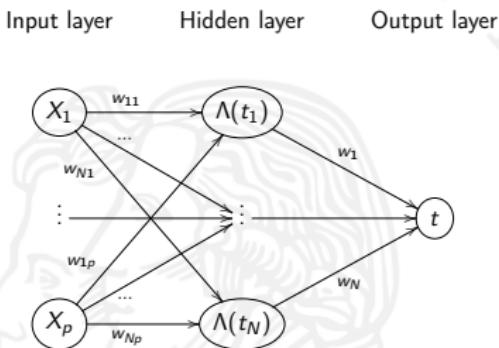
$$\hat{f}(x) = \sum_{j=1}^N w_j \Lambda \left(w_{i0} + \sum_{i=1}^p w_{ij} x_i \right)$$

as an approximate realization of the function f that is,

$$|\hat{f}(x) - f(x)| < \varepsilon \text{ for all } x \in I_p.$$

Cybenko, G. (Dec. 1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314. issn: 0932-4194, 1435-568X. doi: 10.1007/BF02551274. url: <http://link.springer.com/10.1007/BF02551274> (visited on 12/15/2016).

- Consequently a single layer feed-forward network is sufficient:



$$\text{with } t = w_0 + \sum_{j=1}^N w_j \Lambda(t_j), \quad t_j = w_{j0} + \sum_{i=1}^p w_{ij} x_i$$

- the feed-forward property generates the universal approximation property of the neural network and *not* the choice of the activation function (Hornik, 1991)

Hornik, Kurt (Jan. 1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2, pp. 251–257. issn: 08936080. doi: 10.1016/0893-6080(91)90009-T. url: <http://linkinghub.elsevier.com/retrieve/pii/089360809190009T> (visited on 12/15/2016).

Activation functions

- Step function:

$$\Lambda(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}$$

- Sigmoid function:

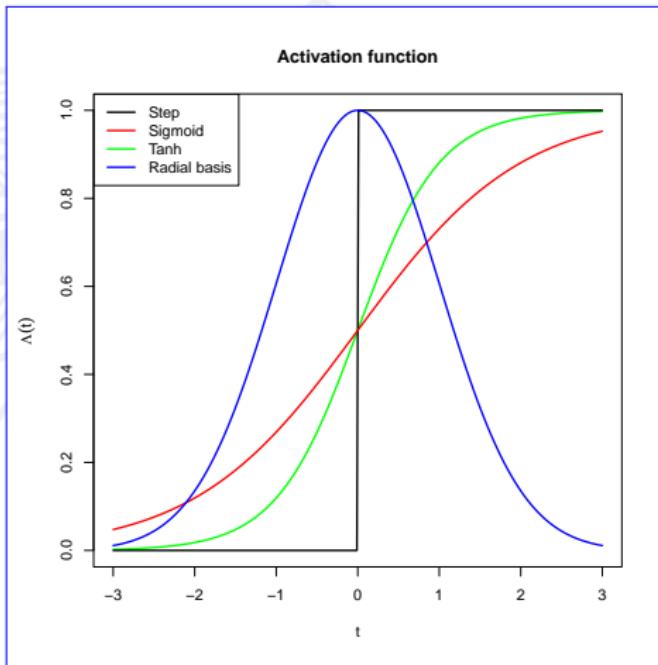
$$\Lambda(u) = \frac{1}{1 + \exp(-u)}$$

- Tangens hyperbolicus function:

$$\Lambda(u) = 0.5 + 0.5 \tanh(u)$$

- Radial basis function:

$$\Lambda(u) = g(|u - c_i|)$$



Error functions

- Error for regression

$$E_R = \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2$$

- Error for classification

$$E_C = \sum_{i=1}^n \underbrace{\sum_{j=1}^C y_{ij} \log(\hat{f}(x_i))}_{\text{cross-entropy}}$$

with $y_{ij} = 1$ if observation i belongs to class j otherwise $y_{ij} = 0$

- ▶ for multiple classes use C output neurons
- ▶ use Softmax (instead of identity) as activation function for output neurons

$$\Lambda_j(t_j^o) = \frac{\exp(t_j^o)}{\sum_{j=1}^C \exp(t_j^o)}$$

- ▶ ensures that output can be interpreted as probabilities

Backpropagation

- use an optimization method to minimize the error, e.g. an iterative gradient descent method

$$w_{ij,t} = w_{ij,t-1} - \lambda_t \frac{\partial E}{\partial w_{ij}}$$

- ▶ λ_t the learning rate with $\lim_{t \rightarrow \infty} \lambda_t = 0$
- ▶ choose learning rate from the inverse Hessian (= matrix of the second derivatives)
- use online learning (one observation) or batch learning (all observations)
- for the sigmoid function $\Lambda(u) = \frac{1}{1+\exp(-u)}$ holds

$$\frac{\partial \Lambda(u)}{\partial w_{ij}} = \Lambda(u)(1 - \Lambda(u)) \frac{\partial u}{\partial w_{ij}}$$

- ▶ simplifies the calculation of the derivative(s)

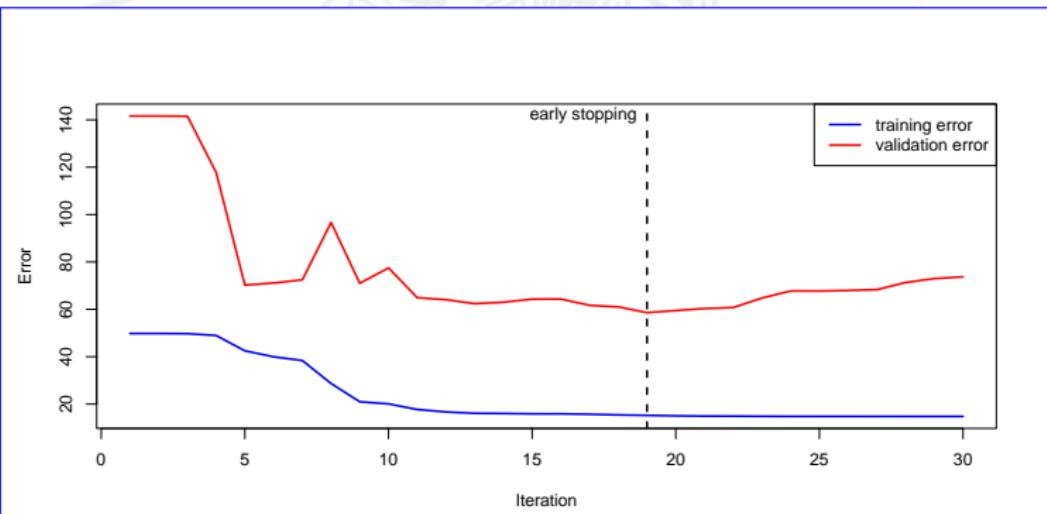
Data decomposition

Original dataset	Training set	determine weights w_{ij}
	Validation set	determine number of neurons N
	Test set	estimate error

- Training set: find for each candidate model the model parameters
- Validation set: find model hyperparameters/model selection (avoid overfitting)
- Test set: estimate out-of-sample error on selected model
- Rule-of-thumb: 60% training, 20% validation, 20% test
- small dataset \Rightarrow use k -fold crossvalidation instead training/validation sets

Early stopping

1. Divide the data into training and validation sets
2. Use a large number of hidden units
3. Use very small random initial values for the weights
4. Use a slow learning rate
5. Compute the validation error rate periodically during training
6. Stop training when the validation error rate “starts to go up”



Weight or soft decay

- To avoid overfitting weight or soft decay ω is an alternative to early stopping

$$E' = E + \omega \left(\sum w_j^2 + \sum w_{ij}^2 \right)$$

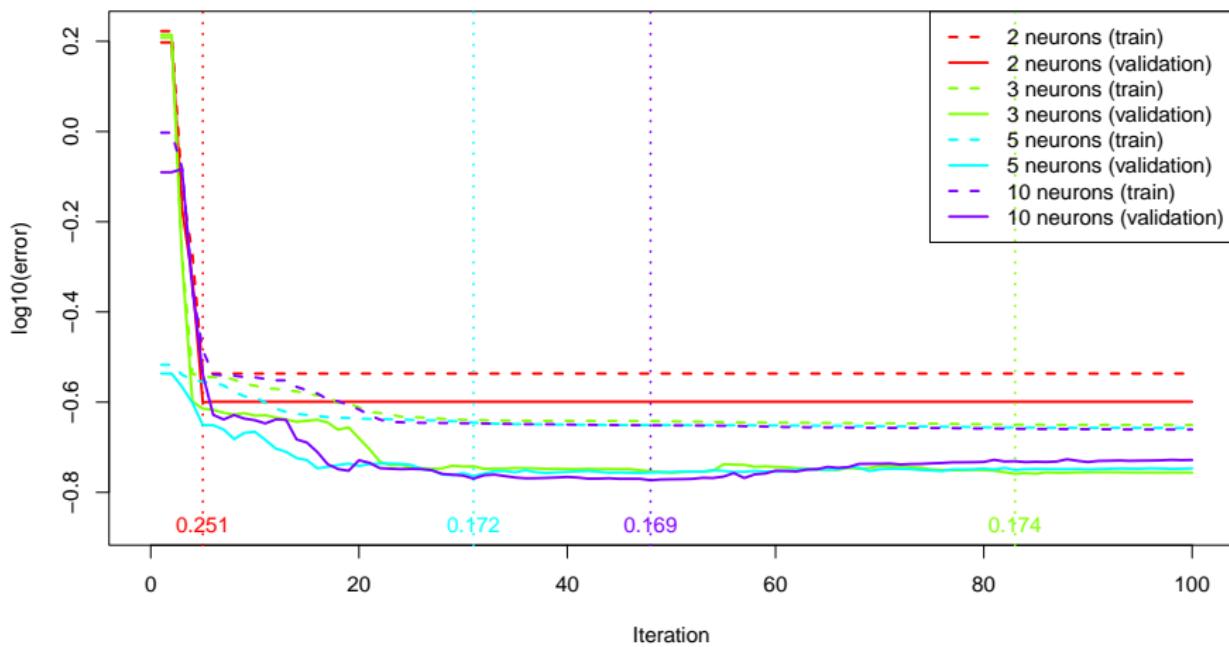
- Large weights are penalized
 - ▶ small weights mean that $\Lambda(t)$ is linear (simple model)
- BUT: under linear transformations of the input data different network weights are generated due to soft decay
- better to use two weight decay parameters

$$E' = E + \omega_1 \sum w_j^2 + \omega_2 \sum w_{ij}^2$$

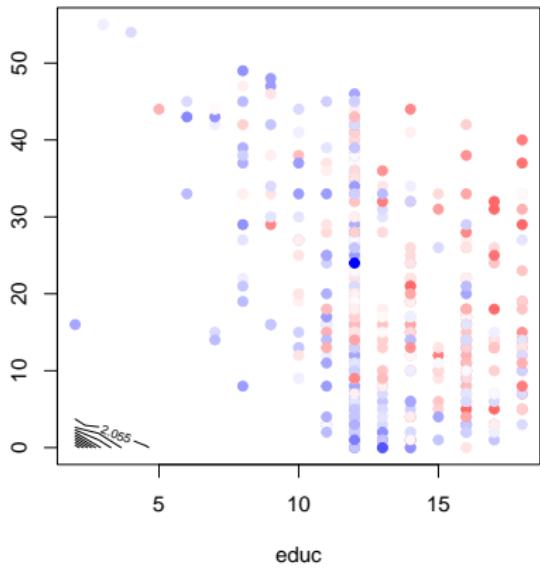
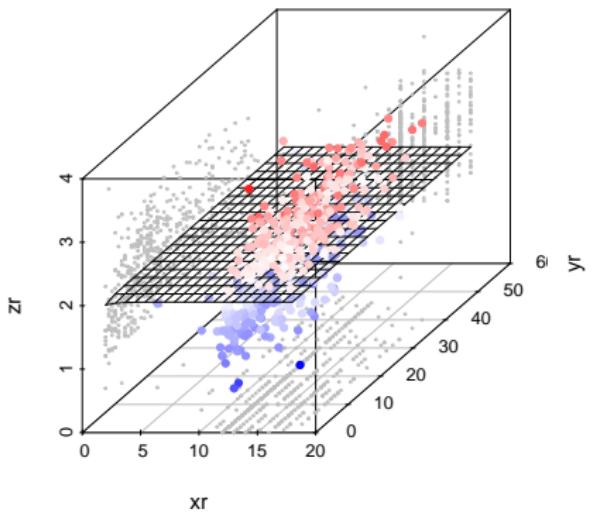
- disadvantage: we need to estimate more parameters from the data

Mincer data with neural networks

Neural networks with early stopping

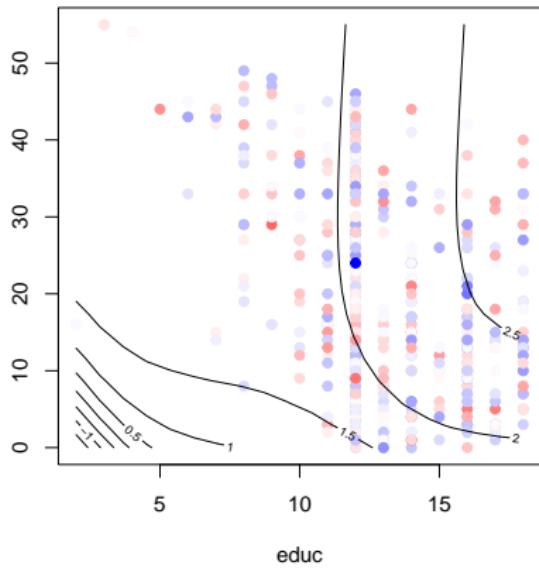
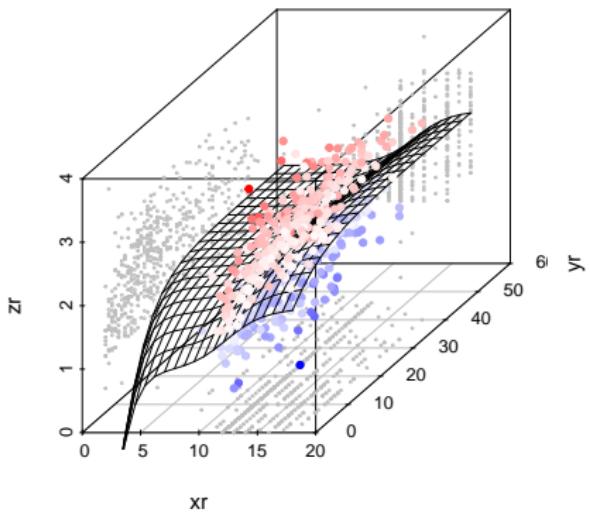


Neural network with 2 hidden neurons ($R^2=-0.000$)



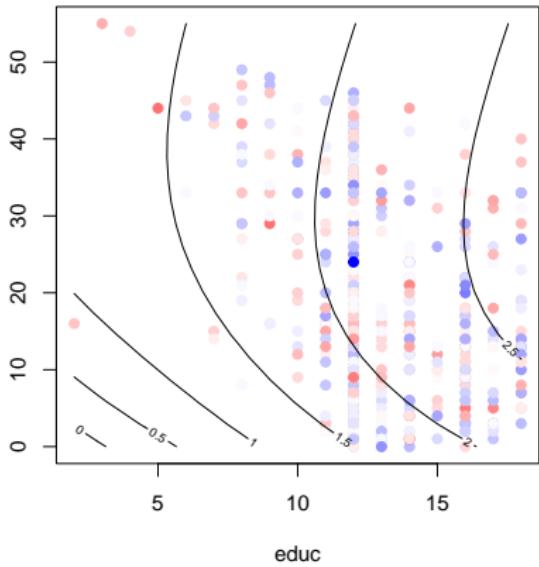
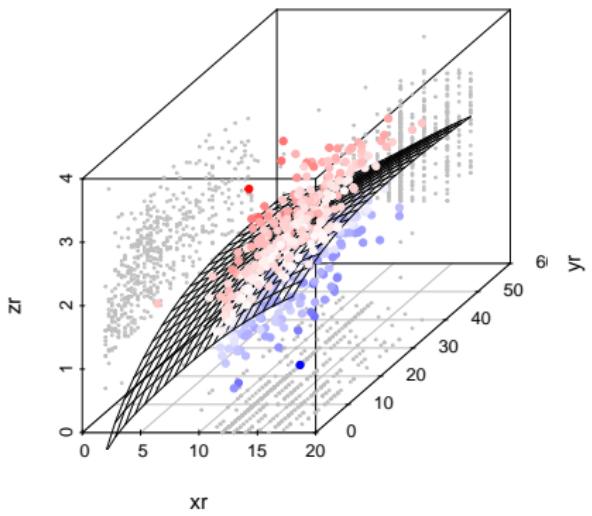
```
nnet.formula(formula = I(wage ~ educ + age + data$exp + data$T), x, maxit = minerrpos[i],
```

Neural network with 3 hidden neurons ($R^2=0.260$)



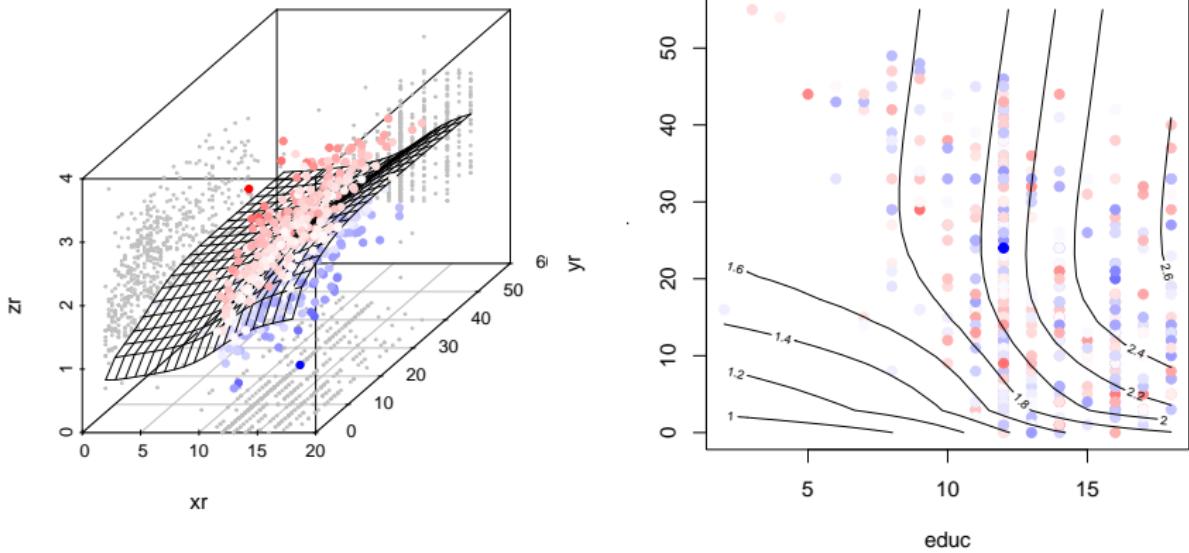
```
nnet.formula(formula = I(wage ~ educ + zr), data = exp0data), x, maxit = minerrpos[i],
```

Neural network with 5 hidden neurons ($R^2=0.247$)



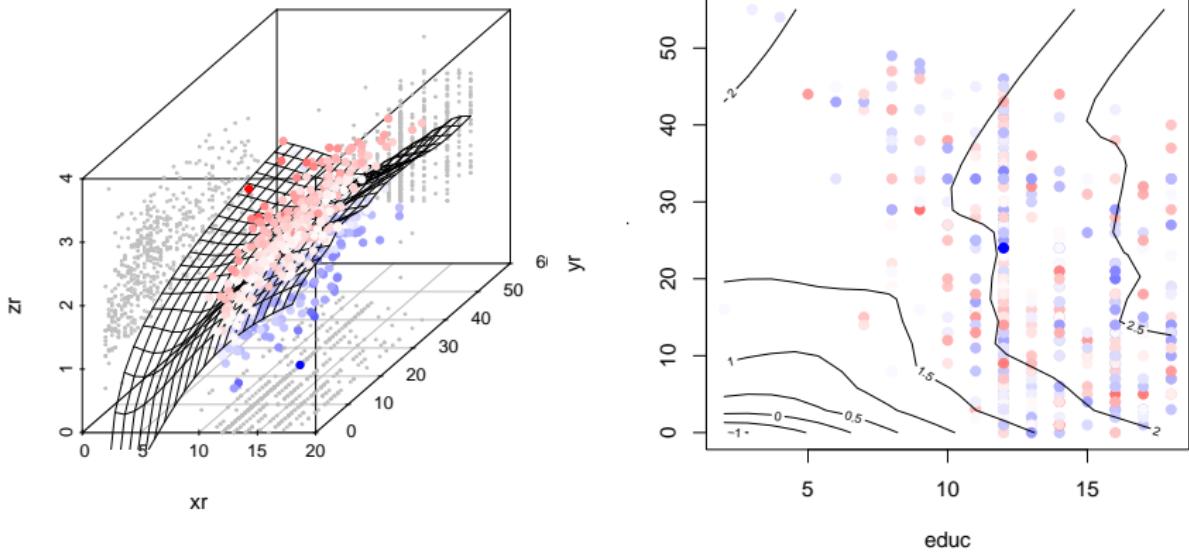
```
nnet.formula(formula = I(wage ~ educ + age + data), x, maxit = minerrpos[i],
```

Neural network with 10 hidden neurons ($R^2=0.260$)



```
nnet.formula(formula = I(wage ~ educ + age + data), x, maxit = minerrpos[i],
```

Neural network with 25 hidden neurons ($R^2=0.267$)



```
nnet.formula(formula = I(wage > size), size, expandData) = x, maxit = minerrpos[i],
```

Hidden neurons	#Weights	R^2
2	9	0.000
3	13	≈ 0.260
5	21	≈ 0.260
10	41	≈ 0.260
25	101	> 0.260
H	$(p+2)H+1$	-

p number of input variables

Example

R Listing 21.1: example_nnet.R

```
1 library("MASS")
2 library("nnet")
3 # run several times
4 model <- nnet(medv~lstat+rm, data=Boston, size=5,
5                           linout=T, maxi
6 summary(model)
7 plot(Boston$lstat, residuals(model),
8          xlab="Lstat", ylab="Residuals")
```

```
R nnet::nnet(formula, data, size=1, linout=F, maxit=100, softmax=F,
             decay=0)
```

Only one hidden layer

 Listing 21.2: example_neuralnet.R

```
1 library("MASS")
2 library("neuralnet")
3 model <- neuralnet(medv~lstat+rm, data=Boston,
4
5 plot(Boston$lstat, residuals(model),
6                 xlab="Lstat", ylab="Residuals")
7 plot(model, rep="best")
```

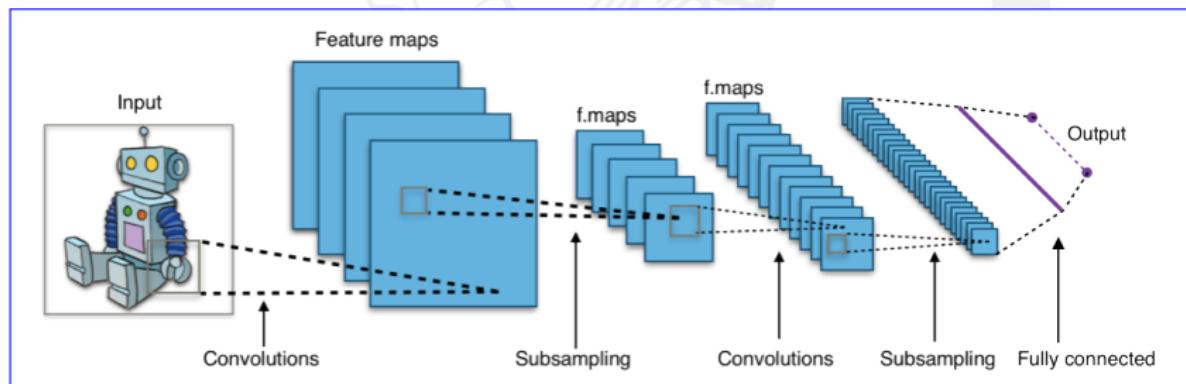


```
® neuralnet::neuralnet(formula, data, hidden=1, linear.output=T,
algorithm=c("rprop+", "backprop", "rprop-",
"sag", "slr"), err.fct=c("sse", "ce"),
act.fct=c("logistic", "tanh"))
```

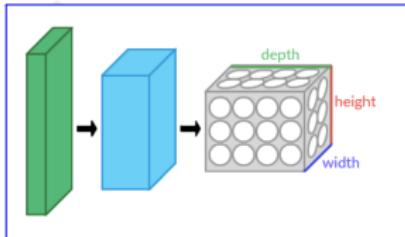
Multiple hidden layers

Convolutional Neural Networks

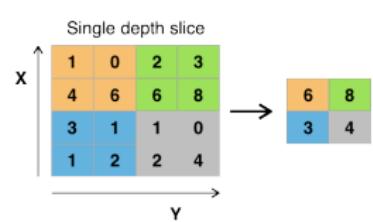
- convolutional networks (CNN) are inspired by biological processes, e.g. in image recognition
- 126 Mio. photoreceptor cells (retina) are connected to 3 Mio. ganglion cells (convolution) which send “raw data” to the brain
 - ▶ 120 Mio. rod cells: single photon activation, black and white
 - ▶ 6 Mio. cone cells: several photons required, three types for colors



- a CNN is mixture of convolutional and pooling layers with a final fully connected network layer
- a convolutional layer is a set of 2D or 3D neurons which are only locally connected and have shared weights
- a pooling layer which uses non-linear downfiltering to reduce spatial size of representation and control overfitting
- 2012 an error rate of 0.23% on the MNIST database for handwritten digits was reported
- one of the three CNNs used in Alpha Go was a 13-layer CNN
- Apple's Siri uses CNNs for language recognition



3D layer, e.g. for width, height and color (RGB)



Maximum pooling

Deep learning frameworks

- **caffe** by the Berkeley Vision and Learning Center (C++)
 - **Keras** by François Chollet, a Google engineer (Python)
 - **Microsoft Cognitive Toolkit** by Microsoft (C++, Java)
 - **MXNet** by Apache Incubator (C++, Python, R)
 - **Tensorflow** by Google (Python)
-
- packages: `caffeR`, `keras`, `mxnet`, `tensorflow`
 - package on github: [Microsoft/CNTK-R](#)

Chollet, F. and Allaire, J.J. (2018). *Deep Learning with R*. O'Reilly Media. isbn: 978-1-61729-554-6.