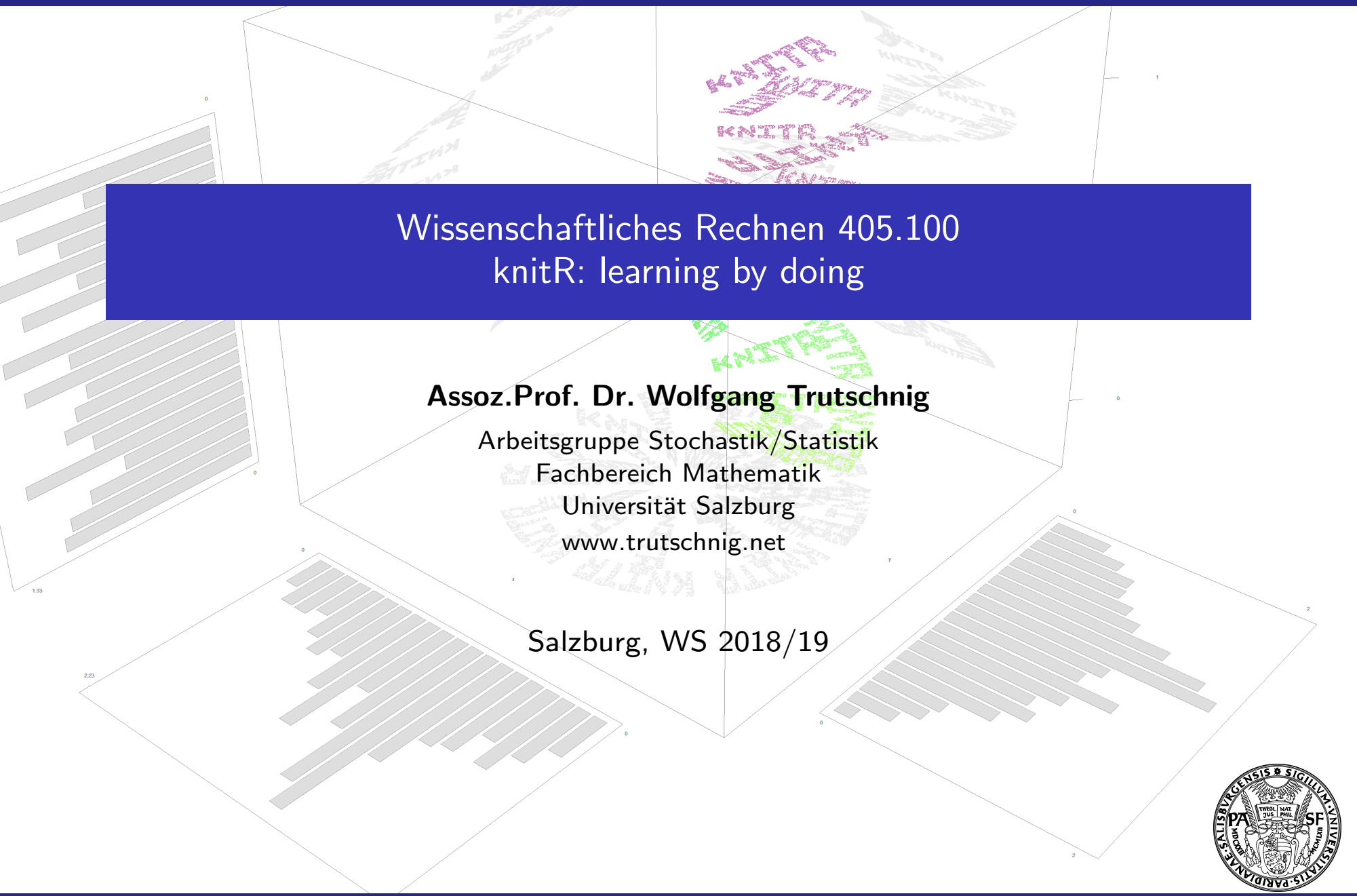


Wissenschaftliches Rechnen 405.100
knitR: learning by doing

Assoz.Prof. Dr. Wolfgang Trutschnig

Arbeitsgruppe Stochastik/Statistik
Fachbereich Mathematik
Universität Salzburg
www.trutschnig.net

Salzburg, WS 2018/19

Starting point

- Very common situation (in research and in the economy):
- An elaborate report including graphics, tables and text has been produced (using Excel, Word, Powerpoint etc.)
- New data arrives, often on a weekly/monthly basis, or new data is added to an existing dataset
- The same report is needed for the new/updated data
- **Do all calculations again, prepare all the tables and graphics again, and reproduce the report?**
- Apart from wasting time and money, which other problems occur?
  - Copy and paste is dangerous...
  - Small (copy and paste) mistakes can have far-reaching consequences
  - Can you assure to execute exactly the same manual data analysis steps as last time (Excel, etc.)?
- Reproducible results?

A possible solution - knitR

- Combine the power and flexibility of R with the typesetting capabilities of LaTeX
- R-Studio (user-interface for R) supports knitR
- **R, R-Studio and LaTeX are freeware** and run on all standard platforms!
- Let knitR do all the calculations, generate the graphics and tables and include numbers into the text
- Invest time in the first creation of the report - save time in all subsequent runs
- Use the saved time for more important and less annoying things than copy and paste (which robots can do much better than we)...

Plan for today:

- Get knitR running on your laptops
- Run a minimal example
- Understand the basic building blocks of knitR
- Try to understand a pre-prepared knitR report
- Adjust/Manipulate/Extend the pre-prepared knitR report

knitR has to be learned hands-on!

- Download and install the necessary software: (Basic) Miktex, Texmaker, R, R-Studio (in this order)

- Each knitR-file has the extension **.Rnw**
- Each knitR-file consists of **two building blocks**:
  1. LaTeX code
  2. R-Code
- There are **two types of R-Code**:
  1. **chunks**, i.e. code as separate paragraph (produce tables, graphics, etc.)
     Each chunk starts with `<<some options>>=` and ends with `@`
  2. inline code, using the command Sexpr
- Knowing (basic) LaTeX and R is enough - the rest is combining blocks
- We start with a first minimal example demonstrating the LaTeX & R structure

```
\documentclass{article}
\begin{document}


<<histo, fig.width=10, fig.height=6, fig.cap='Histogram', echo=FALSE
    >>=
x<-rnorm(100,0,1)
hist(x, probability = TRUE, col="lightblue")
@

The estimates for $\mu$ and $\sigma$ are given by $\overline{x}_n=\
    Sexpr{mean(x)}$
and $s^2_n=\Sexpr{sd(x)}$, whereby $n=\Sexpr{length(x)}$.

\end{document}
```

knitR_mini.Rnw

- Which part is LaTeX, which part is R?
- Download the file knitR_mini.Rnw from www.trutschnig.net/courses, save it in a new folder, and open it with R-Studio

- A button **Compile PDF** appears
- Clicking it has the following effect
- knitR 'translates' the R-Codes into standard LaTeX code and produces a .tex-file
- Plots produced are, by default, saved as pdfs in a folder named 'figure'
- the .tex file is then compiled (pdfLaTeX) and a final pdf is produced
- All files are saved in the folder where the knitR-file is located

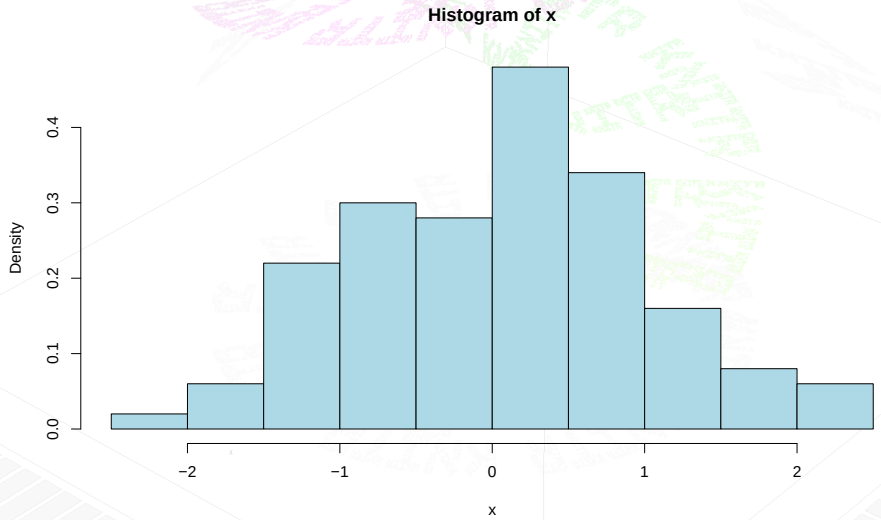


Figure 1: Histogram

The estimates for μ and σ are given by $\overline{x}_n = 0.0479857$ and $s_n^2 = 0.9382978$, whereby $n = 100$.

Learning by doing - exercise(s)

Exercise 1:

- Manipulate/Extend knitR_mini.Rnw in such a way that the resulting pdf looks like knitR_mini_extended.pdf
  Hint: par(mfrow = c(2,1)) can be used to have two plots in one graphic

```
%example loads the ATM data set, aggregates and produces a small
    summary
%install R-packages and (basic) miktex first

%Block 1: basic LaTeX settings

\documentclass[12pt]{article}
\usepackage{amsmath}
\usepackage{graphicx}
\usepackage{hyperref}
\usepackage{eurosym}
\usepackage{color}
\usepackage{float}
\setlength{\textheight}{650pt}
\setlength{\textwidth}{480pt}
\hoffset = -15mm

\begin{document}
```

knitR_ATM/knitR_ATM.Rnw

```
1
2  %Block 2: basic R setup: load all required packages
3  <<setup, include=FALSE, cache=FALSE>>=
4  library(knitr)
5  library(ggplot2)
6  library(doBy)
7  library(gridExtra)
8  library(RColorBrewer)
9  library(xtable)
10 Sys.setlocale("LC_TIME","English")     #set English
11 # set global chunk options
12 opts_chunk$set(fig.path='figure/graphic-', fig.align='center', fig.
       pos='!ht',echo=FALSE,warning = FALSE)
13           #global options for produced figures (name of figures,
                automatic centering, etc.)
14           #echo=FALSE: don't include R-Code in output,
15           #warning=FALSE: print warning in console but not in pdf
16           #fig.pos='!ht': place it here in the doc
17 a<-Sys.time()
18 @
```

knitR_ATM/knitR_ATM.Rnw

```
1 \title{\vspace{-4cm} ATM withdrawals\footnote{This report was
      created on \Sexpr{a}}}
2 \author{Wolfgang Trutschnig}
3     %The footnote prints the acutal time calculated in the chunk
          above
4
5 \maketitle
6
7 \section{Quick overview}
```

knitR_ATM/knitR_ATM.Rnw

```
<<results='asis'>>=
#Download ATM.txt and include table of first six rows in output
A<-read.table("http://www.trutschnig.net/Datensatz.txt",head=TRUE)
A$ymd<-as.Date(A$ymd)
A$month<-as.numeric(substr(A$ymd,6,7))
A$year<-substr(A$ymd,1,4)
beg<-min(A$ymd); end<-max(A$ymd)
mis<-nrow(subset(A,is.na(A$sum_out)==1))
H<-subset(A,A$holiday==1)
V<-subset(A,A$holiday==0.5)
B<-A[1:6,1:5]
B$ymd<-as.character(B$ymd)
print(xtable(B,label="taba",caption="First six lines of the dataset
    "),size="footnotesize",include.rownames=FALSE)
#NB: xtable prepares table output for LaTeX
@
```

knitR_ATM/knitR_ATM.Rnw

```
1 The dataset (see Table \ref{taba}) contains daily withdrawn amounts
      in the period from \Sexpr{beg} till \Sexpr{end} (\Sexpr{mis}
      entries are missing).
2 We expect weekdays and holidays to have a strong influence on
      withdrawn amounts and, additionally, to see an impact of the
      financial crisis starting with autumn 2008.
3 Figure \ref{fig:boxplot}, Figure \ref{fig:boxplot_monthly} as well
      as Table \ref{tabb} confirm this suspicion.
```

knitR_ATM/knitR_ATM.Rnw

```
<<boxplot, fig.width=13, fig.height=6, fig.cap=paste('Boxplot per day
      of week and year, the medians are also printed in Table \\ref{
    tabb}')>>=
  #size of the plot (width and height) in inches, NB: default output
      width is \textwidth
p <- ggplot(data=A, aes(x=factor(nr_weekday), y=sum_out, fill=factor(nr
    _weekday)))
p <- p + geom_boxplot(outlier.size=0)
p <- p + facet_wrap(~year)
p <- p + xlab("weekday")
p <- p + scale_fill_discrete(name = "Weekday")
p <- p + geom_point(data=H, colour="red", size=1.5)
p <- p + geom_point(data=V, colour="blue", size=1.5)
p <- p + theme_bw()
p
  #NB: label of figure is automatically generated as fig:boxplot
@
```

knitR_ATM/knitR_ATM.Rnw

```
<<results='asis'>>=
medna<-function(x){median(x[is.na(x)==0])}
    #little function calculating the median (ignoring missing values)
BB<-summaryBy(data=A,sum_out~nr_weekday,FUN=c(medna))
    #calculate meadian withdrawn amount per weekday
WD<-A[1:7,2:3]
BB<-merge(BB,WD)
names(BB)[2]<-"sum_out"
BB$sum_out<-round(BB$sum_out)
@
```

knitR_ATM/knitR_ATM.Rnw

```
Considering all years together the median withdrawn amount is \
    Sexpr{BB$sum_out[1]} \euro{} on Mondays, \Sexpr{BB$sum_out[2]} \
    euro{} on Tuesdays, \Sexpr{BB$sum_out[3]} \euro{} on Wednesdays,
\Sexpr{BB$sum_out[4]} \euro{} on Thursdays, \Sexpr{BB$sum_out[5]} \
    euro{} on Fridays, \Sexpr{BB$sum_out[6]} \euro{} on Saturdays,
    and \Sexpr{BB$sum_out[7]} \euro{} on Sundays.
```

knitR_ATM/knitR_ATM.Rnw

```
<<results='asis'>>=
AA<-summaryBy(data=A,sum_out~year+nr_weekday,FUN=c(medna))
WD<-A[1:7,2:3]
AA<-merge(AA,WD)
AA<-AA[order(AA$nr_weekday,AA$year),]
names(AA)[3]<-"sum_out"
AA<-subset(AA,select=c(weekday,year,sum_out))
print(xtable(AA,label="tabb",caption="Median withdrawn amount per
      year and day of week"),size="scriptsize",include.rownames=FALSE)
@
```

knitR_ATM/knitR_ATM.Rnw

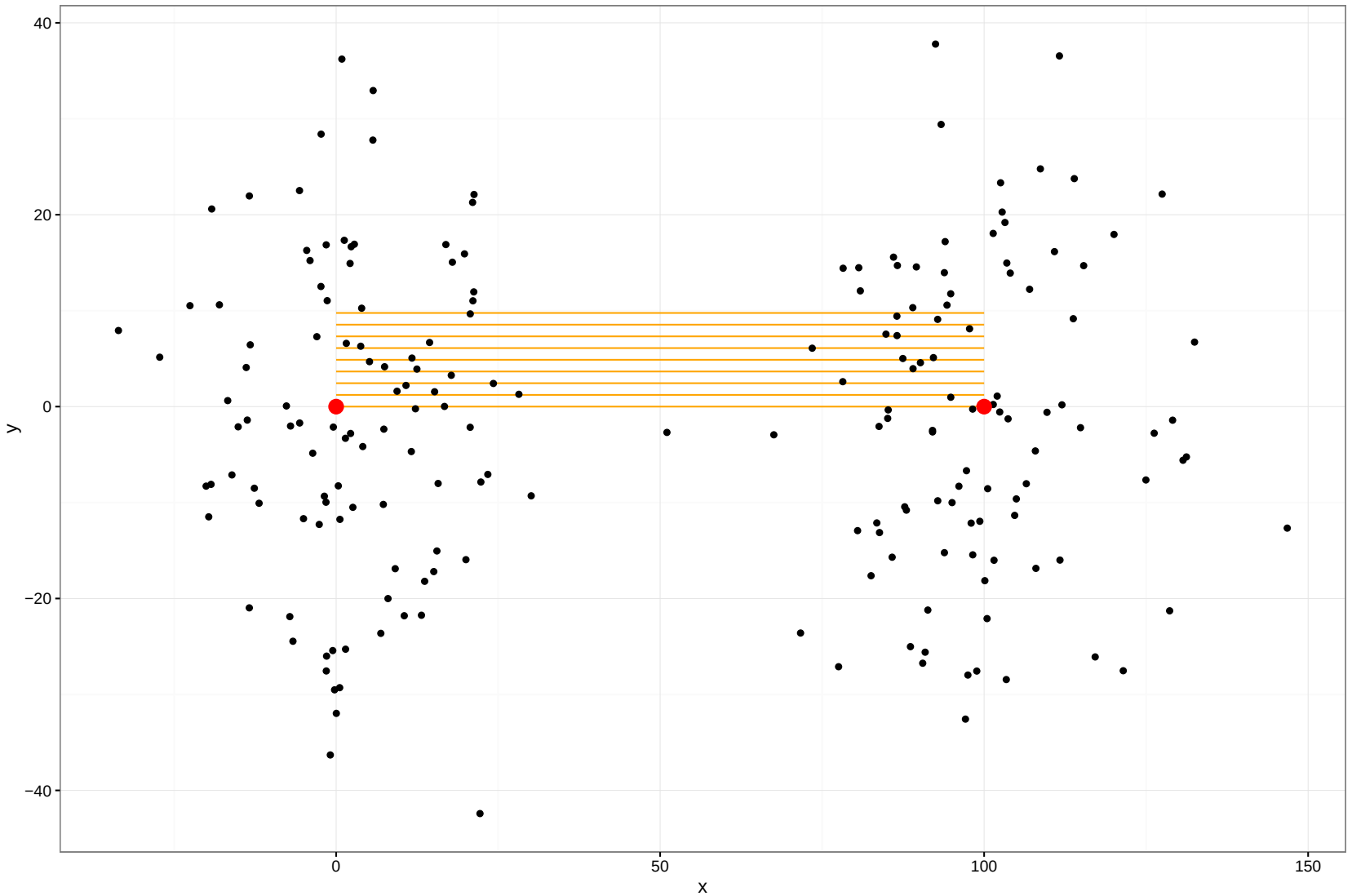
```
<<boxplot_monthly, fig.width=13, fig.height=6, fig.cap='Boxplot per
    month, the impact of the financial crisis starting with autumn
    2008'>>=
B<-A
B$year<-as.factor(A$year)
B$month<-as.factor(A$month)
farben<-c("gray50","magenta","green")
p <- ggplot(data=B,aes(x=month,y=sum_out))
p <- p + geom_boxplot(aes(fill=year),outlier.size=0)
p <- p + scale_fill_manual(values=farben)
p <- p + geom_jitter(colour="gray30")
p <- p + theme_bw()
p
@

\end{document}
```

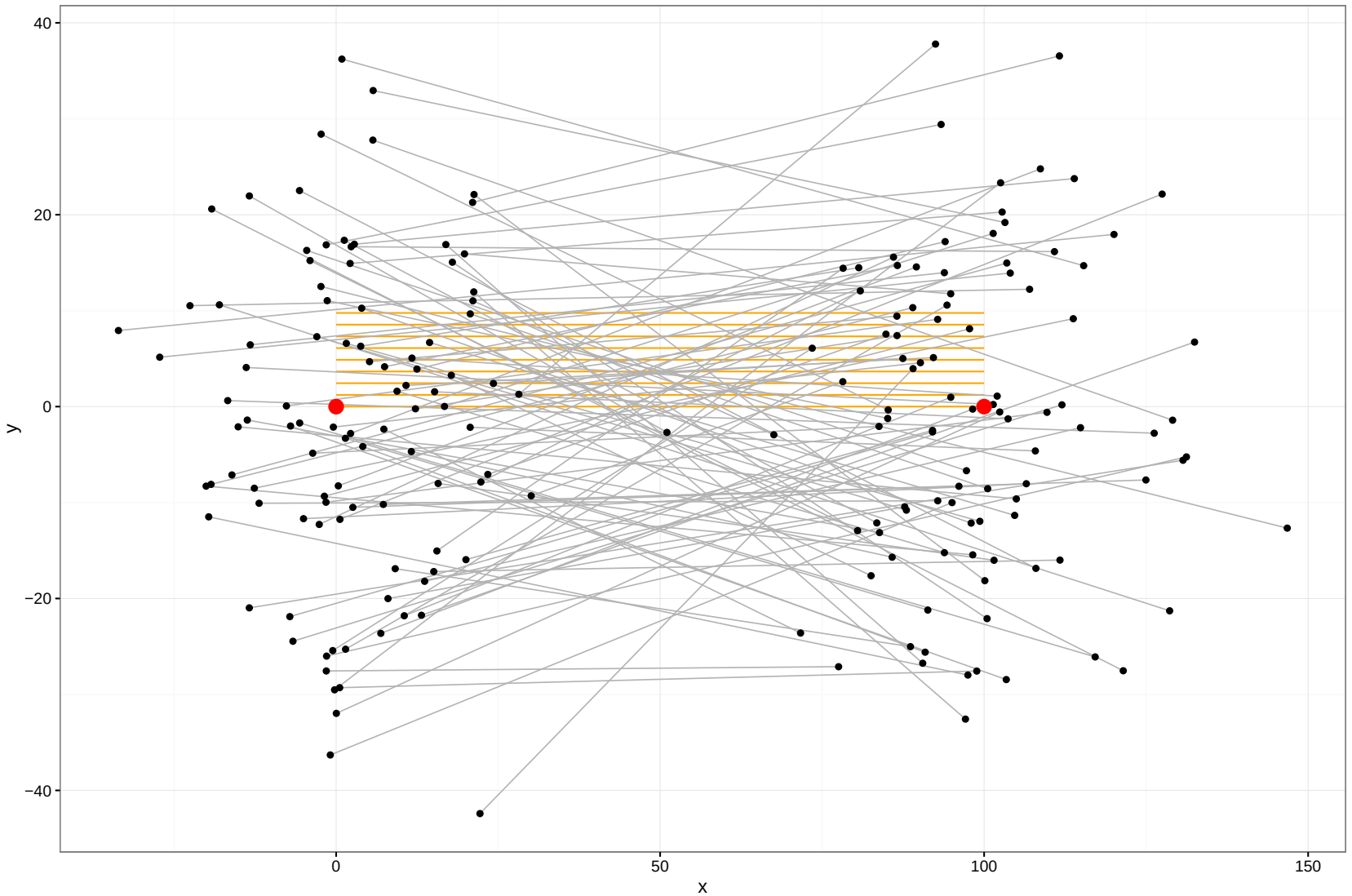
knitR_ATM/knitR_ATM.Rnw

- Ten students of geoinformatics want to test GPS-based distance measurements
- They (consecutively) record the GPS-coordinates of (the outer track of) the 100m starting line in an athletics stadium close by, then (consecutively) walk along the outer track till the finishing line, and again record the GPS-coordinates.
- Each of them repeats this procedure 50 times
- For each of the 500 pairs they calculate the distance in meters
- Given the sample size of $n = 500$ they expect the mean distance to be pretty close to 100m (why?).
- All the bigger the surprise when the mean distance turns out to be roughly 102m
- What went wrong - just bad luck?

A small simulation study using knitR: GPS bias?



A small simulation study using knitR: GPS bias?



- What went wrong - just bad luck?
- We answer the question by means of simulations
- W.l.o.g. we assume that the starting point S and the end point Z have the following exact coordinates: $S = (0, 0), Z = (100, 0)$
- S', Z' will denote the measured coordinates; $F = (X_1, Y_1)$ denotes the measurement error in S, $G = (X_2, Y_2)$ the measurement error in Z.
- In other words

$$\begin{aligned} S' &= S + (X_1, Y_1) = (X_1, Y_1) \\ Z' &= Z + (X_2, Y_2) = (100 + X_2, Y_2) \end{aligned}$$

- The measured distance therefore given by

$$\|S' - Z'\|_2 = \sqrt{(100 + X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

- To simplify matters we assume that the errors follow a normal distribution, i.e. $X_1, X_2, Y_2, Y_2 \sim \mathcal{N}(0, \sigma^2)$.
- Consider the case $\sigma^2 = 15$

- We simulate 10.000 (or more) distance measurements (maybe 500 was not big enough)
- Use the code 'R-Code_GPS.R' auf http://www.trutschnig.net/courses to do the simulations
- We really get 102,2 as mean distance, i.e. a bias of 2,2.

Exercise 2:

- Extend R-Code_GPS.R to a knitR report summarizing the simulations
- Include a boxplot of the distances
- Include the resulting bias
- Include a possible explanation how this overestimation could be explained
- Analyze what happens if σ^2 is increased or reduced (in a separate section)

Exercise 3:

- Produce a nice knitR report for the RTR dataset. The report should include (minimum requirements):
  - At least three graphics (use the ones you already have from previous exercises)
  - At least three tables (including some summary statistics)
  - Some text informing about what the RTR-dataset is about
  - Some values calculated from the data included in the text via Sexpr.