

Output development using R & R markdown

Ziv Shkedy and Thi Huyen Nguyen
Hasselt University

Foundations for inference
Ha Noi, 29/05/2024



ER-BioStat

GitHub  <https://github.com/eR-Biostat>

twitter  @erbiostat



The airquality data

Introduction

What do we do in this session ?

- We conduct a simple analysis for the variable Wind speed in the `airquality` data:
 - Summary statistics.
 - Graphical display: histogram.
 - Confidence interval.
 - Test of hypothesis.
- Focus:
 - How to produce an output ?
 - How to combine text and software output in the same document ?



The airquality data

R Program: Case studies_Prog1_V1.R

Part 1


The dataset

The `airquality` data in R

```
> dim(airquality)
```

```
[1] 153 6
```

The R object for the data: 153 observations and 6 variables.



```
> names(airquality)
```

```
[1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"
```

Variables names:



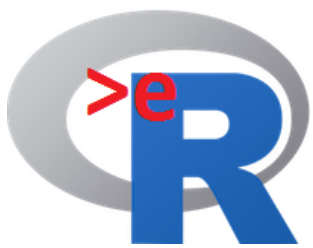
Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.

Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport.

The `airquality` data in R

```
> head(airquality)
```

	zone	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



The airquality data

Part 2

The Rmd program

The output

- We run the R markdown file.
- Produce output format: Word document.

The Rmd program

The screenshot displays the RStudio interface with an R Markdown document open. The document is titled 'Case studies_Prog1_V1.Rmd*'. The code is divided into two main sections by red curly braces:

- Document setup:** This section includes the document title, output format (pdf_document), subtitle, and layout. It also contains a chunk for setting up the R environment with various options and loading numerous packages.
- Many R packages, not all needed:** This section lists the packages loaded in the R environment, including knitr, tidyverse, desolve, minpack.lm, ggpubr, readxl, gamlss, data.table, grid, png, nlme, gridExtra, mvtnorm, e1071, lattice, and ggplot2.

The right-hand pane shows the Environment tab, which lists the objects in the Global Environment. The Data tab shows a list of objects, including hist_bmi, NHANES, NHANES_summary, params, sleep_summary, TotChol_summary, and violin_bmi. The Values tab shows the values of these objects.

The bottom status bar shows the RStudio logo and the text 'R Markdown'.

Document setup.

```
1 ---
2 title: 'Case study 1: analysis of the daily average wind speed in New York in 1973'
3 output:
4   word_document: default
5   pdf_document: default
6   html_document: default
7 subtitle: Foundations for inference using R
8 layout: page
9 ---
10
11
12 {r setup, include=FALSE}
13 options(htmltools.dir.version = FALSE)
14 knitr::opts_chunk$set(echo = TRUE,
15   message = FALSE,
16   warning = FALSE,
17   eval = TRUE,
18   tidy = FALSE)
19 library(knitr)
20 library(tidyverse)
21 library(desolve)
22 library(minpack.lm)
23 library(ggpubr)
24 library(readxl)
25 library(gamlss)
26 library(data.table)
27 library(grid)
28 library(png)
29 library(nlme)
30 library(gridExtra)
31 library(mvtnorm)
32 library(e1071)
33 library(lattice)
34 library(ggplot2)
```

Many R packages, not all needed.

Environment **History** **Connections** **Tutorial**

Import Dataset 162 MiB

R Global Environment

Data

Object	Value
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

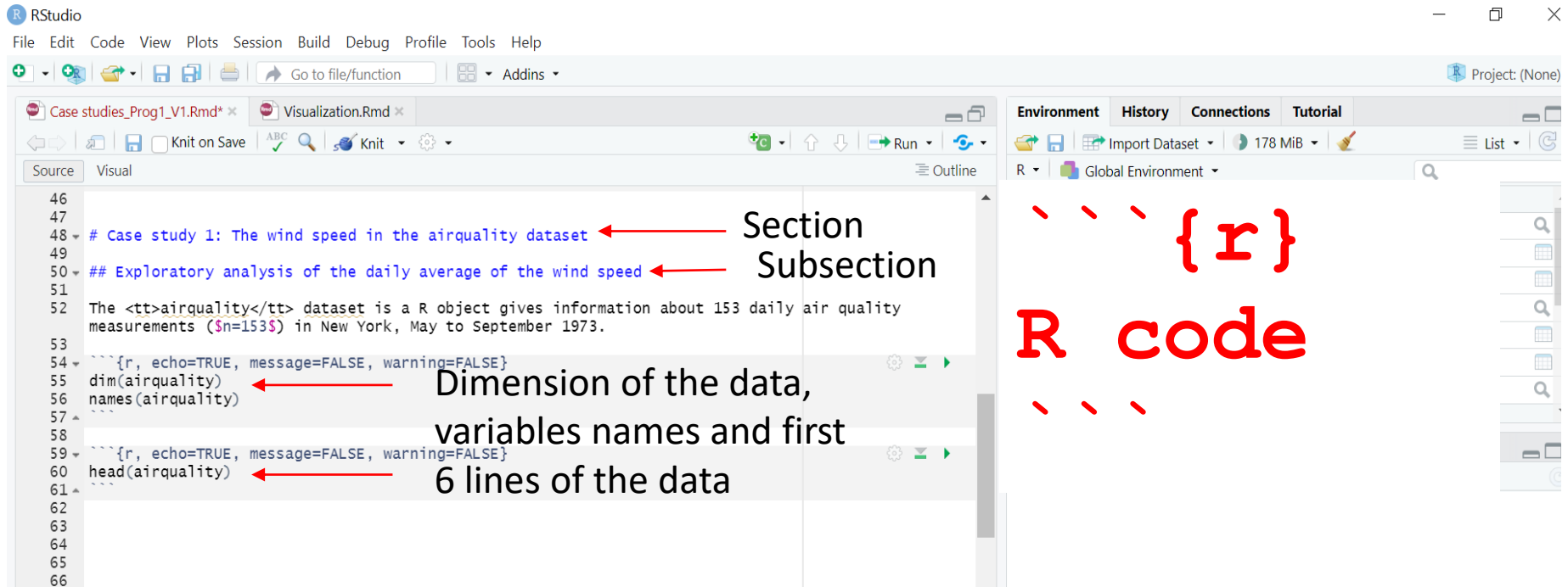
Values

Files **Plots** **Packages** **Help** **Viewer** **Presentation**

Zoom Export

10:55 21/05/2024

The Rmd program



The screenshot shows the RStudio interface with an R Markdown file open. The source editor displays the following code:

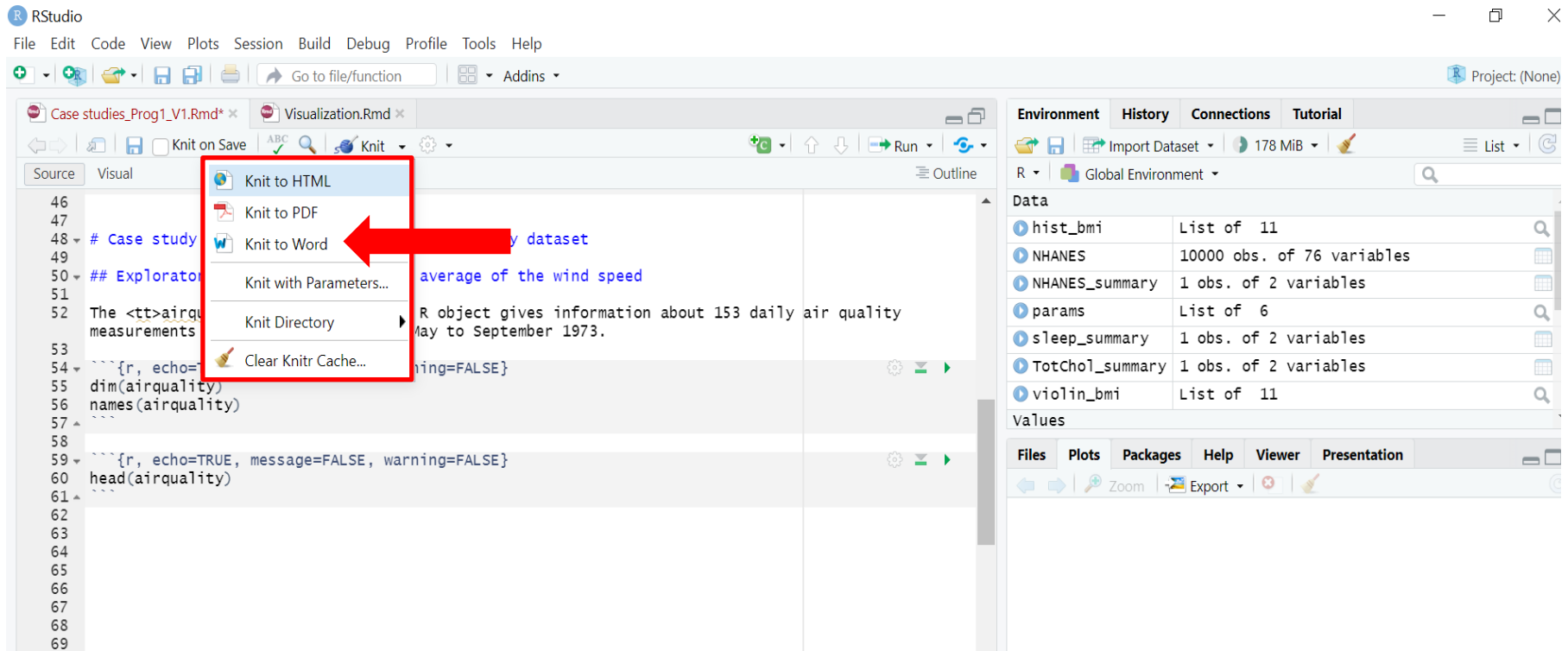
```
46  
47  
48 # Case study 1: The wind speed in the airquality dataset  
49 ## Exploratory analysis of the daily average of the wind speed  
50  
51  
52 The <tt>airquality</tt> dataset is a R object gives information about 153 daily air quality  
   measurements ( $n=153$ ) in New York, May to September 1973.  
53  
54 {r, echo=TRUE, message=FALSE, warning=FALSE}  
55 dim(airquality)  
56 names(airquality)  
57  
58  
59 {r, echo=TRUE, message=FALSE, warning=FALSE}  
60 head(airquality)  
61  
62  
63  
64  
65  
66
```

Annotations with red arrows point to specific parts of the code:

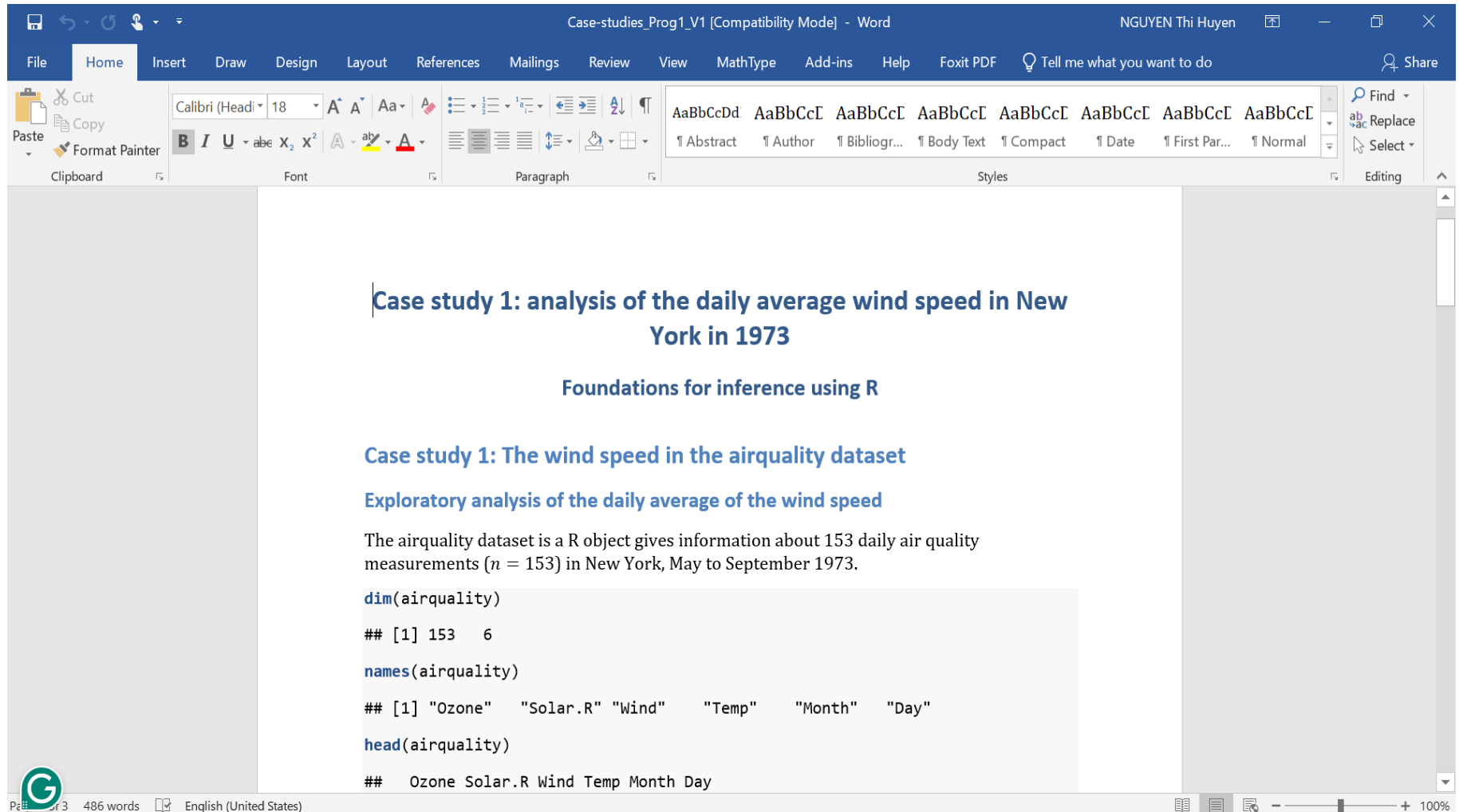
- Section** points to the line `# Case study 1: The wind speed in the airquality dataset`.
- Subsection** points to the line `## Exploratory analysis of the daily average of the wind speed`.
- Dimension of the data, variables names and first 6 lines of the data** points to the R code blocks `dim(airquality)`, `names(airquality)`, and `head(airquality)`.

On the right side of the image, there is a red dashed box containing the text `{r}` and the text **R code**.

Choose the output



The Word document output





The airquality data


Part 3

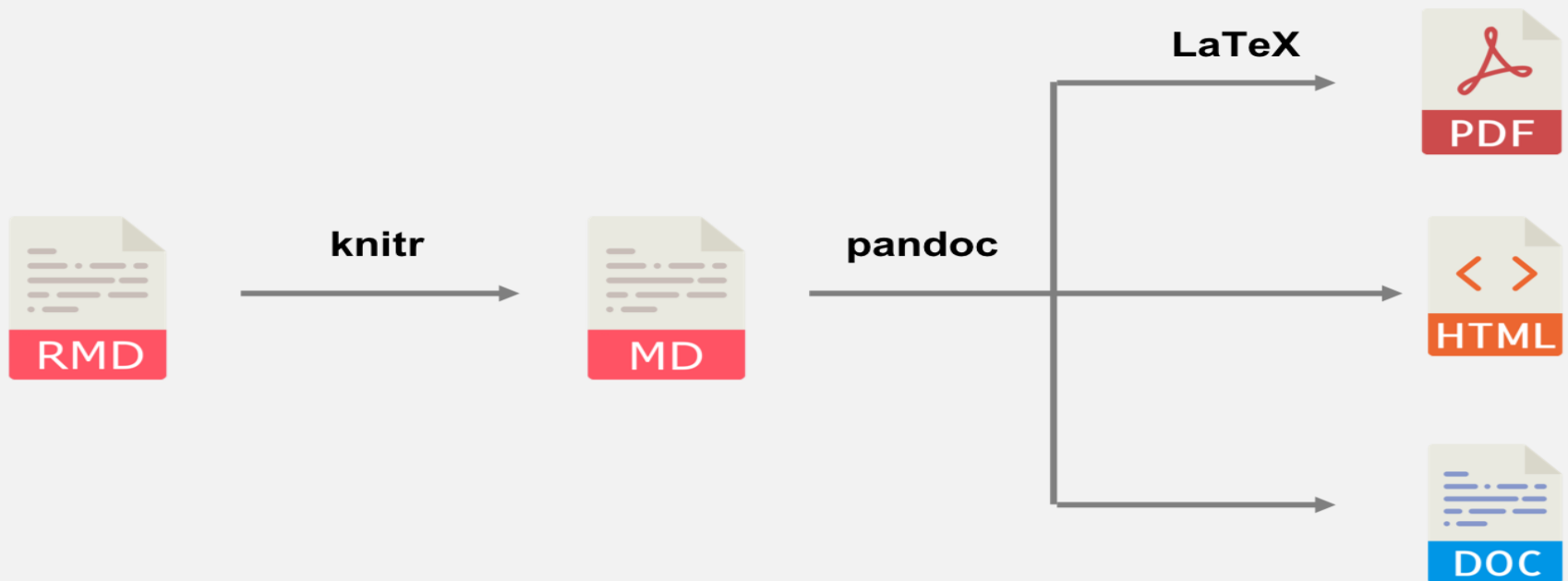
How to produce the Word output ?

Reproducible Research

- Aim: create an output in a Word document.
- Can be used to communicate the analysis' results with other people in the organization.
- Not all potential readers are interested on “how to do the analysis”.
- We DO NOT aim to develop a report for the analysis but to provide a document from which the results can be seen and discuss by different people in the organization.

The Rmd file

- Analyses  high quality report.
- Rmarkdown – Different dynamic and statistic formats (html, pdf, **word**, books, dashboard, e.t.c).



The Word output

- A Word document output.
- Presents the same analysis as in the example.

Case-studies_Prog1_V1 [Compatibility Mode] - Word

NGUYEN Thi Huyen

File Home Insert Draw Design Layout References Mailings Review View MathType Add-ins Help Foxit PDF Tell me what you want to do Share

Clipboard Font Paragraph Styles Editing

Case study 1: analysis of the daily average wind speed in New York in 1973

Foundations for inference using R

Case study 1: The wind speed in the airquality dataset

Exploratory analysis of the daily average of the wind speed

The airquality dataset is a R object gives information about 153 daily air quality measurements ($n = 153$) in New York, May to September 1973.

```
dim(airquality)
## [1] 153 6

names(airquality)
## [1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"

head(airquality)
## Ozone Solar.R Wind Temp Month Day
```

Title

Analysis output

Page 3 486 words English (United States)



Part 3.1:

How to set up the Word file?

The Rmd file

- We use Rmd file to
 - Conduct the analysis.
 - Set up the document.
- We use a Word file in order to
 - Present & communicate the result.

Set up the document

The screenshot displays the RStudio interface with an R Markdown document titled 'Case studies_Prog1_V1.Rmd'. The document is in 'Source' view, showing the following code:

```
1 ---
2 title: 'Case study 1: analysis of the daily average wind speed in New York in 1973'
3 output:
4   word_document: default
5   pdf_document: default
6   html_document: default
7 subtitle: Foundations for inference using R
8 layout: page
9 ---
10
11
12 {r setup, include=FALSE}
13 options(htmltools.dir.version = FALSE)
14 knitr::opts_chunk$set(echo = TRUE,
15   message = FALSE,
16   warning = FALSE,
17   eval = TRUE,
18   tidy = FALSE)
19
20 library(knitr)
21 library(tidyverse)
22 library(deSolve)
23 library(minpack.lm)
24 library(ggpubr)
25 library(readxl)
26 library(gamlss)
27 library(data.table)
28 library(grid)
29 library(png)
30 library(nlme)
31 library(gridExtra)
32 library(mvtnorm)
33 library(e1071)
34 library(lattice)
35 library(ggplot2)
```

A red arrow points to the `word_document: default` line in the `output:` section. Another red arrow points to the text 'Set up the Word document: Word_document: default' which is overlaid on the code editor. The 'Environment' pane on the right shows the 'Global Environment' with a list of data objects:

Data	
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

The 'Values' pane shows the following values:

Files	Plots	Packages	Help	Viewer	Presentation
Case studies_Prog1_V1_KnitrCompatibility Model	Word	NGUYEN Thi Nguyen			

The 'Viewer' pane shows a preview of the document output, which is a Word document. The preview displays the title 'Case study 1: analysis of the daily average wind speed in New York in 1973' and the subtitle 'Foundations for inference using R'. The main content of the document is a paragraph about the air quality dataset, followed by a code chunk showing the output of the `dim` function applied to the `airquality` dataset.

```
## [1] 153 6
names(airquality)
## [1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"
head(airquality)
##   Ozone Solar.R Wind Temp Month Day
```

Titles, authors and dates

The screenshot shows the RStudio interface with a project named 'Case studies_Prog1_V1.Rmd'. The source editor displays R Markdown code. A red box highlights the following code block:

```
---
title: 'Case study 1: analysis of the daily average wind speed in New York in 1973'
output:
  word_document: default
  pdf_document: default
  html_document: default
subtitle: Foundations for inference using R
author: "Ziv Shkedy et al."
date: "29-05-2024"
layout: page
---
```

To the right of this code block, the text 'Subtitle, author, and date' is displayed. Below the code block, the following R code is visible:

```
{r setup, include=FALSE}
options(htmltools.dir.version = FALSE)
knitr::opts_chunk$set(echo = TRUE,
  message = FALSE,
  warning = FALSE,
  eval = TRUE,
```

The right-hand pane shows the 'Environment' tab with a list of data objects:

Object	Description
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

The bottom pane shows the 'Console' tab with the R version and file path:

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
```

Titles, authors and dates

Case-studies_Prog1_V1 [Compatibility Mode] - Word

NGUYEN Thi Huyen

File Home Insert Draw Design Layout References Mailings Review View MathType Add-ins Help Foxit PDF Tell me what you want to do Share

Clipboard Font Paragraph Styles Editing

Case study 1: analysis of the daily average wind speed in New York in 1973 Title

Foundations for inference using R Sub title

Ziv Shkedy et al.

29-05-2024

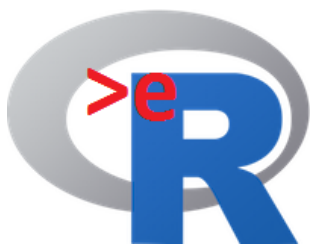
Case study 1: The wind speed in the airquality dataset

Exploratory analysis of the daily average of the wind speed

The airquality dataset is a R object gives information about 153 daily air quality measurements ($n = 153$) in New York, May to September 1973.

```
dim(airquality)
## [1] 153 6
names(airquality)
## [1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"
head(airquality)
## Ozone Solar.R Wind Temp Month Day
```

491 words English (United States)



Part 3.2:

The Word document and the Rmd program in details.

Section, subsection, subsubsection

The screenshot shows the RStudio interface with an R Markdown file open. The source editor displays R code and text. A red box highlights a paragraph of text, and red arrows point from labels to specific parts of the code and text.

Section: Points to the code line `# Case study 1: The wind speed in the airquality dataset`.

Subsection: Points to the code line `## Exploratory analysis of the daily average of the wind speed`.

Free text: Points to the paragraph of text: "The variable of primary interest, `wind`, is the average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport. We use the R package `ggplot2` to explore the

The Environment pane on the right shows a table of data:

Object	Value
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

- In addition to the code, we can add free text in the Rmd file.

Section, subsection, subsubsection

The screenshot shows a Microsoft Word document titled "Case-studies_Prog1_V1 [Compatibility Mode] - Word" by NGUYEN Thi Huyen. The document content is as follows:

Case study 1: analysis of the daily average wind speed in New York in 1973

Foundations for inference using R

Ziv Shkedy et al.

29-05-2024

Case study 1: The wind speed in the airquality dataset ← section

Exploratory analysis of the daily average of the wind speed ← subsection

The airquality dataset is a R object gives information about 153 daily air quality measurements ($n = 153$) in New York, May to September 1973.

```
dim(airquality)
## [1] 153 6
names(airquality)
## [1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"
head(airquality)
## Ozone Solar.R Wind Temp Month Day
```

Annotations in the image:

- A red box highlights the paragraph "The airquality dataset is a R object gives information about 153 daily air quality measurements ($n = 153$) in New York, May to September 1973." with the label "free text" to its left.
- Red arrows point from the text "section" to the bolded title "Case study 1: The wind speed in the airquality dataset".
- Red arrows point from the text "subsection" to the bolded title "Exploratory analysis of the daily average of the wind speed".
- Red arrows point from the text "The code is shown as a part of the output" to the R code block.

Word status bar: Page 3, 491 words, English (United States), 100% zoom.

Code in the Rmd file

The screenshot shows the RStudio interface with a file named 'Case studies_Prog1_V1.Rmd' open. The code editor displays R code for generating a histogram. A red rectangle highlights the following code block:

```
66 {r, echo=TRUE, message=FALSE, warning=FALSE, fig.cap="Histogram with density of wind speed"}
67 ggplot(airquality, aes(x = wind)) +
68   geom_histogram(aes(y = ..density..), fill = "skyblue", color = "black") +
69   geom_density(alpha = 0.2, fill = "orange") +
70   ylab("Density")
```

A red arrow points from the text 'The code for histogram' to the highlighted code block.

The right sidebar shows the 'Environment' pane with a table of data objects:

Object	Description
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotCho1_summary	1 obs. of 2 variables
violin_bmi	List of 11

The bottom status bar shows the R version (4.2.1) and the current file path: 'C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/'.

The output in the Word file

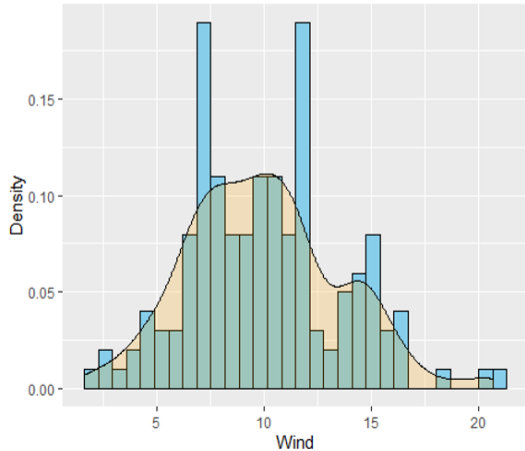
Case-studies_Prog1_V1 [Compatibility Mode] - Word

NGUYEN Thi Huyen

File Home Insert Draw Design Layout References Mailings Review View MathType Add-ins Help Foxit PDF Tell me what you want to do Share

Paste Cut Copy Format Painter Clipboard Font Paragraph Styles Editing

```
ggplot(airquality, aes(x = Wind)) +  
  geom_histogram(aes(y = ..density..), fill = "skyblue", color = "black") +  
  geom_density(alpha = 0.2, fill = "orange") +  
  ylab("Density")
```



Histogram with density of wind speed.

501 words English (United States)

19°C 12:50 21/05/2024

Code in the Rmd file

The screenshot displays the RStudio interface with an R Markdown file open. The source editor shows the following code:

```
83  
84 ## A 95% confidence interval for the population mean  
85  
86 We construct a 95% confidence interval for the population mean using the R function  
<tt>z.test</tt>. We assume that  $\sigma=s$  and the function uses the  $N(0,1)$  to select the  
critical value  $z_{\alpha}$ . For a 95% confidence interval,  $z_{\alpha}=1.96$ .  
87 A 95% confidence interval for the wind speed is [9.399284; 10.515749]. This means that we are  
95% confident that the true average wind speed lies within this range.  
88  
89 {r, echo=TRUE, message=FALSE, warning=FALSE}  
90 library(TeachingDemos)  
91 wind=na.omit(airquality$wind)  
92 z.test(wind,sd=SD.wind)  
93  
94  
95  
96  
97  
98
```

The code block from line 89 to 92 is highlighted with a red box. A text overlay points to this box:

The code for confidence interval.

The Environment pane on the right shows the following data objects:

Object	Details
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

The Console pane at the bottom shows the R prompt and the file path: R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/

The output in the Word file

The screenshot displays the Microsoft Word interface with the title bar 'Case-studies_Prog1_V1 [Compatibility Mode] - Word' and the user name 'NGUYEN Thi Huyen'. The ribbon includes tabs for File, Home, Insert, Draw, Design, Layout, References, Mailings, Review, View, MathType, Add-ins, Help, Foxit PDF, and a search bar. The 'Home' tab is active, showing font and paragraph settings. The document content is as follows:

A 95% confidence interval for the population mean

We construct a 95% confidence interval for the population mean using the R function `z.test`. We assume that $\sigma = s$ and the function uses the $N(0,1)$ to select the critical value Z_α . For a 95% confidence interval, $Z_\alpha = 1.96$. A 95% confidence interval for the wind speed is [9.399284; 10.515749]. This means that we are 95% confident that the true average wind speed lies within this range.

```
library(TeachingDemos)
wind=na.omit(airquality$wind)
z.test(wind,sd=SD.wind)
```

The output of the R code is shown in a separate window, with the 95% confidence interval highlighted by a red box:

```
##
## One Sample z-test
##
## data: wind
## z = 34.961, n = 153.00000, Std. Dev. = 3.52300, Std. Dev. of the sample
## mean = 0.28482, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  9.399284 10.515749
## sample estimates:
## mean of wind
##  9.957516
```

To the right of the output window, the text 'The output for confidence interval.' is displayed.

The bottom status bar shows '501 words', 'English (United States)', and a zoom level of '80%'. The taskbar at the bottom includes icons for Windows, Google Chrome, Presentation, Visualization, Downloads, RStudio, eR_biostat_i..., Meeting 21..., Case-studie..., and system icons for weather (19°C), date (21/05/2024), and time (12:56).

Code in the Rmd file

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Case studies_Prog1_V1.Rmd* Visualization.Rmd*

Knit on Save ABC Knit Run

Source Visual Outline

```
99 ## Test of hypothesis about the population mean
100 Testing the hypotheses whether the wind speed is equal to 9 versus a two-sided alternative
101 hypothesis at the significant level of 0.05 can be formulated by:
102  $H_0: \mu = 9$  Vs.  $H_1: \mu \neq 9$ .
```

$H_0: \mu = 9$ Vs. $H_1: \mu \neq 9$.

```
103
104 We use the z.test() function and specify mu=9.
105
106 z.test(wind, SD.wind, mu=9)
107
108 Since n-value = 0.0007742 which is much smaller than  $\alpha = 0.05$ , there is sufficient evidence.
109
110 # Test of hypothesis about the population mean
```

The code for testing hypotheses.

Environment History Connections Tutorial

Import Dataset 152 MiB

R Global Environment

Data

hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

Files Plots Packages Help Viewer Presentation

Zoom Export

Console Terminal Render Background Jobs

R 4.2.1 C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/

Windows Taskbar: Presentati... Visualizati... Downloads RStudio eR_biosta... Meeting ... Case-stu... Testing o... 19°C 12:57 21/05/2024

The output in the Word file

Case-studies_Prog1_V1 [Compatibility Mode] - Word

NGUYEN Thi Huyen

File Home Insert Draw Design Layout References Mailings Review View MathType Add-ins Help Foxit PDF Tell me what you want to do Share

Clipboard Font Paragraph Styles Editing

Test of hypothesis about the population mean

Testing the hypotheses whether the wind speed is equal to 9 versus a two-sided alternative hypothesis at the significant level of 0.05 can be formulated by:

$$H_0: \mu = 9 \text{ Vs. } H_1: \mu \neq 9.$$

We use the `z.test()` function and specify `mu=9`.

```
z.test(wind, SD.wind, mu=9)
```

```
##  
## One Sample z-test  
##  
## data: wind  
## z = 3.3619, n = 153.00000, Std. Dev. = 3.52300, Std. Dev. of the sample  
## mean = 0.28482, p-value = 0.0007742  
## alternative hypothesis: true mean is not equal to 9  
## 95 percent confidence interval:  
## 9.399284 10.515749  
## sample estimates:  
## mean of wind  
## 9.957516
```

Since p-value = 0.0007742 which is much smaller than $\alpha = 0.05$, there is sufficient evidence to say that the mean of the wind speed is not equal to 9.

The output for testing hypotheses.

Page 3 of 3 501 words English (United States)

19°C 12:56 21/05/2024

Discussion

- R Studio + R markdown:
- Easy to use.
- Text + code.
- Output:
 - Standard: HTML, PDF, DOC.
 - Example: Word.doc.