



**University of  
Zurich<sup>UZH</sup>**



MAKERERE UNIVERSITY

# **Data Analysis with R:**

## **Day 3 - Slides**

Sonja Hartnack, Terence Odoch & Muriel Buri

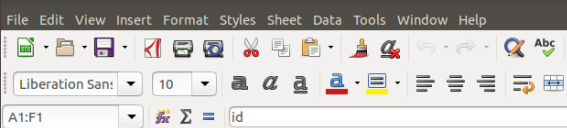
October 2018

## Examples of different data types

- numeric variable
- integer variable
- variable with two levels (binary factor)
- ordered variable with **more than** two levels (ordinal)
- unordered variable with **more than** two levels (nominal)

## Rules for importing data into R (from Excel)

- 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).



	A	B	C	D	E	F	G	H	I	J
1	id	fev1	age	height	sex	resp	symptoms			
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				

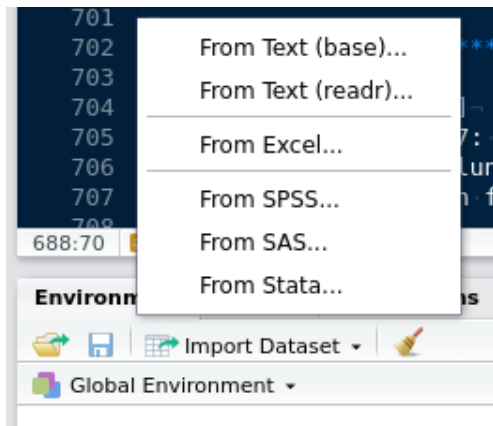
# Rules for naming variables

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 /, #, @, &, \*, ...

## How to import external data files into R?

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



# How to import external data files into R?

- 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)
- > Import Dataset > **From Excel...** > Excel Files (.xlsx)

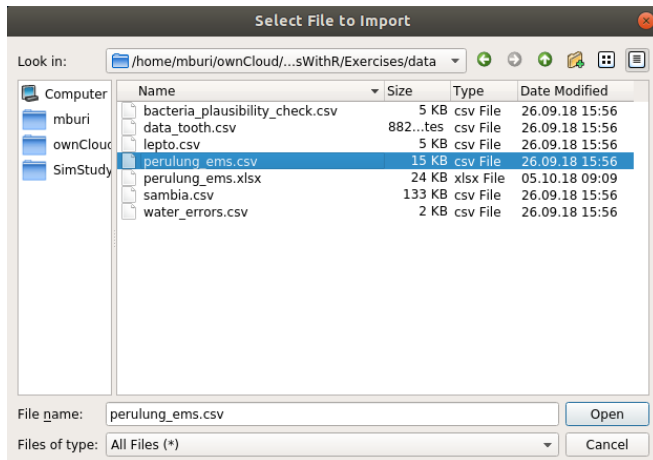
```
install.packages("readxl")  
library("readxl")  
perulung_ems <- read_excel("perulung_ems.xlsx")  
lung <- data.frame(perulung_ems)  
head(lung)
```

## How to import .txt and .csv files into R? (1/2)

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

## How to import .txt and .csv files into R? (1/2)

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





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

Import Dataset

Name: perulung\_ems

Input File: id;fev1;age;height;sex;respsymptoms

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

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

Import Cancel

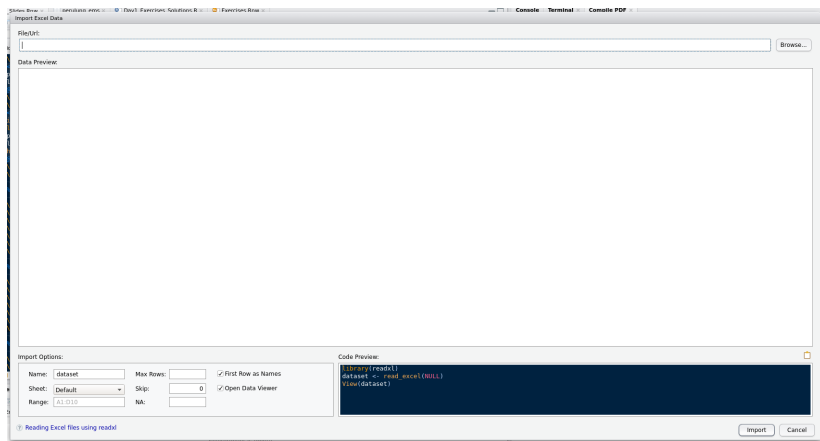
```
perulung_ems <- read.csv("perulung_ems.csv", row.names = 1,  
                        sep = ";")  
lung <- data.frame(perulung_ems)
```

## How to import .xlsx files into R? (1/3)

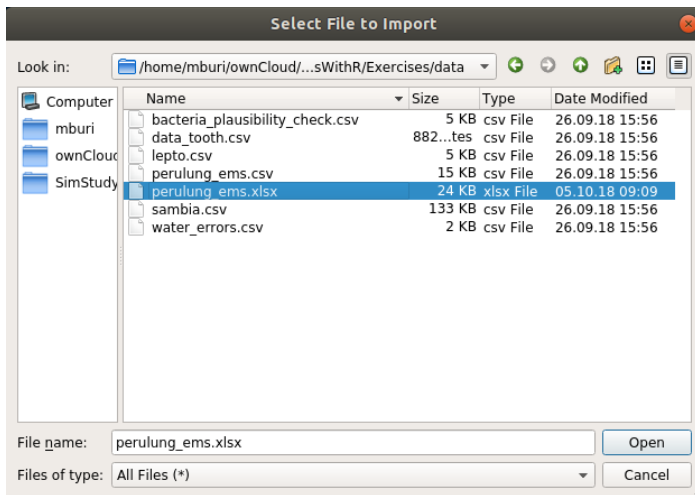
- Environment (upper right corner)
- > Import Dataset > From Excel... > Excel Files (.xlsx)

# How to import .xlsx files into R? (1/3)

- Environment (upper right corner)
- > Import Dataset > From Excel... > Excel Files (.xlsx)



## How to import .xlsx files into R? (2/3)



# How to import .xlsx files into R? (3/3)

Import Excel Data

File/Lib:

Data Preview:

id	fvc1	age	height	sex	respresymptoms
(double)	(double)	(double)	(double)	(double)	(double)
1	1.56	9.593	126.8	0	0
2	1.18	7.483	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	2.31	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	8.488	128.8	1	0
10	1.41	9.629	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	135.7	0	0
14	1.57	8.822	125.5	1	0
15	1.51	9.633	125.9	1	0
16	1.25	8.643	122.3	0	0
17	1.55	9.167	126.5	1	0
18	2.23	9.035	123.0	1	1
19	2.09	10.090	136.5	0	0
20	1.38	8.658	119.3	0	1
21	1.76	9.602	129.3	1	0
22	1.71	9.676	130.3	0	0

Previewing first 50 entries.

Import Options:

Name:  Max Rows:   
Sheet:  Skip:  ☒ First Row as Names  
Range:  NA:  ☒ Open Data Viewer

Code Preview:

```
library(readxl)
perulung_ems <- read_excel("Exercises\data\perulung_ems.xlsx")
View(perulung_ems)
```

[Reading Excel files using readxl](#)

```
perulung_ems <- read_excel("perulung_ems.xlsx")
lung <- data.frame(perulung_ems)
head(lung)
```

## Exercise 7: perulung

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

# Lecture Slides for Day 3

## Why do we need Statistics?

### **Repeatability of results:**

**Statistical science** allows us to estimate what might happen if an experiment was repeated - but without having to actually repeat it!



## Why do we need Statistics?

- Study results must be shown to be robust, i.e. real and not due to random chance
- Best way to demonstrate this is to repeat the same experiment/study many times each with **different** subjects (animals) drawn from the **same study population** and show that the result is truly repeatable
- It is generally totally impractical, in terms of both time and resources, to repeat an experiment many times!

## Why do we need Statistics?

- Instead of repeating the experiment many times **probability theory i.e. statistics** is used to **estimate** what might have happened if the experiment had been repeated
- A mathematical model is used to fill this “data gap”
- Generally the most difficult task in statistics is to decide what “model” is most appropriate for a given experiment

# What is Statistics? - A definition

A set of analytical tools designed to quantify uncertainty

- If an experiment or procedure is repeated, how likely is it that the new results will be similar to those already observed?
- What is the likely variation in results if the experiment was repeated?

# What is Statistics? - A definition

The key scientific purpose of statistics

- to provide **evidence** of the existence of some “effect” of scientific interest
- i.e. evidence based medicine

## As a reminder: The importance of study design

Even the most sophisticated statistical analyses cannot rescue a poorly designed study

→ unreliable results

→ inability to answer the main research question

## Putting Statistics in Context

- The vast majority of analyses can be done in a straightforward fashion - just remember and always use common sense as a guide - be skeptical!
- It is very easy to get “lost” in the statistical software and technical jargon, which differs markedly between different software packages. Terminology can also differ greatly between textbooks...
- Wikipedia is as good a resource as any for finding out about different statistical tests and terminology

# Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses

## Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)



# Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)
- Helps to decide what kind of formal statistical analyses might be most appropriate for the data available

# Exploratory Data Analysis

- It is crucially important to explore your data fully before considering any “formal” statistical analyses
- What explorations are done depends on the objective of the study - the research question(s)
- Helps to decide what kind of formal statistical analyses might be most appropriate for the data available
- What a simple descriptive analysis **does not** provide is evidence of whether the observed treatment effects are large enough to be notable once sampling variation has been accounted - that is the role of formal analyses, e.g. hypothesis testing

# Summary Statistics

## Continuous (Integers / Numeric)

- Mean - a measure of location. Always examine the average value of the response variable(s) for the different “treatment” effects in your data
- Median - a robust single value summary of a set of data (50% quantile point) - most useful in highly skewed data or data with outliers
- Standard deviation (sd) - a measure of spread, how variable the data are
- Standard error of the mean (se) - an estimate of how far the sample mean is likely to be from the population mean
- and others: min, max, range, IQR, ...



```
mean(x) # mean
```

```
median(x) # median
```

```
sd(x) # standard deviation
```

```
min(x) # minimum
```

```
max(x) # maximum
```

```
range(x) # range
```

```
IQR(x) # interquartile range
```

## Continuous Data Summaries

**standard deviation**

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

**standard error**

$$se = \frac{s}{\sqrt{n}}$$

## Correlation coefficient

### Combination of continuous and continuous

Correlation coefficient a measure association between two continuous variables (common but somewhat limited)

#### Pearson's correlation coefficient r

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$\bar{X}$ : mean of variable x

$\bar{Y}$ : mean of variable y



```
# Test for Association/Correlation Between  
# Paired Samples  
cor.test(data$x, data$y, method = "pearson")  
cor.test(data$x, data$y, method = "spearman")  
  
# Scatterplot(s)  
pairs(data$x ~ data$y)  
pairs(data)
```

# Summary Statistics

## Continuous and ordinal variables



```
tapply(data$x.cont, data$y.fac, mean)
```

```
tapply(data$x.cont, data$y.fac, median)
```

```
tapply(data$x.cont, data$y.fac, sd)
```





- Median - a robust single value summary of a set of data (50% quantile point) - most useful in highly skewed data or data with outliers
- e.g. 10th and 90th percentile - a measure of spread, how variable the data are

```
quantile(x, probs = c(0.1, 0.9))
```

- proportions - e.g. percentage per grade

```
prop.table(table(data$x.fac))  
prop.table(table(data$x.fac, data$y.fac))
```



- proportions - percentage within the different categories
- contingency tables e. g. 2 x 2

```
table(data$x.fac)
table(data$x.fac, data$y.fac)
prop.table(table(data$x.fac))
prop.table(table(data$x.fac, data$y.fac))
```

## Exercise 8



## How to deal with missing values in R? (1/3)

- In R, missing values are represented by the symbol **NA** (not available).
- Impossible values (e. g., dividing by zero) are represented by the symbol **NaN** (not a number).
- Ask yourself why a **NA** and / or **NaN** occurs!

# How to deal with missing values in R? (2/3)

- Testing for Missing Values

```
vec1 <- c(1, 2, 3, NA)
is.na(vec1) # returns a vector (FALSE, FALSE, FALSE, TRUE)
# The TRUE indicates the position of the NA in vec1.
```

- Recoding Values to Missing

```
# recode specific values (e. g. 0.001) to missing for variable x
# select rows where x is 0.001 and recode value in column x with NA
dat$x[dat$x == 0.001] <- NA
```

## How to deal with missing values in R? (3/3)

- Excluding Missing Values from specific function calls

```
a <- c(1, 2, NA, 3)
mean(a) # returns NA
mean(a, na.rm=TRUE) # returns 2
```

- Check for complete cases with function `complete.cases(...)`

```
# list rows of data that have missing values
dat[!complete.cases(dat),]
subdat <- dat[complete.cases(dat),]
```

- Create new dataset without missing data with function `na.omit(...)`

```
new.dat <- na.omit(dat)
```

## How to check your data for plausibility?

- Ask yourself what can go wrong?
- Implausible values?
- Impossible values?
- Logical errors?

## Exercise 9A: Plausibility Checks

