

Output development using R & R markdown

Ziv Shkedy and Thi Huyen Nguyen
Hasselt University

Foundations for inference
Ha Noi
03/03/25-07/03/25



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 conduct the analysis in R ?
 - 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:



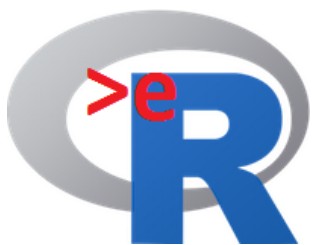
Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.

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

Analysis plan + analysis in R

Analysis plan

- Response: wind speed.
- Analysis:
 - Calculate mean and SD.
 - Construct 95% C.I. for the population mean.
 - Test the hypothesis for the population mean.

Case 1: using a $N(0,1)$

If $X \sim N(\mu, \sigma^2)$

then: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

And $Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

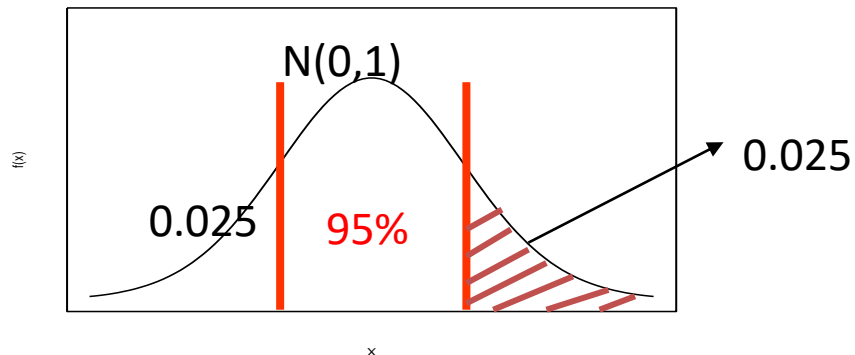
1. X has a normal distribution with unknown μ and known σ^2 .

CI for case 1

Step 1: example, choose $1-\alpha = 0.95$

Step 2: Case 1, thus:
$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Step 3: critical point:



CI for case 1

Step 4: calculate the point estimator : \bar{X}

Step 5: calculate the CI

For this, we know:

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96\right) = 0.95$$

or, after the conversion of the formula:

$$P\left(\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

CI for case 1

$$P\left(\underbrace{\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_L \leq \mu \leq \underbrace{\bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_R\right) = 0.95$$

L

R

$$P(L \leq \mu \leq R) = 1 - \alpha$$

So, a $(1-\alpha)$ CI for μ is :

$$\left[\bar{x} - z \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z \sqrt{\frac{\sigma^2}{n}} \right]$$

Two sided hypothesis testing

The mean under H_0 is not equal to the mean under H_1 :

$$H_0 : \mu = \mu_{H_0}$$

null hypothesis

$$H_1 : \mu \neq \mu_{H_0}$$

alternative
hypothesis

two sided test problem

Test statistic

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

R functions for the analysis

- `mean()`
- `var()`
- `z.test()`

R program for the analysis: `Airquality.R`


R program

```
wind<-airquality$Wind
```

```
M.wind<-mean(wind)
```

```
SD.wind<-sqrt(var(wind))
```

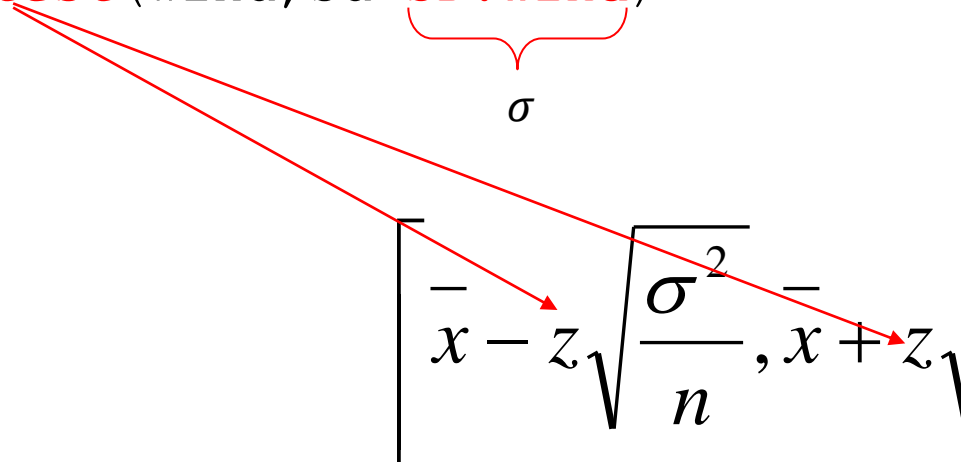
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$


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

R program

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

σ


$$\left[\bar{x} - z \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z \sqrt{\frac{\sigma^2}{n}} \right]$$

R program

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

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

$$H_0: \mu = 9$$

$$H_1: \mu \neq 9$$

Output

```
> wind<-airquality$Wind
> M.wind<-mean(wind)
> SD.wind<-sqrt(var(wind))
> M.wind
[1] 9.957516
> SD.wind
[1] 3.523001
```

Output (C.I)

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

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
```

$$\left[\bar{x} - z \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z \sqrt{\frac{\sigma^2}{n}} \right] = [9.399, 10.515]$$

Output (test of hypothesis)

```
> 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
```



$H_0: \mu = 9$

$H_1: \mu \neq 9$

A two sided test

Case 3: using a $t_{(n-1)}$ distribution

An estimate for the standard error of the sample mean.

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

→ The standard deviation.
→ The sample size.

If the sample comes from a population whose **distribution is normal** and the **variance is unknown**, and the **sample size is small**, then :

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$$

A Student t-distribution with (n-1) degrees of freedom is denoted by $t_{(n-1)}$.

R functions for the analysis

- In our case, $n=153$ ($n>30$!!!) but, for illustration, we will use t distribution for the test.
- `mean()`
- `var()`
- `t.test()`

R program

t.test(wind, mu=9)

$$t_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} \sim t_{(n-1)}$$

$$H_0: \mu = 9$$

$$H_1: \mu \neq 9$$

- Two sided t test with unknown variance.
- **n>30.**

Output (C.I)

```
> t.test(wind,mu=0)
```

One Sample t-test

```
data: wind
```

```
t = 34.961, df = 152, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
9.394804 10.520229
```

```
sample estimates:
```

```
mean of x
```

```
9.957516
```

$$\left[\bar{x} - t \sqrt{\frac{s^2}{n}}, \bar{x} + t \sqrt{\frac{s^2}{n}} \right] = [9.394, 10.520]$$

Not relevant

Output (test of hypothesis)

```
t.test(wind,mu=9,var.equal=TRUE)
```

One Sample t-test

```
data: wind
```

```
t = 3.3619, df = 152, p-value = 0.0009794
```

```
alternative hypothesis: true mean is not equal to 9
```


```
95 percent confidence interval:
```

```
9.394804 10.520229
```

```
sample estimates:
```

```
mean of x
```

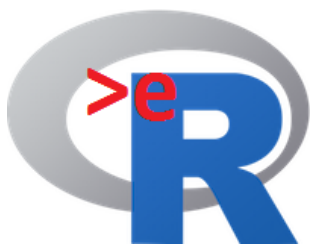
```
9.957516
```



$H_0: \mu = 9$

$H_1: \mu \neq 9$

A two sided test



The airquality data

Part 3

The Rmd program

R program for the analysis: Case Study_Prog1_V1.Rmd

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' and is in the 'Source' view. The code is as follows:

```
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)
35 library(data.table)
```

Annotations in the image include:

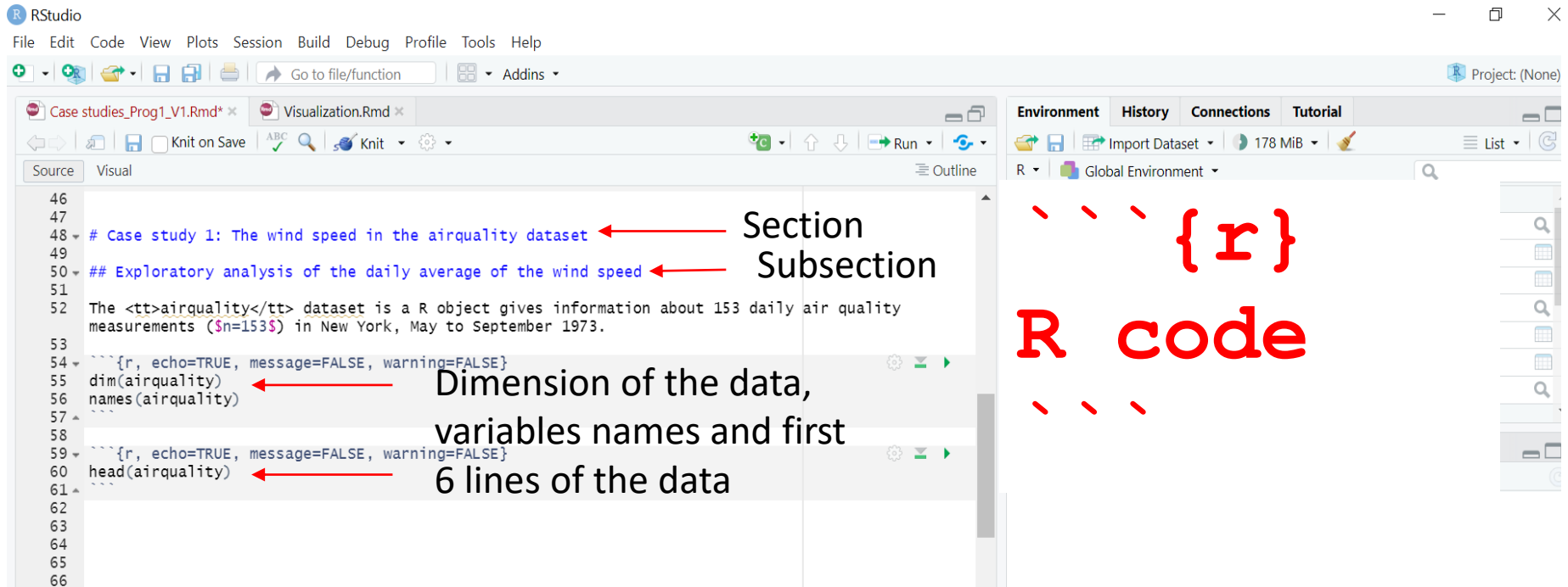
- A red bracket on the right side of the document setup section (lines 2-8) with the text "Document setup."
- A red bracket on the right side of the library loading section (lines 19-34) with the text "Many R packages, not all needed."

The right-hand pane shows the 'Environment' tab with a table of loaded 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 status bar shows the system clock as 10:55 on 21/05/2024, and the temperature as 18°C.

The Rmd program



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

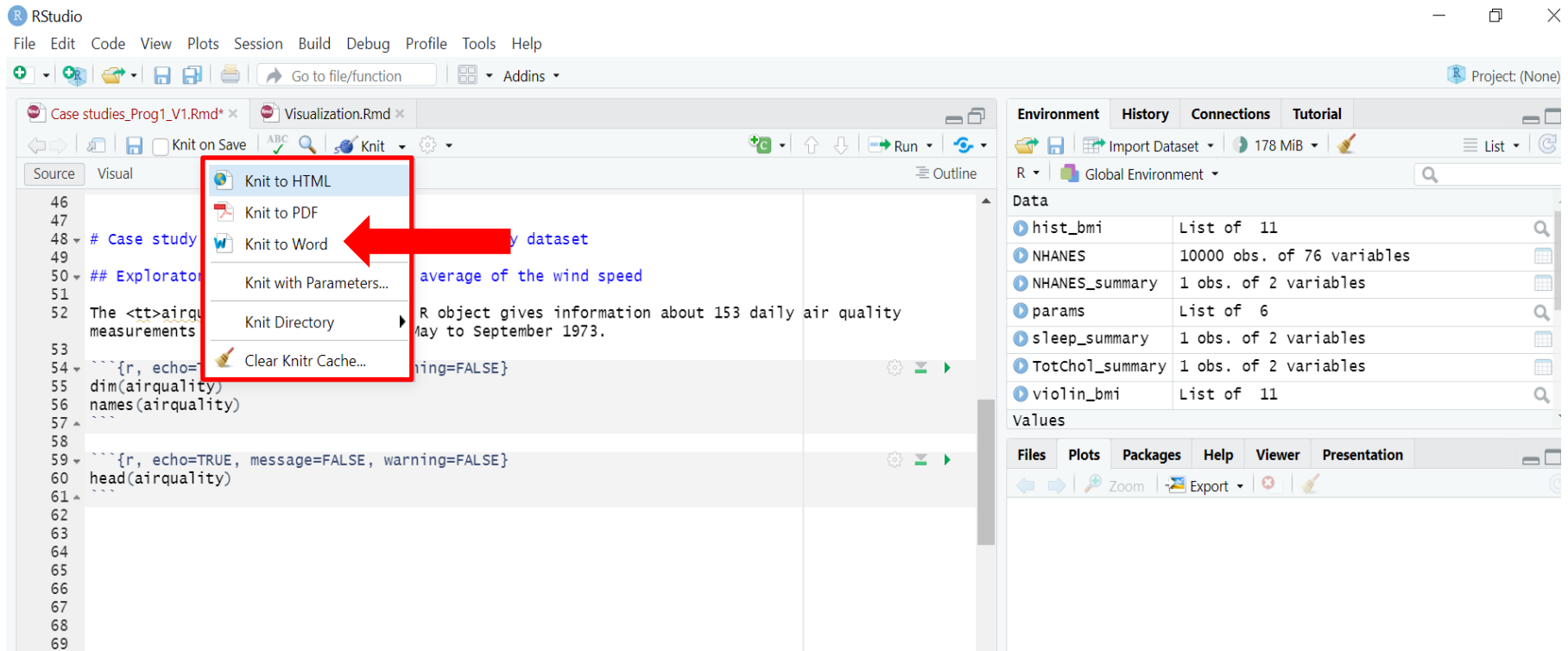
```
46  
47  
48 # Case study 1: The wind speed in the airquality dataset  
49 ## Exploratory analysis of the daily average of the wind speed  
50  
51  
52 The <tt>airquality</tt> dataset is a R object gives information about 153 daily air quality  
   measurements ( $n=153$ ) in New York, May to September 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 on the right side of the code editor:

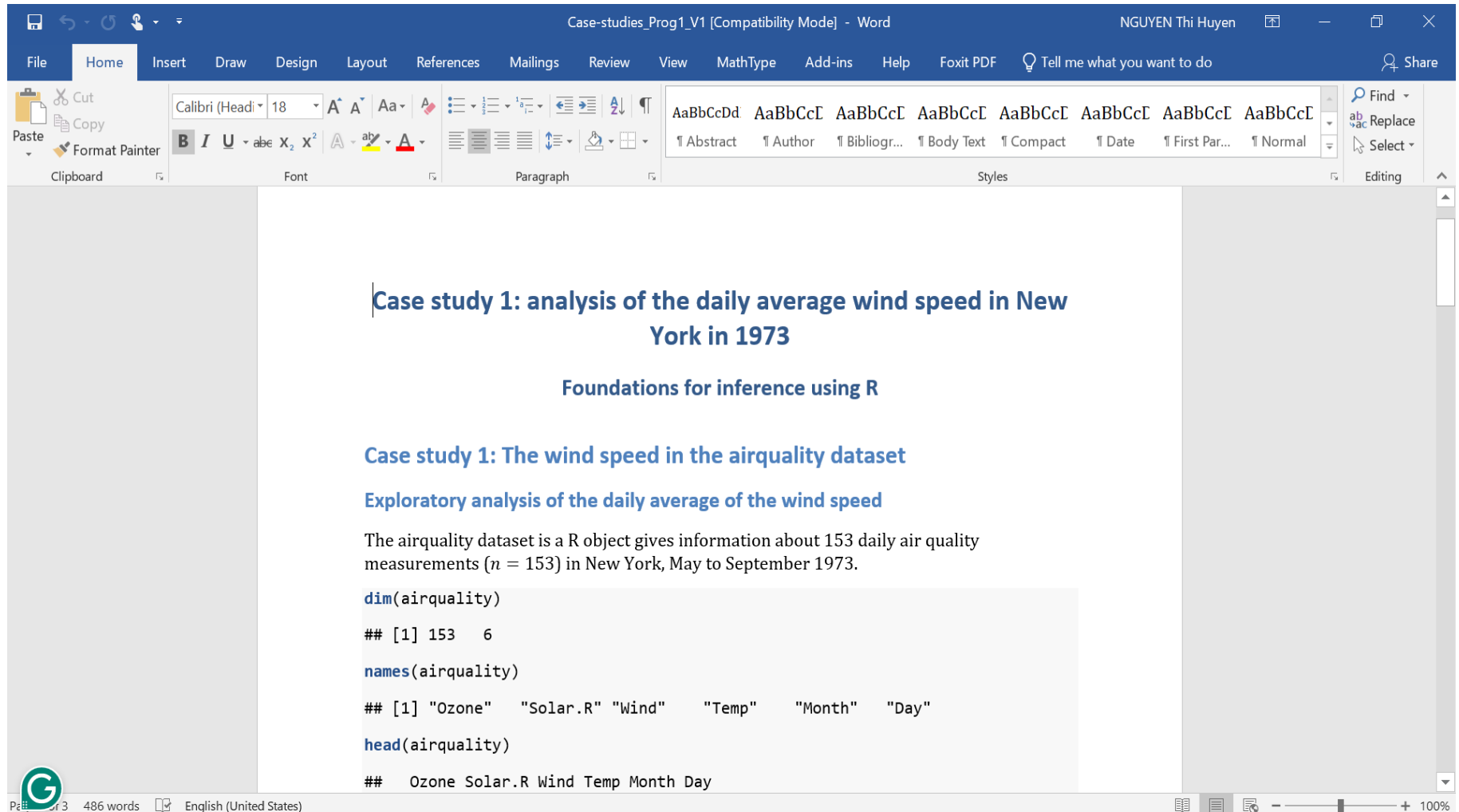
- Section**: Points to line 48 with a red arrow.
- Subsection**: Points to line 49 with a red arrow.
- Dimension of the data, variables names and first 6 lines of the data**: Points to lines 55, 56, and 60 with red arrows.

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

Choose the output



The Word document output



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
```

Page 3 486 words English (United States)

100%



The airquality data


Part 4

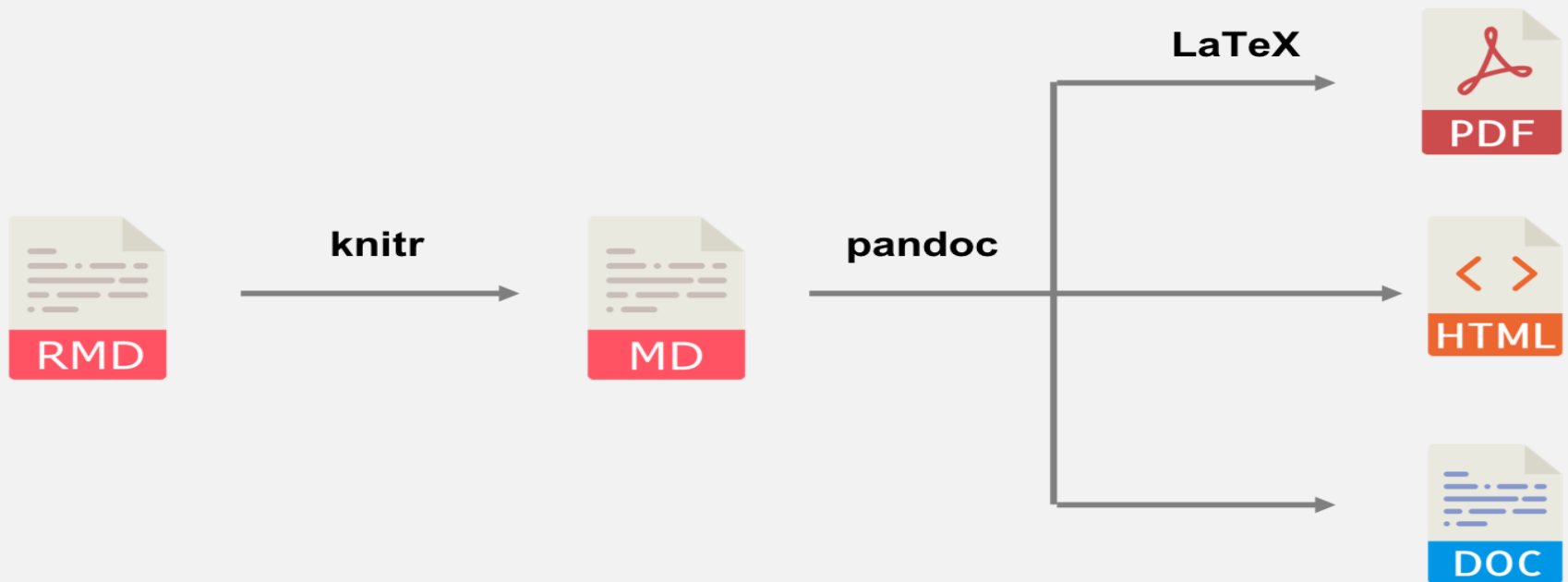
How to produce the Word output ?

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

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
```

Page 3 486 words English (United States)



Part 4.1:

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 the following components:

- Source Editor:** Contains R code for setting up a Word document. A red arrow points to the `word_document: default` line in the `output:` section.
- Environment Panel:** Shows the global environment with a list of data objects and their sizes.
- Viewer Panel:** Displays a preview of the generated Word document.
- Console:** Shows the output of the R code execution.

Source Editor 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)
```

Environment Panel Data:

| Object | Size |
|-----------------|----------------------------|
| 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 |

Viewer Panel Preview:

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
```

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 table of loaded 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 'R 4.2.1' and the current working directory '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 4.2:

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. 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`.
- free text:** Points to a paragraph of text enclosed in a red box: "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 right-hand pane shows the Environment tab with a list of objects in the Global Environment:

| 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 |

The bottom status bar shows the R version (4.2.1) and the current file path.

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

Section, subsection, subsubsection

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

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

free text

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
```

The code is shown as a part of the output

Page 3 491 words English (United States) 100%

Code in the Rmd file

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for an Rmd file. The code includes a header, a paragraph about wind speed data, and a ggplot2 histogram with density. A red box highlights the following code block:

```
67 {r, echo=TRUE, message=FALSE, warning=FALSE, fig.cap="Histogram with density of wind speed"}>
68 ggplot(airquality, aes(x = wind)) +
69   geom_histogram(aes(y = ..density..), fill = "skyblue", color = "black") +
70   geom_density(alpha = 0.2, fill = "orange") +
71   ylab("Density")
72 }
```
- Environment Panel:** Shows the Global Environment with a list of objects: hist_bmi, NHANES, NHANES_summary, params, sleep_summary, TotCho1_summary, and violin_bmi.
- Console:** Displays the R version (4.2.1) and the current working directory (C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/).
- Text Overlay:** A red arrow points from the text "The code for histogram" to the highlighted ggplot2 code block.

The output in the Word file

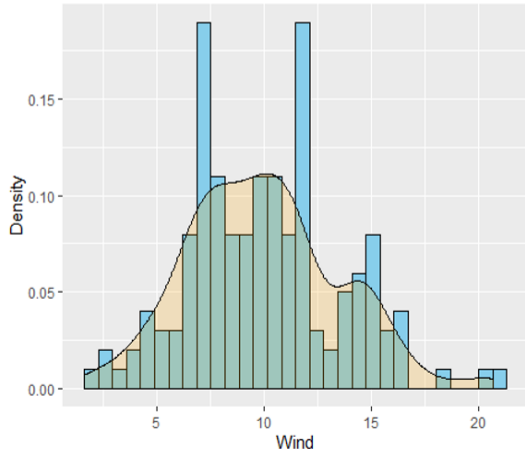
Case-studies_Prog1_V1 [Compatibility Mode] - Word

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 starting at line 89 is highlighted with a red box. The text "The code for confidence interval." is overlaid on the right side of the code block.

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 settings (Calibri, size 18) and paragraph styles. The main text area contains the following content:

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 code block, with a red rectangle highlighting the 95 percent confidence interval:

```
##
## 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 code block, the text 'The output for confidence interval.' is displayed.

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

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


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 # 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)
```

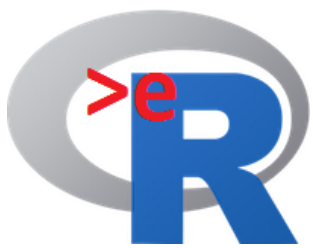
```
##  
## One Sample z-test  
##  
## data: wind  
## z = 3.3619, n = 153.00000, Std. Dev. = 3.52300, Std. Dev. of the sample  
## mean = 0.28482, p-value = 0.0007742  
## alternative hypothesis: true mean is not equal to 9  
## 95 percent confidence interval:  
## 9.399284 10.515749  
## sample estimates: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

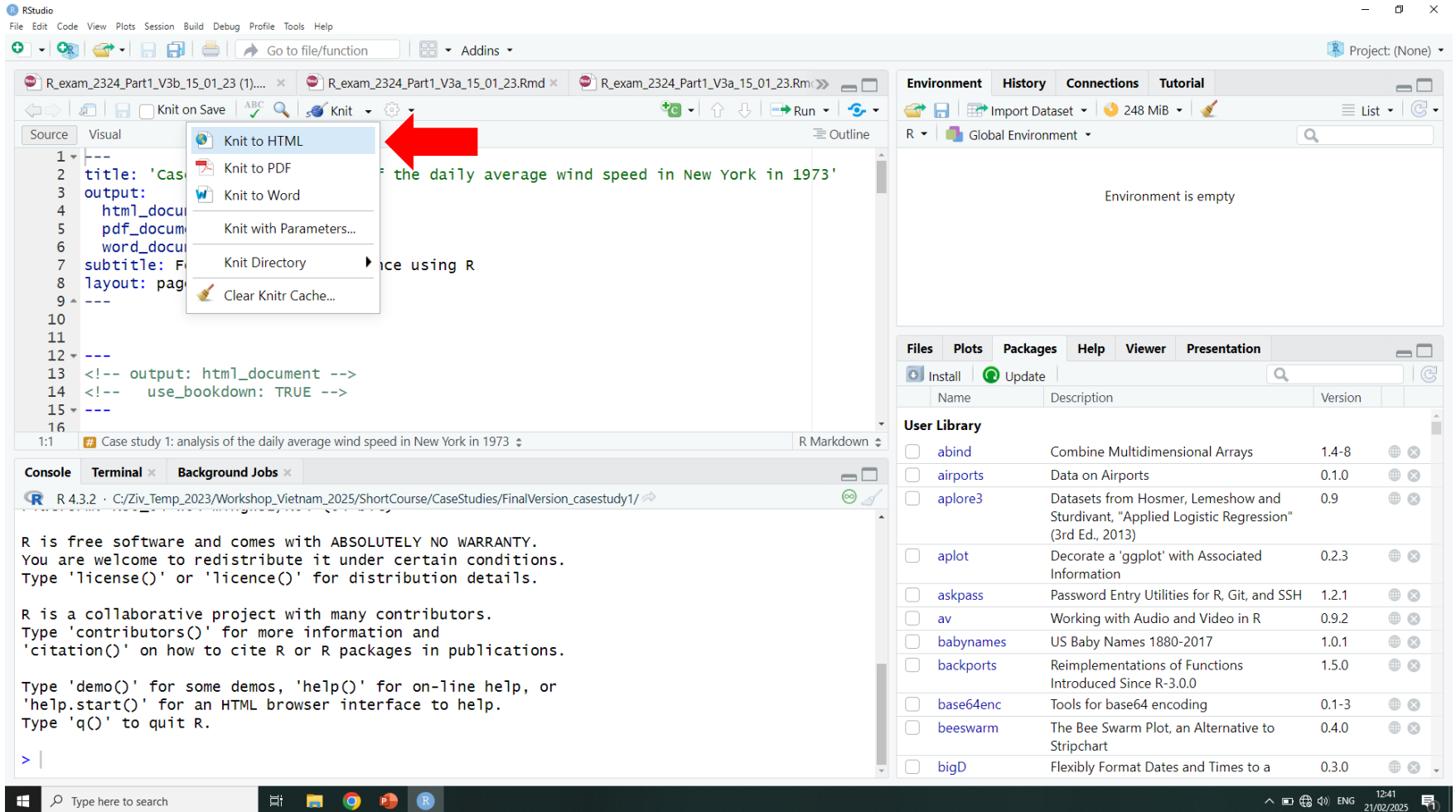
Windows Taskbar: Presentation, Visualization, Downloads, RStudio, eR_biostat_i..., Meeting 21..., Case-studie..., 19°C, 12:56, 21/05/2024



Part 5:

How to set up the HTML file?

Setup the HTML document



The screenshot shows the RStudio interface with the 'Knit' menu open. A red arrow points to the 'Knit to HTML' option. The source editor shows a YAML header for a document titled 'Case study 1: analysis of the daily average wind speed in New York in 1973'. The console shows the R startup message.

Source Editor:

```
1 ---  
2 title: 'Case study 1: analysis of the daily average wind speed in New York in 1973'  
3 output:   
4   html_document  
5   pdf_document  
6   word_document  
7 subtitle:   
8 layout: page  
9 ---  
10  
11  
12  
13 <!-- output: html_document -->  
14 <!-- use_bookdown: TRUE -->  
15  
16
```

Knit Menu:

- Knit to HTML
- Knit to PDF
- Knit to Word
- Knit with Parameters...
- Knit Directory
- Clear Knitr Cache...

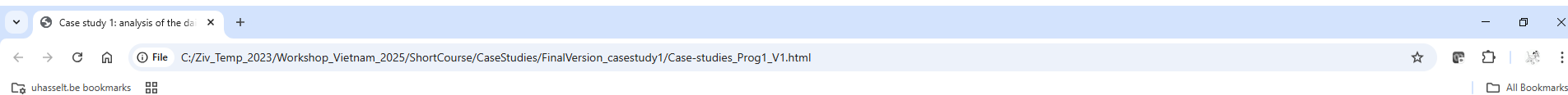
Environment Panel:

Environment is empty

User Library:

| Name | Description | Version |
|------------------------------------|--|---------|
| <input type="checkbox"/> abind | Combine Multidimensional Arrays | 1.4-8 |
| <input type="checkbox"/> airports | Data on Airports | 0.1.0 |
| <input type="checkbox"/> aplore3 | Datasets from Hosmer, Lemeshow and Sturdivant, "Applied Logistic Regression" (3rd Ed., 2013) | 0.9 |
| <input type="checkbox"/> aplot | Decorate a 'ggplot' with Associated Information | 0.2.3 |
| <input type="checkbox"/> askpass | Password Entry Utilities for R, Git, and SSH | 1.2.1 |
| <input type="checkbox"/> av | Working with Audio and Video in R | 0.9.2 |
| <input type="checkbox"/> babynames | US Baby Names 1880-2017 | 1.0.1 |
| <input type="checkbox"/> backports | Reimplementations of Functions Introduced Since R-3.0.0 | 1.5.0 |
| <input type="checkbox"/> base64enc | Tools for base64 encoding | 0.1-3 |
| <input type="checkbox"/> beeswarm | The Bee Swarm Plot, an Alternative to Stripchart | 0.4.0 |
| <input type="checkbox"/> bigD | Flexibly Format Dates and Times to a | 0.3.0 |

The output



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)
```

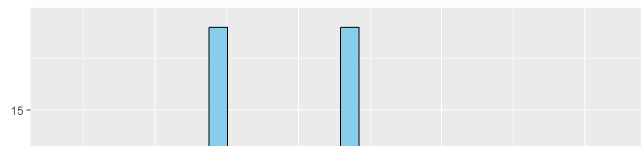
```
## [1] 153 6
```

```
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
```

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 data. Figure 1 shows histogram of wind speed.

```
ggplot(airquality, aes(x = Wind)) +
  geom_histogram(fill = "skyblue", color = "black")+
  ylab("Frequency")
```



The Rmd file

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

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

Red annotations highlight 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`.
- free text**: points to the paragraph starting with "The `airquality` dataset is a R object gives information about 153 daily air quality measurements ($n=153$) in New York, May to September 1973."
- R code**: points to the R code block starting with `{r, echo=TRUE, message=FALSE, warning=FALSE}`.

The Environment pane shows "Environment is empty". The Packages pane lists the following user libraries:

| Name | Description | Version |
|-----------|--|---------|
| abind | Combine Multidimensional Arrays | 1.4-8 |
| airports | Data on Airports | 0.1.0 |
| aplore3 | Datasets from Hosmer, Lemeshow and Sturdivant, "Applied Logistic Regression" (3rd Ed., 2013) | 0.9 |
| aplot | Decorate a 'ggplot' with Associated Information | 0.2.3 |
| askpass | Password Entry Utilities for R, Git, and SSH | 1.2.1 |
| av | Working with Audio and Video in R | 0.9.2 |
| babynames | US Baby Names 1880-2017 | 1.0.1 |
| backports | Reimplementations of Functions Introduced Since R-3.0.0 | 1.5.0 |
| base64enc | Tools for base64 encoding | 0.1-3 |
| beeswarm | The Bee Swarm Plot, an Alternative to Stripchart | 0.4.0 |
| bigD | Flexibly Format Dates and Times to a | 0.3.0 |

The Console pane shows the R startup message:

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> |
```

The output

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)
```

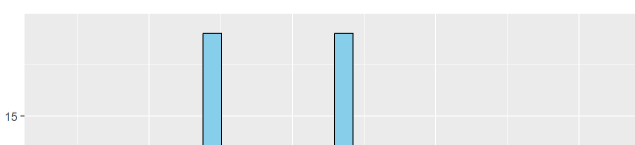
```
## [1] 153 6
```

```
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 |

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 data. Figure 1 shows histogram of wind speed.

```
ggplot(airquality, aes(x = Wind)) +  
  geom_histogram(fill = "skyblue", color = "black")+  
  ylab("Frequency")
```



section

subsection

free text

R code and

Discussion

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