



# Introduction to R and data exploration

Martijn Wieling University of Groningen

### This lecture

- RStudio and R
- R as calculator
- Variables
- Functions and help
- Importing data in R in a dataframe
- Accessing rows and columns
- Adding columns to the data
- Goal of statistics
- Data exploration (descriptive statistics)
  - Numerical measures
  - Visual exploration

#### **Our tool: RStudio (frontend to R)**

B RStudio				
<u>File E</u> dit <u>C</u> ode <u>V</u> iew <u>P</u> lots <u>S</u> ession <u>B</u> uild <u>D</u> ebug <u>T</u> ools <u>H</u> elp				
🐑 🖣 🚽 🔚 🔚 🔚 🌈 Go to file/function 🔡 🗏 🖼 🗸 🛛 Addins 🗸			<b>(8</b> ) F	²roject: (None) 🔻
0 code1.R *		Environment History		
🔶 🖒 🙇 🕞 🖸 Source on Save 🛛 🔍 🎢 📲	🕈 Run 🛛 😎 📑 Source 👻 🚍	🕣 🔒 🖃 Import Dataset 🕶 🔬		🗉 List 🕶 🛛 🞯
1 # store a value in a variable		🛑 Global Environment 🗸	Q,	
2 a = 5		Data		
4 # store a series of values in a vector		🕐 dat 38	3 obs. of 3 variables	
5 b = c(1,2,3,4,a)		Values		
7 # apply a function (mean) to the data in the vector		a 5		
8 mean(b)		b nu	um [1:5] 1 2 3 4 5	
10 # look at help page of the function				
11 ?mean				
<pre>12 13 # import data which was exported to csv by Excel (which uses separator ";") 14 dat = read.csv2('thnl.csv') 15</pre>	)			
16 # look at structure of the data			=	
17 str(dat)		Files Plots Packages Help Vi	ewer	
		(= -> 🏠 🚔 🔊	Q,	G
20		R: Arithmetic Mean + Find in Topic		
19:1 [Top Level] ≑	R Script ≑	mean (hase)	R Docu	mentation
		incan (cacc)		nontation
Console C/Users/Martijn/Desktop/ R		Arithmetic Mean		
> a = 5	<u> </u>			
		Description		=
> h = c(1,2,3,4,a)				
		Generic function for the (trimmed) a	rithmetic mean.	
> # apply a function (mean) to the data in the vector > mean(b)		Usade		
[1] 3				
> # look at help page of the function		mean(x,)		
> ?mean		## Default S3 method:		
> # import data which was exported to csv by Excel (which uses separator ":")		<pre>mean(x, trim = 0, na.rm =</pre>	FALSE,)	
<pre>&gt; dat = read.csv2('thnl.csv')</pre>		Argumente		
> # look at structure of the data		Arguments		
> str(dat)		x An R object. Currently there	are methods for numeric/logical vectors and <u>date, date-time</u> and <u>time int</u>	erval
data.trame : 38 obs. of 3 variables: \$ Participant: Factor w/ 19 levels "VENI-NL 1"."VENI-NL 10": 1 2 3 4 5 6 7 8	9 10	objects. Complex vectors a	re allowed for trim = O, only.	
\$ Sound : Factor w/ 2 levels "T", "TH": 1 1 1 1 1 1 1 1 1 1		trim the fraction (O to 0.5) of obs	ervations to be trimmed from each end of ${f x}$ before the mean is computed	. Values
Frontness : num 0.781 0.766 0.884 0.748 0.748	E	of trim outside that range a	re taken as the nearest endpoint.	
	v	no rm a logical value indicating wh	other WA values should be stripped before the computation proceeds	-

#### **RStudio: quick overview**



#### **Basic functionality: R as calculator**

# Addition (this is a comment: preceded by '#')
5 + 5

#### # [1] 10

# Multiplication

5 \* 3

#### # [1] 15

# Division

5/3

# [1] 1.6667

### **Basic functionality: using variables**

- a <- 5 # store a single value; instead of '<-' you can also use '='</pre>
- a # display the value

#### # [1] 5

 $b \leq c(2, 4, 6, 7, 8) \#$  store a series of values in a vector b

#### # [1] 2 4 6 7 8

b[4] <-a # assign value 5 (stored in 'a') to the 4th element of vector b b[1] <-NA # assign NA (missing) to the first element of vector b b <-b \* 10 # multiply all values in vector b with 10 b

# [1] NA 40 60 50 80



# What is the value of b after the commands: b <- c(1,2,3); b <- c(b,b\*2)



Mentimeter

### **Basic functionality: using functions**

mn <- mean(b) # calculating the mean and storing in variable mn
mn</pre>

#### # [1] NA

# mn is NA (missing) as one of the values is missing
mean(b, na.rm = TRUE) # we can use the function parameter na.rm to ignore NAs

#### # [1] 57.5

# But which parameters does a function have: use help! help(mean) # alternatively: ?mean

### **Basic functionality: a help file**

mean {base}

R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

mean(x, ...)

```
## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

#### Х

An R object. Currently there are methods for numeric/logical vectors and <u>date</u>, <u>date-time</u> and <u>time interval</u> objects. Complex vectors are allowed for trim = 0, only.

#### trim

the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.

#### na.rm

a logical value indicating whether NA values should be stripped before the computation proceeds.

• • •

further arguments passed to or from other methods.



#### What is the purpose of the R function 'paste'?

🞽 Mentimeter





Press ENTER to show correct

### Try it yourself!

- There are many resources for R which you can easily find online
- Here we use "swirl" an online platform for interactive R courses
- Start RStudio, install and start swirl:

```
install.packages("swirl", repos = "http://cran.rstudio.com/")
library(swirl)
swirl()
```

- Follow the prompts and install the course *R* programming: The basics of programming in *R*
- Choose that course to start with and finish Lesson 1 of that course

### **Getting data into R: exporting a data set**

	<b>≤) -</b> (°= -   <del>-</del>		- Data Daviana Mana		thnl - Mic	rosoft Excel					
Normal	Page Page Break Custom Ful Layout Preview Views Screw	II Gridlines	✓ Formula Bar ✓ Headings Show Zoo	6 Zoom to Selection Mew Arrange F Window All P	reeze anes + Unhide	) View Side by Side ( Synchronous Scrolling ) Reset Window Position Vindow	Save Switch Workspace Windows	Macros Macros			
	L16 - (	f <sub>x</sub>	2	-		-	-				
	A	В		D	E	F	G	Н		J	K
1	Participant	Sex	Frontness.T	Frontness.TH							
2	VENI-NL_1	М	0,78051833	0,738011653							
3	VENI-NL_10	М	0,76620705	0,766849495							
4	VENI-NL_11	М	0,88366063	0,878713529							
5	VENI-NL_12	М	0,74756871	0,760936038	Vo	u can	ontor	vour	data i	n Evo	
6	VENI-NL_13	F	0,74761212	0,774196171	10	u can	CIILCI	your	uala I		<b>₽</b>  ,
7	VENI-NL_14	М	0,7518633	0,749126658	an	d save	e it as	a 'cs'	v' file		
8	VENI-NL_15	F	0,73293967	0,836400371		mma	sena	rated.	value	file)	
9	VENI-NL_16	М	0,69605361	0,664942896		//////	Jepu			· · · ·	-
10	VENI-NL_17	М	0,79925562	0,81018554	YO	u can	then	load t	nis da	ta into	) R.
11	VENI-NL_18	F	0,81542047	0,876608068							
12	VENI-NL_19	F	0,7144495	0,804125147							
13	VENI-NL_2	М	0,73680675	0,74903671							
14	VENI-NL_20	F	0,79514215	0,854225483							
15	VENI-NL_21	F	0,80580602	0,791117852							
16		Е	074571014	0 747007402							▼ 
Ready										🔲 🗆 💾 200'	%+

### **Getting data into R: importing a data set**

setwd("C:/Users/Martijn/Desktop/Statistics/Intro-R") # set working directory
dat <- read.csv2("thnl.csv") # read.csv2 reads Excel csv file from work dir
str(dat) # shows structure of the data frame dat (note: wide format)</pre>

# 'data.frame': 19 obs. of 4 variables: # \$ Participant : chr "VENI-NL\_1" "VENI-NL\_10" "VENI-NL\_11" "VENI-NL\_12" ... # \$ Sex : chr "M" "M" "M" "M" ... # \$ Frontness.T : num 0.781 0.766 0.884 0.748 0.748 ... # \$ Frontness.TH: num 0.738 0.767 0.879 0.761 0.774 ...

dim(dat) # number of rows and columns of data set

# [1] 19 4

### Investigating imported data set: using head

head(dat) # show first few rows of dat

#		Participant	Sex	Frontness.T	Frontness.TH
#	1	veni-nl_1	М	0.78052	0.73801
#	2	VENI-NL_10	М	0.76621	0.76685
#	3	VENI-NL_11	М	0.88366	0.87871
#	4	VENI-NL_12	М	0.74757	0.76094
#	5	VENI-NL_13	М	0.74761	0.77420
#	6	VENI-NL_14	М	0.75186	0.74913



## How can you show the first two lines of a dataframe (dat)? Mentimeter (multiple answers possible)



-

### Investigating imported data set: using RStudio viewer

RStudi	o Code View	Plots Se	ession Build Deb	ua Profile Tools Help								-		
- Q			Go to file/fun	iction								3	Project	t: (Non
dat ×	c					- 7	Environment	History	Connections	Tutorial				_
(ac)	an   ♥ Filter				Q		a 🛛 🕬	Import Data	set • 🔿 127	MiB 🔹 🍯			List	- 0
	Participant	Sex	° Frontness.T °	Frontness.TH			R - 🛑 Gloi	bal Environm	ent •			Q,		
1	VENI-NL_1	м	0.7805183	0.7380117			Data							-
2	VENI-NL_10	м	0.7662070	0.7668495			🗢 dat		19 obs.	of 4 vari	ables			
3	VENI-NL 11	м	0.8836606	0.8787135			\$ Part	ticipant	: chr "	VENI-NL_1"	"VENI-NL_1	.0" "VEN	I-NL_	11
4	VENI-NL_12	м	0.7475687	0.7609360			\$ Sex	these T	: chr "	781 0 766	"M"	8 0 749		
5	VENI-NL_13	м	0.7476121	0.7741962			\$ From	itness.T	H: num 0	.738 0.760	0.879 0.76	1 0.774		
6	VENI-NL_14	м	0.7518633	0.7491267										
7	VENI-NL_15	F	0.7329397	0.8364004										
8	VENI-NL_16	м	0.6960536	0.6649429										
9	VENI-NL_17	м	0.7992556	0.8101855										
10	VENI-NL_18	F	0.8154205	0.8766081										
11	VENI-NL_19	F	0.7144495	0.8041251										
12	VENI-NL_2	м	0.7368067	0.7490367										
13	VENI-NL_20	F	0.7951421	0.8542255										
14	VENI-NL_21	F	0.8058060	0.7911179										
15	VENI-NL_3	F	0.7457121	0.7470974										
16	VENI-NL_4	F	0.7250988	0.7432163										
17	VENI-NL_5	F	0.7943390	0.8175356										
18	VENI-NL_6	м	0.8219972	0.8052455										
19	VENI-NL_9	F	0.7551367	0.7570195										
showing 1	1 to 19 of 19 ent	tries, 4 total	columns											
Console						0 D	Files Plots	Packages	Help Vi	ewer Present	tation			Ð

#### Subsetting the data: indices and names

dat[1, ] # values in first row

# Participant Sex Frontness.T Frontness.TH

# 1 VENI-NL 1 M 0.78052 0.73801

dat[1:2, c(2, 3)] # values of first two rows for second and third column

# Sex Frontness.T
# 1 M 0.78052
# 2 M 0.76621

dat[c(1, 2, 3), "Participant"] # values of first three rows for column 'Participant'

# [1] "VENI-NL\_1" "VENI-NL\_10" "VENI-NL\_11"

tmp <- dat[5:8, c(1, 3)] # store columns 1 and 3 for rows 5 to 8 in tmp</pre>



# How can you view the value in the 3rd colomn and 4th row of dataframe 'dat'?



19/43

Mentimeter

÷ →

### Subsetting the data: conditional indexing

tmp <- dat[dat\$Sex == "M", ] # only observations for male participants
head(tmp, n = 2) # show first two rows</pre>

#		Participant	Sex	Frontness.T	Frontness.TH
#	1	VENI-NL_1	М	0.78052	0.73801
#	2	veni-nl_10	М	0.76621	0.76685

```
# more advanced subsetting: include rows for which frontness for the T sound is
# higher than 0.74 AND participant is either 1 or 2 N.B. use '|' instead of '&' for
# logical OR
dat[dat$Frontness.T > 0.74 & dat$Participant %in% c("VENI-NL_1", "VENI-NL_2"), ]
```

# Participant Sex Frontness.T Frontness.TH

# 1 VENI-NL\_1 M 0.78052 0.73801



#### What is the result of: dat[dat\$Sex=='M'|dat\$Sex=='F']

🞽 Mentimeter



 $\leftarrow \rightarrow$ 

-

### Supplementing the data: adding columns

```
# new column Diff containing difference between TH and T positions
dat$Diff <- dat$Frontness.TH - dat$Frontness.T</pre>
```

```
# new column DiffClass, initially all observations set to THO
dat$DiffClass <- "THO"</pre>
```

```
# observations with Diff larger than 0.02 are categorized as TH1, negative as TH-
dat[dat$Diff > 0.02, ]$DiffClass <- "TH1"
dat[dat$Diff < 0, ]$DiffClass <- "TH-"</pre>
```

```
dat$DiffClass <- factor(dat$DiffClass) # convert string variable to factor</pre>
```

head(dat, 2)

#	Participant	Sex	Frontness.T	Frontness.TH	Diff	DiffClass
# 1	VENI-NL_1	М	0.78052	0.73801	-0.04250668	TH-
# 2	VENI-NL_10	М	0.76621	0.76685	0.00064245	THO



Mentimeter

# What is the effect of: dat\$Test = paste(dat\$DiffClass,dat\$Sex)



 $\leftarrow \rightarrow$ 

### Try it yourself!

- Run **swirl()** and finish the following lessons of the *R Programming* course:
  - Lesson 6: Subsetting vectors
  - Lesson 12: Looking at data

#### **Statistics**

- **Goal of statistics** is to gain understanding from data
  - Descriptive statistics (this lecture): describe data without further conclusions
  - Inferential statistics: describe data (**sample**) and its relation to larger group (**population**)

#### Numerical variables: central tendency and spread

mean(dat\$Diff) # mean

#### # [1] 0.016263

median(dat\$Diff) # median

#### # [1] 0.01093

min(dat\$Diff) # minimum value

#### # [1] -0.042507

max(dat\$Diff) # maximum value

#### # [1] 0.10346

#### Numerical variables: measures of spread

sd(dat\$Diff) # or: sqrt((1/(length(dat\$Diff)-1)) \* sum((dat\$Diff - mean(dat\$Diff))^2))

#### # [1] 0.038213

var(dat\$Diff) # or: sd(dat\$Diff)^2

#### # [1] 0.0014603

quantile(dat\$Diff) # quantiles

#	0%	25%	50%	75%	100%
# -0.0	)425067 -0	0.0038419	0.0109299	0.0248903	0.1034607

```
summary(dat$Diff) # summary
```

#	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
#	-0.04251	-0.00384	0.01093	0.01626	0.02489	0.10346

### **Categorical variables: frequency tables**

table(dat\$Sex)
# # F M # 9 10
<pre>with(dat, table(Sex)) # alternative</pre>
# Sex # F M # 9 10
table(dat\$DiffClass)
# # TH- THO TH1 # 6 7 6



# What is the effect of table(dat\$Sex,dat\$DiffClass)?

Mentimeter



29/43

-

← →

### **Exploring relationships between pairs of variables**

# correlation: relation between two numerical variables

```
cor(dat$Frontness.T, dat$Frontness.TH)
```

#### # [1] 0.71054

# crosstable: relation between two categorical variables
table(dat\$Sex, dat\$DiffClass) # or: with(dat, table(Sex,DiffClass))

```
# TH- THO TH1
# F 1 3 5
# M 5 4 1
```

# means per category: relation between numerical and categorical variable
c(mean(dat[dat\$Sex == "M", ]\$Diff), mean(dat[dat\$Sex == "F", ]\$Diff))

#### # [1] -0.0034299 0.0381446



# Suppose the result of cor(x,y) = NA, what do you know?



Mentimeter

 $\leftarrow \rightarrow$ 

31/43

#### **Data exploration with visualization**

- Many basic visualization options are available in R
  - **boxplot()** for a boxplot
  - **hist()** for a histogram
  - qqnorm() and qqline() for a quantile-quantile plot
  - **plot()** for many types of plots (scatter, line, etc.)
  - **barplot()** for a barplot (plotting frequencies)

### **Exploring numerical variables: box plot**

par(mfrow = c(1, 2)) # set graphics option: 2 graphs side-by-side boxplot(dat\$Diff, main = "Difference") # boxplot of difference values boxplot(dat[, c("Frontness.T", "Frontness.TH")]) # frontness per group



Difference

### **Exploring numerical variables: histogram**

hist(dat\$Diff, main = "Difference histogram")



#### Difference histogram

dat\$Diff

### Exploring numerical variables: Q-Q plot

qqnorm(dat\$Diff) # plot actual values vs. theoretical quantiles
qqline(dat\$Diff) # plot reference line of normal distribution





#### **Exploring numerical relations: scatter plot**

plot(dat\$Frontness.T, dat\$Frontness.TH, col = "blue")



### Visualizing categorical variables (frequencies): bar plot

counts <- table(dat\$Sex) # frequency table for sex</pre>

barplot(counts, ylim = c(0, 15))



### **Exploring categorical relations: segmented bar plot**

counts <- table(dat\$Sex, dat\$DiffClass)
barplot(counts, col = c("pink", "lightblue"), legend = rownames(counts), ylim = c(0, 10))</pre>





# Which plot parameters can you use to change axis labels?



39/43

Mentimeter

### Try it yourself!

- Run swirl() and finish the following lesson of the *R Programming* course:
  - *Lesson 15*: Base graphics

#### Recap

- $\cdot\,$  In this lecture, we've covered the basics of  ${\rm I\!R}$
- Now you should be able (with help of this presentation) to use **R** for:
  - Data manipulation, exploration and visualization
- Associated lab session and additional swirl resources:
  - https://www.let.rug.nl/wieling/Statistics/Intro-R/lab
  - Install swirl course *Exploratory Data Analysis* 
    - install\_from\_swirl('Exploratory\_Data\_Analysis')
    - Finish Lessons 1 5 (download associated slides)
    - If interested, you can finish the full *Exploratory Data Analysis* course



#### Please provide your opinion about this lecture in Mentimeter at most 3 words/phrases!



### **Questions?**

Thank you for your attention!

http://www.martijnwieling.nl m.b.wieling@rug.nl

