University of Zurich^UZH

Makerere University

Data Analysis with R:
Lecture Slides: Day 1 - Monday

Sonja Hartnack, Terence Odoch & Muriel Buri

22nd of July 2019

Goals of the course

To be able to...

- import data sets to R
- describe data with R
- apply basic statistical tests in R
- some ideas for more advanced statistical tools ...
- simulate a data set similar to own research

General remarks

Course schedule:

- Starting at 9:00am / 9:30am (?)
- Tea breaks in between
- Lunch break
- Teaching until 4.30pm ($\sim$ 5pm)

Obtaining a certificate is conditional on:

- active participation in class
- attending at least 75 % of the course (lecture & exercises)
- assignments during now and October
- short final exam in October (format to be defined)

Getting to know each other

- My name is ...
- I am doing a Master / a PhD in ...
- I hope to learn in this course how to ....
- My personal goal for this course is ...

How do we reach these goals

- hands on exercises with R:
  - `chickwts`
  - `ToothGrowth`
  - `bacteria`
  - `perulung`
  - ... and others.
- interactive discussions & student's present their own solutions
- ask us a lot of questions but also ask google for help!
- group work
- short motivational lectures

RStudio and R

Do you all have RStudio and R installed on your computers?

Get started with data set: chickwts

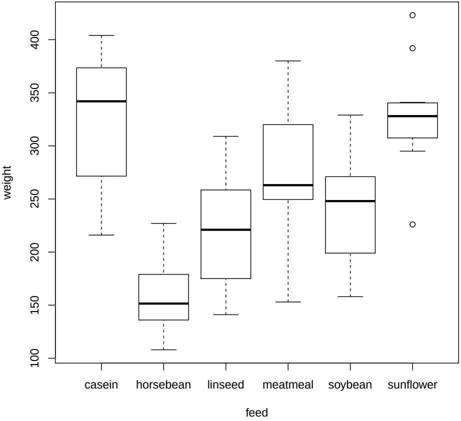
An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.

```r
# load data set "chickwts"
data("chickwts", package = "datasets")
# the head(...) function shows the first 6 observations
head(chickwts)

##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean

# FUNCTION - open bracket - DATA SET / VARIABLE - close bracket
```

Ideas for plotting the data

Ideas for plotting the data

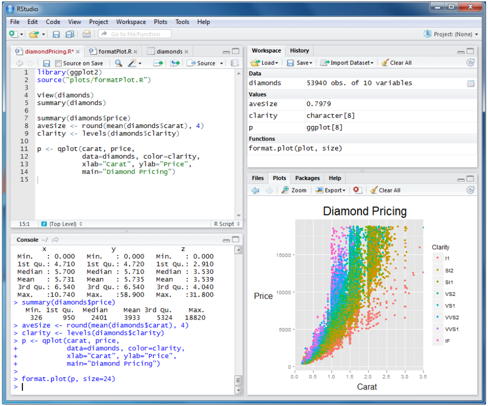
```r
# use x axis to show the categorical variable (feed),
# y axis to represent the continuous variable (weight)
# boxplot (y.cont.variable ~ x.cat.variable, data = dataset)
# ?boxplot
boxplot(weight ~ feed, data = chickwts)
```



Exercise: Statistical Terminologies

Functionalities in RStudio

- Source
- Console
- Environment, History, Files
- Files, Plots, Packages, Help



```
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")

```

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

```
Data
diamonds    53940 obs. of 10 variables
Values
aveSize     0.7979
clarity     character[8]
p           ggplot[8]
Functions
format.plot(plot, size)
```

Good housekeeping!

- Define manually a new folder called **rcourse** in your personal documents on your personal computer

- Know in which directory you are

```
getwd()

## [1] "/home/mburi/Documents/git_svn/DataAnalysisWithR/Lectures"
```

- Set directory path

```
# back- and forslash is dependent on the system
setwd("C:/Users/muriel/Documents/rcourse/")
setwd("C:\\Users\\muriel\\Documents\\rcourse\\")
```

- Always clean up before starting with new R-Script

```
rm(list=ls()) # empty workspace, delete previously saved variables
```

How to get help in R

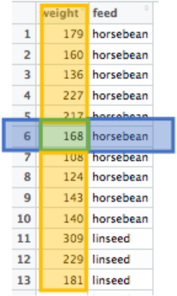
```
?chickwts
?boxplot
```

Also, have a look at the examples at the end of the help pages.

Exercise: Getting to know R and `chickwts`

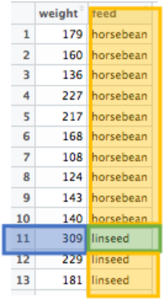
A data frame in R: `chickwts`

chickwts[ROWS , COLUMNS]



	weight	feed
1	179	horsebean
2	160	horsebean
3	136	horsebean
4	227	horsebean
5	217	horsebean
6	168	horsebean
7	108	horsebean
8	124	horsebean
9	143	horsebean
10	140	horsebean
11	309	linseed
12	229	linseed
13	181	linseed

chickwts[6, 1]



	weight	feed
1	179	horsebean
2	160	horsebean
3	136	horsebean
4	227	horsebean
5	217	horsebean
6	168	horsebean
7	108	horsebean
8	124	horsebean
9	143	horsebean
10	140	horsebean
11	309	linseed
12	229	linseed
13	181	linseed

chickwts[11, 2]

Rows and columns of a data frame: `chickwts`

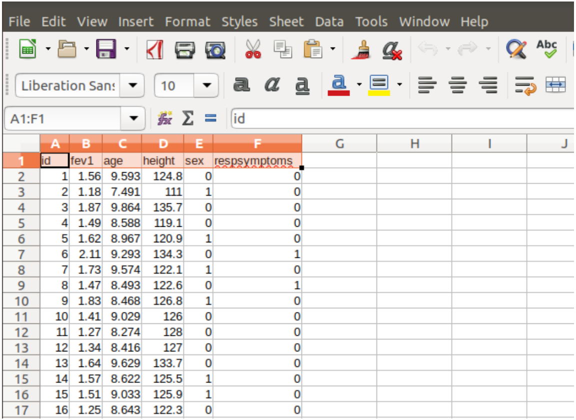
Values of ...

```r
# Load (internal) data set from R
data("chickwts")

# ... all columns of sixth observation:
chickwts[6, ]

# ... all columns of sixth to eleventh observation:
chickwts[c(6:11), ]

# ... all columns of sixth, eleventh and twentieth observation:
chickwts[c(6, 11, 20), ]

# ... all rows of first column (weight):
chickwts[ , 1]

# ... all rows of second column (feed):
chickwts[ , 2]

# or use the "$" sign as a reference to column "feed":
chickwts$feed
```

Exercise: Summary Statistics for the `chickwts` data set

Rules for importing data into R

- First row of excel sheet contains **variable names**: `y, ap, hilo, week, ID, trt.`
- Columns of excel sheet represent **variables**.
- Rows of excel sheet represent **observations per individual** (except for the first row).


File Edit View Insert Format Styles Sheet Data Tools Window Help
Liberation Sans
10
Abc
A1:F1
Σ
=
id

	A	B	C	D	E	F	G	H	I	J
1	id	fev1	age	height	sex	respsymptoms				
2	1	1.56	9.593	124.8	0	0				
3	2	1.18	7.491	111	1	0				
4	3	1.87	9.864	135.7	0	0				
5	4	1.49	8.588	119.1	0	0				
6	5	1.62	8.967	120.9	1	0				
7	6	2.11	9.293	134.3	0	1				
8	7	1.73	9.574	122.1	1	0				
9	8	1.47	8.493	122.6	0	1				
10	9	1.83	8.468	126.8	1	0				
11	10	1.41	9.029	126	0	0				
12	11	1.27	8.274	128	0	0				
13	12	1.34	8.416	127	0	0				
14	13	1.64	9.629	133.7	0	0				
15	14	1.57	8.622	125.5	1	0				
16	15	1.51	9.033	125.9	1	0				
17	16	1.25	8.643	122.3	0	0				

Rules for naming variables

Variable names should ...

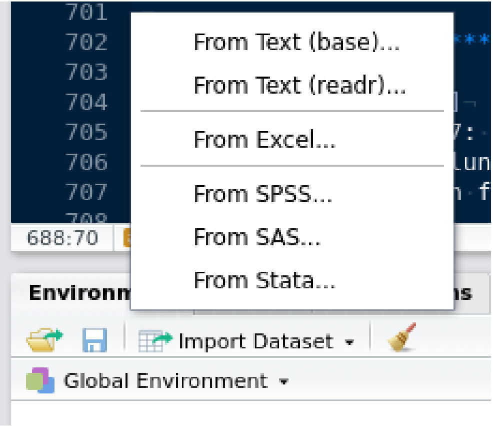
- start with a letter (not a number):
  `y, ap, hilo, week, ID, trt`

- longer variables names should be separated with dots:
  `time.in.weeks`

- do not use special characters, such as /, #, @, &, $\star$, ...

How to import external data files into R?

- > Import Dataset > **From Text (base)...** > CSV Files (.csv) or



How to import .txt and .csv files into R? (1/3)

- Environment (upper right corner)
- > Import Dataset > **From Text (base)...** > CSV Files (.csv)

```
perulung_ems <- read.csv("perulung_ems.csv", row.names = 1,
                         sep = ";")
lung <- data.frame(perulung_ems)
```

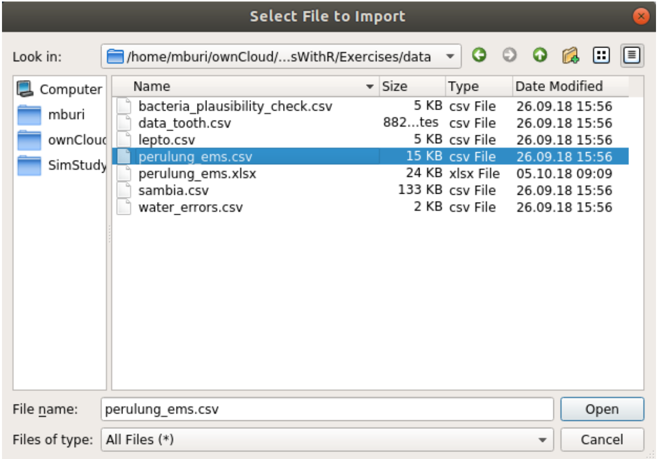
- > Import Dataset > **From Text (base)...** > Text Files (.txt)

How to import .txt and .csv files into R? (2/3)

- Environment (upper right corner)
- $>$ Import Dataset $>$ From Text (base)... $>$ CSV Files (.csv)

How to import .txt and .csv files into R? (2/3)

- Environment (upper right corner)
- > Import Dataset > From Text (base)... > CSV Files (.csv)



Select File to Import

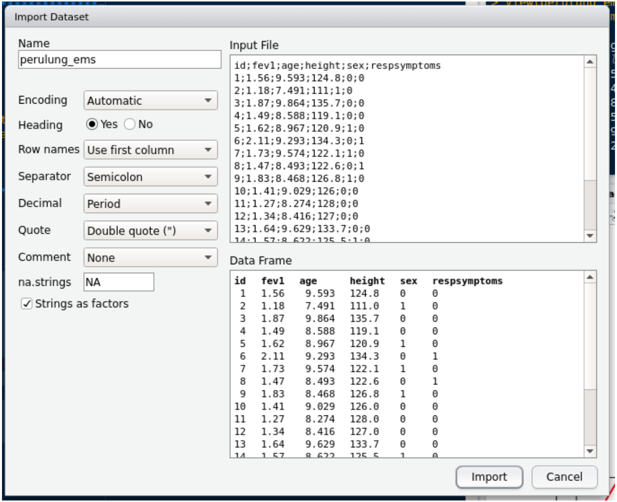
Look in: /home/mburi/ownCloud/...sWithR/Exercises/data

Name	Size	Type	Date Modified
bacteria_plausibility_check.csv	5 KB	csv File	26.09.18 15:56
data_tooth.csv	882...tes	csv File	26.09.18 15:56
lepto.csv	5 KB	csv File	26.09.18 15:56
perulung_ems.csv	15 KB	csv File	26.09.18 15:56
perulung_ems.xlsx	24 KB	xlsx File	05.10.18 09:09
sambia.csv	133 KB	csv File	26.09.18 15:56
water_errors.csv	2 KB	csv File	26.09.18 15:56

File name: perulung_ems.csv

Files of type: All Files (*)

How to import .txt and .csv files into R? (3/3)



Import Dataset

Name: perulung_ems

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Use first column

Separator: Semicolon

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NA

☑ Strings as factors

Input File

```
id;fev1;age;height;sex;respsymptoms
1;1.56;9.593;124.8;0;0
2;1.18;7.491;111;1;0
3;1.87;9.864;135.7;0;0
4;1.49;8.588;119.1;0;0
5;1.62;8.967;120.9;1;0
6;2.11;9.293;134.3;0;1
7;1.73;9.574;122.1;1;0
8;1.47;8.493;122.6;0;1
9;1.83;8.468;126.8;1;0
10;1.41;9.029;126;0;0
11;1.27;8.274;128;0;0
12;1.34;8.416;127;0;0
13;1.64;9.629;133.7;0;0
14;1.57;8.622;125.5;1;0
```

Data Frame

```
id  fev1   age     height  sex   respsymptoms
 1  1.56   9.593   124.8   0     0
 2  1.18   7.491   111.0   1     0
 3  1.87   9.864   135.7   0     0
 4  1.49   8.588   119.1   0     0
 5  1.62   8.967   120.9   1     0
 6  2.11   9.293   134.3   0     1
 7  1.73   9.574   122.1   1     0
 8  1.47   8.493   122.6   0     1
 9  1.83   8.468   126.8   1     0
10  1.41   9.029   126.0   0     0
11  1.27   8.274   128.0   0     0
12  1.34   8.416   127.0   0     0
13  1.64   9.629   133.7   0     0
14  1.57   8.622   125.5   1     0
```

```
perulung_ems <- read.csv("perulung_ems.csv", row.names = 1,
                         sep = ";")
lung <- data.frame(perulung_ems)
```

Exercise: Data import to R and summary statistics `perulung_ems.csv`

Data from a study of lung function among children living in a deprived suburb of Lima, Peru. Data taken from Kirkwood and Sterne, 2nd edition.
Variables:

- `fev1`: in liter, ”forced expiratory volume in 1 second” measured by a spirometer. This is the maximum volume of air which the children could breath out in 1 second
- `age`: in years
- `height`: in cm
- `sex`: 0 = girl, 1 = boy
- `respsymp`: respiratory symptoms experienced by the child over the previous 12 months