R Basics and Examples - A short introduction

Markus Baaske

Faculty of Mathematics and Computer Science (FSU Jena)

May 7, 2018

The R Project for Statistical Computing

The R project `http://www.r-project.org` develops a free software environment for statistical computing and graphics. R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS, is mostly used for statistics but can also be used as a programming (script) language alone.
R is organized as a core distribution of base packages which can be extended by further packages loaded into the a user workspace (or interpreter global environment). Some useful links are

- Tutorials on using R can be found at `http://www.r-tutor.com/`
- Meta search and package documentation `https://www.rdocumentation.org/`
- R CRAN repository for contributed packages: `https://cran.r-project.org/`
- A short reference card `https://cran.r-project.org/doc/contrib/Short-refcard.pdf`

R Practice

1. Generate standard normal random variables, check characteristics and use different sample sizes!
2. Construct a numeric vector and sort this in ascending order using function *sort()*.
3. Delete first row of data frame ‘g‘. Add some new category. Change number of students ‘Anz‘ in in ‘MPV‘.
4. Load data set *airquality*:

   ```
   > data(airquality)
   > head(airquality)
   ```

   Check for ‘NAs‘ and omit these by functions *is.na()* and *na.omit()* What is it about? Write a function which calculates the means and standard deviations of categories ‘Ozone‘,‘Solar.R‘, ‘Wind‘ and ‘Temp‘ from data set *airquality* for each month separately. THe function returns a matrix of dimensions 5×2 where the first columnstores the means and the second the standard errors of all month.

R Practice - solutions

1. ```
   >   gg <- rnorm(n=100)
   >   s <- sample(gg,size=20)
   >   summary(gg)
   ```
   ```
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -2.9055 -0.5283  0.1666  0.1327  0.8187  2.1925
   ```
   ```
   >   summary(s)
   ```
   ```
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -1.6818 -0.4303  0.2663  0.2677  0.8671  2.1925
   ```
2. ```
   > # sort log normal random variables
   > x <- rlnorm(10,0,1)
   > sort(x)
   ```
   ```
    [1] 0.2427637 0.2945478 0.3115378 0.3797676 0.5420980 0.546
    [8] 1.0269044 1.2180035 3.3641965
   ```
3. ```
   > g <- data.frame(StG=c("GTB","MPV","BGM"),Anz=c(75,11,62))
   > z <- g[-1,]
   > z <- cbind(z,Sex=c(10,12))
   > z$Anz[1] <- 20
   > z <- rbind(data.frame(StG="BGOK",Anz=57,Sex=3),z)
   ```
   ```

Collect characteristics

Example: get means and standard deviations from 'airquality'

```
> # load dataset
> data(airquality)
> t(                              # transpose final matrix
+  apply(airquality[1:4],2,       # iterate over columns 1 to 4
+  function(x){                   # and execute for MARGIN=2
+   c("means"=mean(na.omit(x)),"errs"=sd(na.omit(x)))
+  }))

             means      errs
Ozone    42.129310 32.987885
Solar.R 185.931507 90.058422
Wind      9.957516  3.523001
Temp     77.882353  9.465270
```

R Practice - data frames

> head(data.frame)

```
1 function (..., row.names = NULL, check.rows = FALSE, check.nam
2     fix.empty.names = TRUE, stringsAsFactors = default.strings
3 {
4     data.row.names <- if (check.rows && is.null(row.names))
5         function(current, new, i) {
6             if (is.character(current))
```

> print(z)

```
   StG  Anz  Sex
1 BGOK   57    3
2  MPV   20   10
3  BGM   62   12
```

> str(z)

```
'data.frame':        3 obs. of  3 variables:
 $ StG: Factor w/ 4 levels "BGOK","BGM","GTB",..: 1 4 2
 $ Anz: num  57 20 62
 $ Sex: num  3 10 12
```

R Practice - R data set ‘airquality‘

```
> data(airquality)
> head(airquality)
```

```
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

```
Format:
 A data frame with 154 observations on 6 variables.
‘[,1]’  ‘Ozone’    numeric  Ozone (ppb)
‘[,2]’  ‘Solar.R’  numeric  Solar R (lang)
‘[,3]’  ‘Wind’     numeric  Wind (mph)
‘[,4]’  ‘Temp’     numeric  Temperature (degrees F)
‘[,5]’  ‘Month’    numeric  Month (1-12)
‘[,6]’  ‘Day’      numeric  Day of month (1-31)
```

R Practice - R data set ‘airquality‘

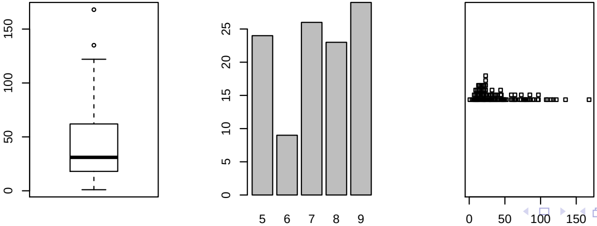
```
> # First check on `missing values`
> sum(is.na(airquality))       # count number of `NAs`
[1] 44
> air <- na.omit(airquality)   # omit these values
> attach(air)                  # attach variables to environment
> c(nrow(air),ncol(air))       # dimensions of the data
[1] 111   6
> any(is.na(air))              # nothing left
[1] FALSE
> # find Ozone values where *Temp* is in between 50-60
> air[(Temp > 50 & Temp < 60),]
   Ozone Solar.R Wind Temp Month Day
8     19      99 13.8   59     5   8
15    18      65 13.2   58     5  15
18     6      78 18.4   57     5  18
21     1       8  9.7   59     5  21
> range(Wind)                  # numeric range
[1]  2.3 20.7
```

R Practice - R data set ‘airquality‘

1. Are there bindings (equal data values) in any of the variables?
2. Do you see any outliers?
3. How many measurements are recorded for each month?

R Practice - R data set ‘airquality‘

```
> op <- par(mfrow=c(2,3))
> class(Ozone)
[1] "integer"
> boxplot(Ozone)
> table(Month)
Month
 5  6  7  8  9
24  9 26 23 29
> barplot(table(Month))
> stripchart(Ozone,method="stack")
> par(op)
```

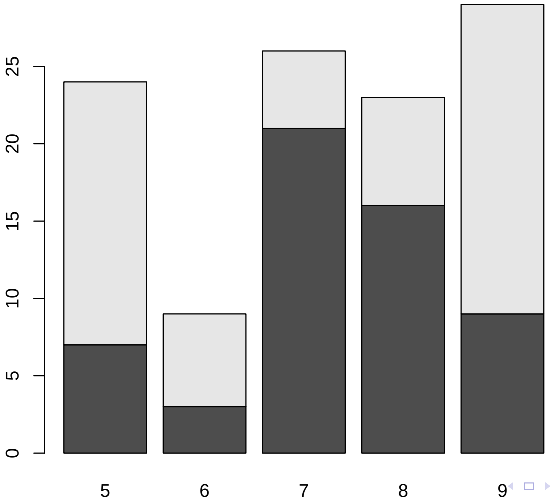


R Practice - R data set ‘airquality‘

1. Find abolute frequencies of measrurements of ‘Ozone‘ for each month where the ozone level drops below the median?
2. Find suitable graphical tools to show statistical relations between some categories. Use ‘*boxplot*‘ and generic ‘*plot*‘ function.

R Practice - R data set ‘airquality‘

```
> # What does this figure show?
> barplot(table(Ozone < median(Ozone),Month)) # a two-way table
```

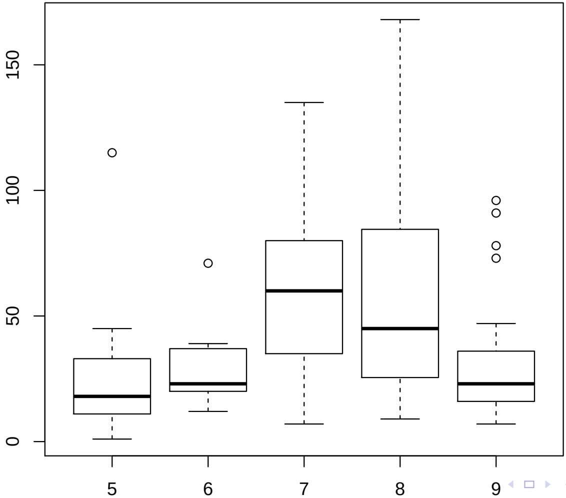


R Practice - R data set ‘airquality‘

1. Find suitable graphical tools to show statistical relations between some categories. Use ‘*boxplot*‘ and generic ‘*plot*‘ function.

R Practice - R data set ‘airquality‘

```
> # ozone level per month
> boxplot(Ozone~Month, data=air)
```

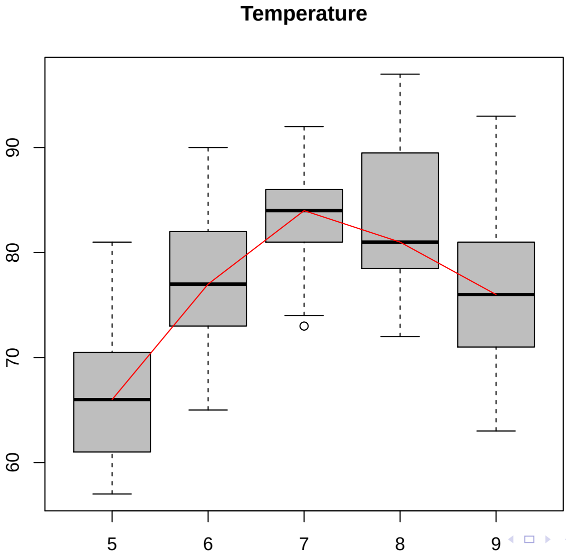


R Practice - R data set ‘airquality‘

1. Describe graphically the distribution of ‘Ozone‘ for each month and compare this to ‘Temp‘ also for each month. Use ‘*boxplot*‘ and ‘*lines*‘ function. What is your conclusion?

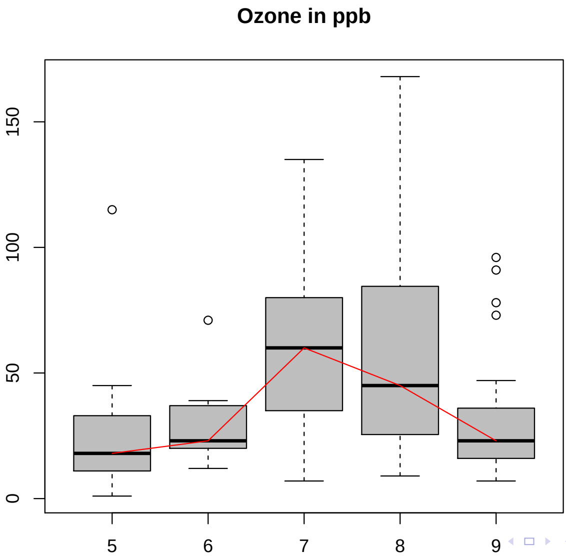
R Practice - R data set ‘airquality‘

```
> bpt <- boxplot(Temp~Month,data=air,col="gray",main="Temperatur
> lines(1:5, bpt$stats[3,], col="red")
```



R Practice - R data set 'airquality'

```
> bpt <- boxplot(Ozone~Month,data=air,col="gray",main="Ozone in
> lines(1:5, bpt$stats[3,], col="red")
```



R Practice - R data set ‘airquality‘

1. What kind of distribution has ‘Ozone‘ for the time of all measurements and in months June and July?

R Practice - R data set ‘airquality‘

```
> op <- par(mfrow=c(1,3),mar=c(6.1, 4.1, 1.1, 1.1),cex=3.0, cex.
> hist(Ozone, col="gray", xlab="Ozone in [ppb]",
+  ylab="freqency",main="all")
> hist(Ozone[Month==6], col="gray", xlab="Ozone in [ppb]",
+  ylab="freqency",main="June")
> hist(Ozone[Month==7], col="gray", xlab="Ozone in [ppb]",
+  ylab="freqency",main="July")
> par(op)
```