

PhD Workshop Series in
Statistics and Applied Data Science (DSP.005)
part two

knitr

Ass.-Prof. Dr. Wolfgang Trutschnig

Arbeitsgruppe Stochastik/Statistik

Fachbereich Mathematik

Universität Salzburg

www.trutschnig.net

Salzburg, December 2016



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



```

1 \documentclass{article}
2 \begin{document}
3
4
5 <<histo,fig.width=10,fig.height=6,fig.cap='Histogram',echo=FALSE
  >>=
6 x<-rnorm(100,0,1)
7 hist(x,probability = TRUE,col="lightblue")
8 @
9
10 The estimates for  $\mu$  and  $\sigma$  are given by  $\overline{x}_n = \text{Sexpr}\{\text{mean}(x)\}$ 
11 and  $s^2_n = \text{Sexpr}\{\text{sd}(x)\}^2$ , whereby  $n = \text{Sexpr}\{\text{length}(x)\}$ .
12
13 \end{document}

```

knitrR_mini.Rnw

- ▶ Which part is LaTeX, which part is R?



- ▶ Download the file `knitrR_mini.Rnw` from www.trutschnig.net/courses, save it in a new folder, and open it with R-Studio
- ▶ A button **Compile PDF** appears (if not, something is not installed properly)
- ▶ Clicking it has the following effect
- ▶ knitrR '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 knitrR-file is located



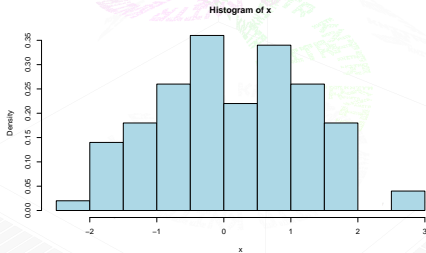


Figure 1: Histogram

The estimates for μ and σ are given by $\bar{x}_n = 0.1099123$ and $s_n^2 = 1.0648799$, whereby $n = 100$.



Learning by doing - two exercises

Exercise 1:

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

Exercise 2:

- ▶ Download the file `knitrR_ATM.Rnw` from www.trutschnig.net/courses, save it in a new folder, and open it with R-Studio
- ▶ Install all R-packages loaded at the beginning of the first chunk and press 'Compile pdf'
- ▶ Work through the `.Rnw` and find out what the Code does
- ▶ Replace the `ggplot`-graphics by analogous graphics using the standard `boxplot` command
- ▶ Extend the report such that the output looks like `knitrR_ATM_extended.pdf`



```
1 %example loads the ATM data set, aggregates and produces a small  
2 summary  
3  
4 %Block 1: basic LaTeX settings  
5  
6 \documentclass[12pt]{article}  
7 \usepackage{amsmath}  
8 \usepackage{graphicx}  
9 \usepackage{hyperref}  
10 \usepackage{eurosym}  
11 \usepackage{color}  
12 \usepackage{float}  
13 \setlength{\textheight}{650pt}  
14 \setlength{\textwidth}{480pt}  
15 \hoffset = -15mm  
16  
17 \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.
13   pos='!ht', echo=FALSE, warning = FALSE)
14   #global options for produced figures (name of figures,
15     automatic centering, etc.)
16   #echo=FALSE: don't include R-Code in output,
17   #warning=FALSE: print warning in console but not in pdf
18   #fig.pos='!ht': place it here in the doc
19 a<-Sys.time()
20 @
```

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



```
1 <<results='asis'>>=
2 #Download ATM.txt and include table of first six rows in output
3 A<-read.table("http://www.trutschnig.net/Datensatz.txt",head=TRUE)
4 A$ymd<-as.Date(A$ymd)
5 A$month<-as.numeric(substr(A$ymd,6,7))
6 A$year<-substr(A$ymd,1,4)
7 beg<-min(A$ymd); end<-max(A$ymd)
8 mis<-nrow(subset(A, is.na(A$sum_out)==1))
9 H<-subset(A,A$holiday==1)
10 V<-subset(A,A$holiday==0.5)
11 B<-A[1:6,1:5]
12 B$ymd<-as.character(B$ymd)
13 print(xtable(B,label="taba",caption="First six lines of the dataset
14   "),size="footnotesize",include.rownames=FALSE)
15 #NB: xtable prepares table output for LaTeX
16 @
```

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



```

1 <<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}')>>=
2 #size of the plot (width and height) in inches, NB: default output
  width is \textwidth
3 p <- ggplot(data=A,aes(x=factor(nr_weekday),y=sum_out,fill=factor(nr
  _weekday)))
4 p <- p + geom_boxplot(outlier.size=0)
5 p <- p + facet_wrap(~year)
6 p <- p + xlab("weekday")
7 p <- p + scale_fill_discrete(name="Weekday")
8 p <- p + geom_point(data=H,colour="red",size=1.5)
9 p <- p + geom_point(data=V,colour="blue",size=1.5)
10 p <- p + theme_bw()
11 p
12 #NB: label of figure is automatically generated as fig:boxplot
13 @

```

knitr_ATM/knitr_ATM.Rnw



```

1 <<results='asis'>>=
2 medna<-function(x){median(x[is.na(x)==0])}
3   #little function calculating the median (ignoring missing values)
4 BB<-summaryBy(data=A,sum_out~nr_weekday,FUN=c(medna))
5   #calculate meadian withdrawn amount per weekday
6 WD<-A[1:7,2:3]
7 BB<-merge(BB,WD)
8 names(BB)[2]<-"sum_out"
9 BB$sum_out<-round(BB$sum_out)
10 @

```

knitr_ATM/knitr_ATM.Rnw



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

knitr_ATM/knitr_ATM.Rnw



```

1 <<results='asis '>>=
2 AA<-summaryBy( data=A, sum_out~year+nr_weekday ,FUN=c( medna ))
3 WD<-A[1:7 ,2:3]
4 AA<-merge( AA,WD)
5 AA<-AA[ order( AA$nr_weekday ,AA$year ) ,]
6 names( AA ) [3]<- "sum_out"
7 AA<-subset( AA, select=c( weekday ,year ,sum_out ))
8 print( xtable( AA, label=" tabb" ,caption=" Median withdrawn amount per
year and day of week" ) ,size=" scriptsize" ,include.rownames=FALSE)
9 @

```

knitr_ATM/knitr_ATM.Rnw



```
1 <<boxplot_monthly, fig.width=13, fig.height=6, fig.cap='Boxplot per
  month, the impact of the financial crisis starting with autumn
  2008 '>>=
2 B<-A
3 B$year<-as.factor(A$year)
4 B$month<-as.factor(A$month)
5 farben<-c("gray50", "magenta", "green")
6 p <- ggplot(data=B, aes(x=month, y=sum_out))
7 p <- p + geom_boxplot(aes(fill=year), outlier.size=0)
8 p <- p + scale_fill_manual(values=farben)
9 p <- p + geom_jitter(colour="gray30")
10 p <- p + theme_bw()
11 p
12 @
13
14 \end{document}
```

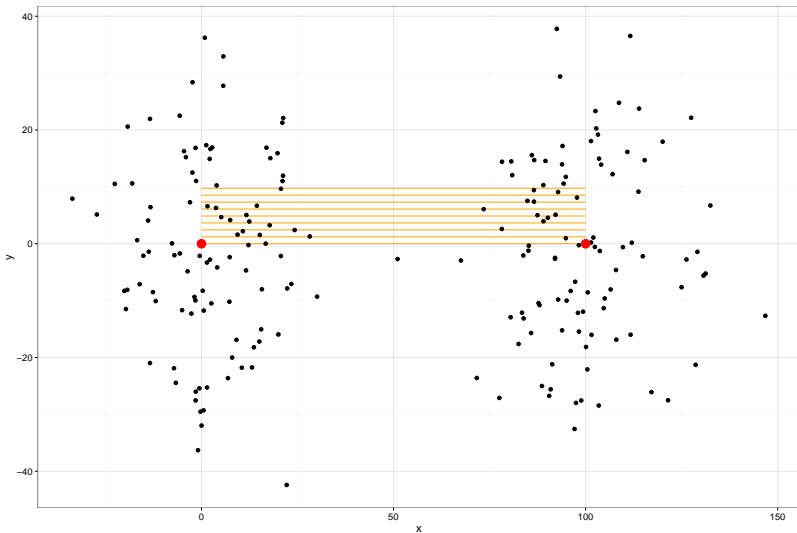
knitr_ATM/knitr_ATM.Rnw



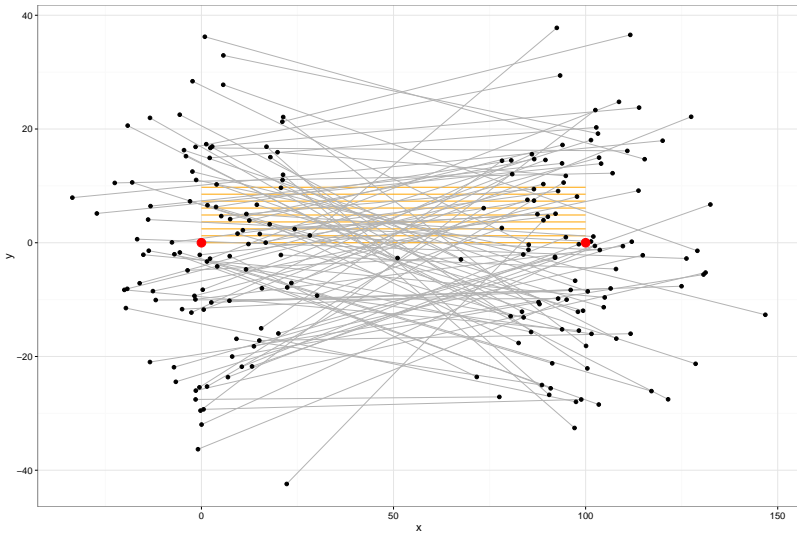
- ▶ Ten students of geoinformatics want to test GPS-based distance measurements
- ▶ They (consecutively) record the GPS-coordinates of (the outer track of) the 100m starting line in an athletics stadium close by, then (consecutively) walk along the outer track till the finishing line, and again record 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

$$S' = S + (X_1, Y_1) = (X_1, Y_1)$$

$$Z' = Z + (X_2, Y_2) = (100 + X_2, Y_2)$$

- ▶ 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 3:

- ▶ Extend knitR_GPS.Rnw to a report like knitR_GPS_extended.Rnw
- ▶ Include a possible explanation of the observed GPS-bias

