

Publication-Ready Data Analysis with R

E. F. Haghish

24/2/2021



Material

- ▶ The material for this lecture are hosted on GitHub:
 - ▶ <https://github.com/haghighi/promenta>
- ▶ an example of and **organized project** for running a simple CFA analysis and producing a dynamic document
- ▶ The current slides
- ▶ Rmarkdown template for creating PDF slides within RStudio

Overview

- ▶ Automate the process of data analysis
 - ▶ organizing a computational project
 - ▶ reproduce the *entire* analysis
- ▶ Automate the process of reporting
 - ▶ producing a sensible analysis reports for a manuscript
 - ▶ dynamic tables, dynamic graphs, and dynamic text
 - ▶ discuss its necessity

Other important issues that are beyond this lecture

- ▶ Integrating version control
- ▶ Pre-planned analysis
- ▶ Automating third-party statistics software such as **MPlus** within **R** or **Stata**



Figure 1: Relevant literature

Statement of problem

Garfield (1995) defines learning statistics as follows:

1. learning to communicate using statistical language
2. solving statistical problems
3. drawing conclusions
4. supporting conclusions with statistical reasoning

requires:

- ▶ in-depth understanding of statistical concepts
- ▶ statistical reasoning
- ▶ computer programming skills

- ▶ Statistics generally causes inconvenience for researchers of different fields (Baloglu, 2003)
- ▶ 80% of graduate students suffer from statistics anxiety (Onwuegbuzie, 2004)
 - ▶ math anxiety
 - ▶ computer anxiety
 - ▶ programming anxiety

- ▶ Proper statistical education has been *avoided*
 - ▶ Teaching through GUI instead of programming
 - ▶ SPSS, MPlus, AMUS, LISREL... gained popularity in social sciences
 - ▶ R and Python gained popularity in natural sciences
 - ▶ How about a 5-ECTS introductory R programming course for undergrads?
- ▶ The complexity of the methods is increasing annually
 - ▶ The journals' appetite for intricate statistics is growing
- ▶ The gap between statistical education and statistical practice is increasing
- ▶ There is no statistical software that does **everything for everyone**
 - ▶ Particular analyses might be available in a special software

Problem?

We are lacking

- ▶ Basic coding education (no more mouse-and-click)
- ▶ Skills for planing and organizing data analysis
- ▶ Tracking our potential errors in different steps of research
- ▶ Communicating statistical decisions and reasons

Which results in lacking **reproducibility**

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p -values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α . Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Figure 3: Estimating the reproducibility of psychological science

- ▶ Before jumping into a conclusion, it might be worth considering
 - ▶ How far statistical approaches to psychological research can go?
 - ▶ What are the limitations of statistical methods
 - ▶ Limitations of quantitative research?
 - ▶ ...

Reproducibility in Psychological Science: When Do Psychological Phenomena Exist?

Seppo E. Iso-Ahola*

Department of Kinesiology, School of Public Health, University of Maryland, College Park, MD, United States

Scientific evidence has recently been used to assert that certain psychological phenomena do not exist. Such claims, however, cannot be made because (1) scientific method itself is seriously limited (i.e., it can never prove a negative); (2) non-existence of phenomena would require a complete absence of both logical (theoretical) and empirical support; even if empirical support is weak, logical and theoretical support can be strong; (3) statistical data are only one piece of evidence and cannot be used to reduce psychological phenomena to statistical phenomena; and (4) psychological phenomena vary across time, situations and persons. The human mind is unreproducible from one situation to another. Psychological phenomena are not particles that can decisively be tested and discovered. Therefore, a declaration that a phenomenon is not real is not only theoretically and empirically unjustified but runs counter to the propositional and provisional nature of scientific knowledge. There are only “temporary winners” and no “final truths” in scientific knowledge. Psychology is a science of subtleties in human affect, cognition and behavior. Its phenomena fluctuate with conditions and may sometimes be difficult to detect and reproduce empirically. When strictly applied, reproducibility is an overstated and even questionable concept in psychological science. Furthermore, statistical measures (e.g., effect size) are poor indicators of the theoretical importance and relevance of phenomena (cf. “deliberate practice” vs. “talent” in expert performance), not to mention whether phenomena are real or unreal. To better understand psychological phenomena, their theoretical and empirical properties should be examined via multiple parameters and criteria. Ten such parameters are suggested.

Computer science faces an ethics crisis. The Cambridge Analytica scandal proves it.



NEW YORK TIMES

Facebook founder Mark Zuckerberg speaks at a conference in San Jose, Calif., in 2017. Cambridge Analytica scraped up Facebook data from more than 50 million people.

By Yonatan Zunger | MARCH 22, 2018

CAMBRIDGE ANALYTICA BUILT a weapon. They did so understanding what uses its buyers had for it, and it worked exactly as intended. To help clients manipulate voters, the company built psychological profiles from data that it surreptitiously harvested from the accounts of 50 million Facebook users. But what Cambridge Analytica did

Reproducibility vs. Replication

- ▶ The two terms have been used interchangeably (Loscalzo, 2012)
- ▶ They have different meanings in different fields of science
- ▶ Replication requires re-implementing experiments by other research groups (Baggerly & Berry, 2009)
 - ▶ using either the same methodology or alternatives
- ▶ Problems with replication?

Reproducibility

- ▶ Baggerly & Berry (2009):
 - ▶ reproducibility is replicating the computation by an independent researcher
 - ▶ using the same data, programmed code, procedure, and methodology
 - ▶ and without requiring any further assistance or information from the author (King, 1995)
 - ▶ the least standard for evaluating the quantitative results
 - ▶ reproducibility does not guarantee (Peng, 2011; Stodden, et. al., 2014):
 - ▶ quality
 - ▶ sound methodology
 - ▶ accurate data collection
 - ▶ validity of the findings
- ▶ reproducibility grants limited transparency (Gentleman & Lang, 2012)
 - ▶ validate the computational procedure
 - ▶ check or adapt the claims in the scientific publication

Sources of error in research

- ▶ Errors can happen at any stage of research
 - ▶ study design
 - ▶ data collection
 - ▶ digitizing the data from questionnaires to a computer
 - ▶ cleaning the data
 - ▶ preparing the data for analysis
 - ▶ choice of methodology
 - ▶ adjustment options, analytical assumptions, algorithms, etc. . .
 - ▶ interpreting the results
 - ▶ reporting the results in the publication
 - ▶ copy-pasting from statistical software to MS Word
 - ▶ any problem with that?
 - ▶ ...
- ▶ Or afterwards, such as publication bias, etc. . .

Collaboration on computational research

- ▶ The majority of statistical contributions do not appear in the manuscripts
 - ▶ no code, no data checking, no quality assurance, ...
- ▶ Lacking reproducibility means no collaboration on statistical analyses
- ▶ Collaboration on statistical analysis is like collaboration on software:
 - ▶ well-structured
 - ▶ automatized
 - ▶ well-documented
 - ▶ dependencies are carefully planned, organized, documented

Costs

- ▶ You need to learn new tricks and let go of old habits
- ▶ No one gives you credit for being transparent
- ▶ Transparency means your mistakes can be revealed by others
 - ▶ Shame or gratitude?
 - ▶ What you cannot reproduce your own analysis?
 - ▶ How would you feel about sharing your code?
- ▶ Reproducibility is human problem, not computers

Automated Data Analysis

- ▶ Automated data analysis means making data analysis reproducible
 - ▶ writing analysis code to track **the entire data analysis**
 - ▶ setup and organize your analytic project
 - ▶ Operating system and statistical software
 - ▶ Add-on packages
 - ▶ Data management
 - ▶ Nesting analysis code (and why should you)
 - ▶ Communicating the analysis



Figure 6: The procedure we are intending to automatize

Organizing the computation

- ▶ The rule is to be disciplined, **very disciplined**
 - ▶ Keep track of changes in code, data, and analysis results
- ▶ There is no template to be applicable to all projects
 - ▶ with different types of data, there will be different procedures and workflows
- ▶ Rule of thumb:
 - ▶ protect your raw data
 - ▶ keep track of all the code for preparing the data for analysis
 - ▶ keep track of all the analysis code
 - ▶ create separate directories for storing raw data, code, analysis results/reports, documents, etc.

Example 1: R package

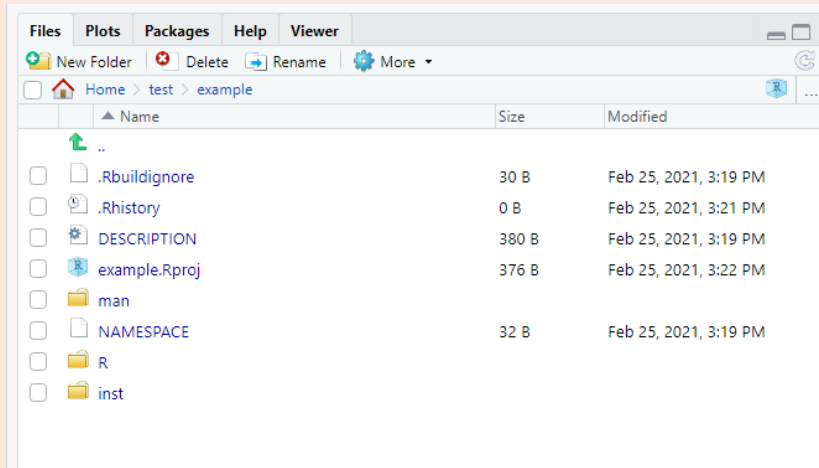


Figure 7: R Package Structure

Example 2: UiN Project







Name	Date modified	Type	Size
 anadata	11.02.2021 08:44	File folder	
 clean	08.02.2021 12:28	File folder	
 code	22.02.2021 10:30	File folder	
 docs	08.02.2021 11:35	File folder	
 raw	12.02.2021 22:43	File folder	
 README	19.02.2021 09:44	Text Document	1 KB

Figure 8: Young in Norway Study

Example 3: My personal preference








Name	Date modified	Type	Size
 code	25.02.2021 15:30	File folder	
 data	22.02.2021 11:46	File folder	
 docs	25.02.2021 15:30	File folder	
 report	25.02.2021 15:30	File folder	
 results	25.02.2021 15:30	File folder	
 MAIN	22.02.2021 12:00	DO File	1 KB
 README	09.02.2021 13:23	Text Document	1 KB

Figure 9: My way of organizing a computational project

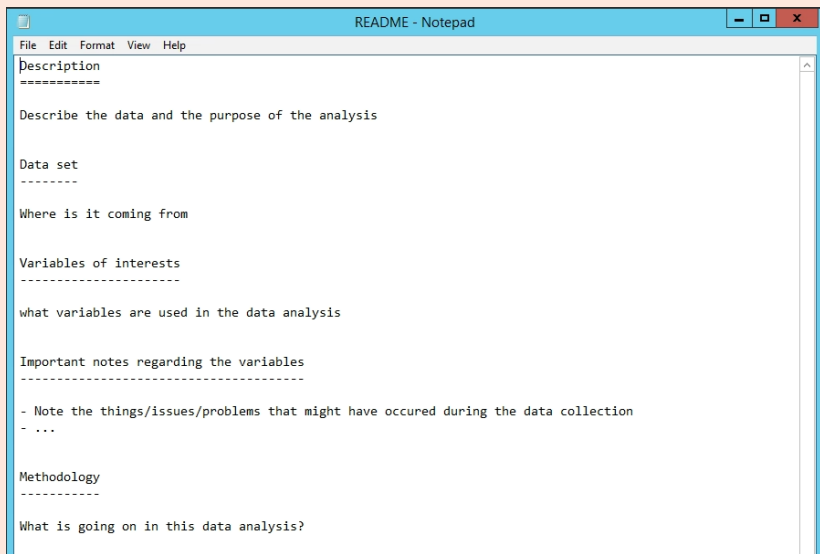
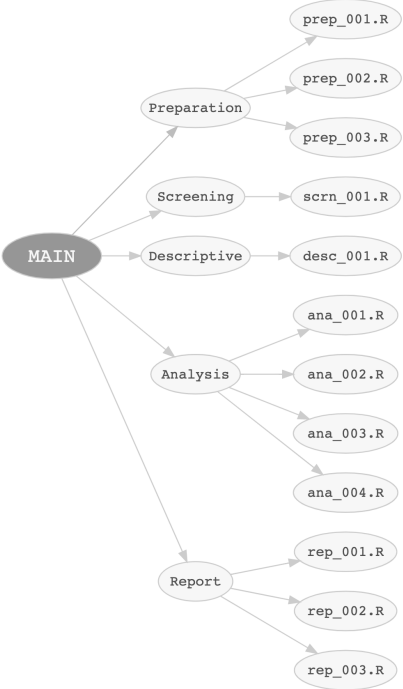


Figure 10: Writing a README file that is worth reading

Nesting script files

- ▶ The idea is borrowed from computer science
 - ▶ For example, see how Linux kernel is compiled
 - ▶ see the **Makefile** in <https://github.com/torvalds/linux>
 - ▶ the file provides all of the orders to compile Linux from the source code
- ▶ We apply the same discipline to approach reproducibility
 - ▶ There will be a single file that provides the instructions to rerun the entire data analysis
 - ▶ I name that file **MAIN**, you name it ...
 - ▶ the file will source all other script files used for preparing, analyzing, and reporting the analysis
- ▶ Nesting works best with relative file paths (instead of absolute paths)
 - ▶ begin the **MAIN** file by setting the working directory:
 - ▶ Use `setwd()` in **R**, `cd` in **Stata** and **SPSS**



Nesting script files

R

```
source('./code/preparation/prep_001.R')
```

Stata

```
do './code/preparation/prep_001.do'
```

SPSS

```
INSERT FILE='./code/preparation/prep_001.sps'.
```

Notes: general suggestions

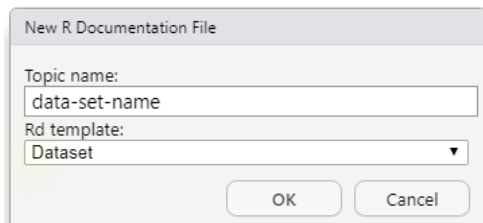
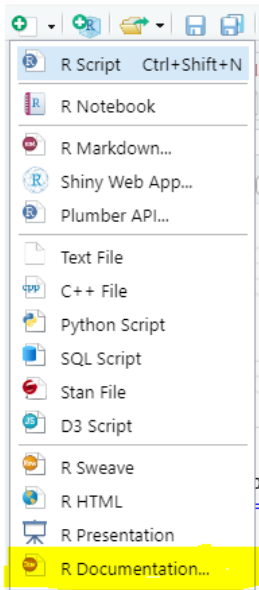
1. The raw data is kept untouched
- ▶ Store time-consuming operations in a different directory (e.g. *anadata*)
2. Organize your code under subdirectories (if you write many files?)
3. Save the results (analysis outputs) in separate directories and name them properly within the code
4. Name and document your data file, especially if it is going to receive further updates in the future
5. Document the software dependencies (Operating system, R/Stata/etc. version, **ALL add-on packages'** versions)
- ▶ check for example `lavaan` change history: <https://lavaan.ugent.be/history/>
6. Document the data set
- ▶ use `datadoc` (Haghighi, 2020) command for **Stata** or `Rd` documentation from **RStudio**

Notes: data documentation

- ▶ CRAN recommends the following documentation section for a data set
 1. Title, the label of the dataset, and where it was published (package name)
 2. Description
 3. Format, including a table summarizing the variables' types and labels
 4. Notes attached to the dataset or the variables (for Stata only)
 5. The source of the data; that is, where they are coming from
 6. References, if any
 7. Examples, if needed

R example - RStudio

File Edit Code View Plots



Automated Analysis Reporting

Avoiding manual reporting

I noted that errors can happen in the process of reporting

Sources of error in research

- ▶ Errors are everyday and can happen at any stage of research
 - ▶ ...
 - ▶ interpreting the results
 - ▶ reporting the results in the publication
 - ▶ copy-paste from statistical software to MS Word
 - ▶ updating the report after making a change in the data or analysis
 - ▶ ...

Avoiding manual reporting

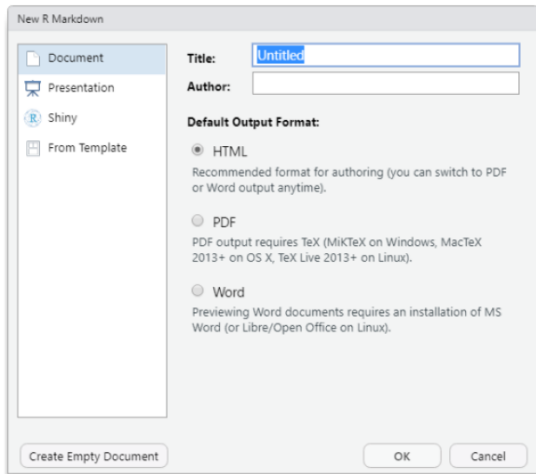
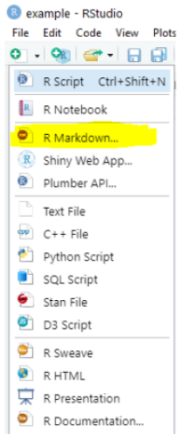
- ▶ A solution is to do the data analysis and write the analysis report at the same time
- ▶ This is a paradigm borrowed from computer science, for solving software documentation problem
 - ▶ documentation is written within code files using special comment signs
 - ▶ next, a program extracts and renders the documentation and updates the documents (Knuth 1983)
- ▶ There are software for generating data analysis reports:
 - ▶ for **R**, use `rmarkdown` (Yihui, et. al., 2018)
 - ▶ for **Stata**, use `markdoc` (haghighi, 2016)
 - ▶ for **SPSS**, no *equivalent* exists, although `StatTag` offers adding dynamic values from SPSS
 - ▶ **StatTag** is not recommended because it does not provide any insight about the reproducibility
 - ▶ both provide a restricted framework to examine the reproducibility of the code

rmarkdown package

- ▶ rmarkdown is a general purpose literate programming software
- ▶ developed particularly for R
- ▶ rmarkdown is versatile:
 - ▶ generate publication-ready analysis report in various document formats (PDF, Docx, ODT, HTML, LaTeX, etc.)
 - ▶ includes a syntax highlighter
 - ▶ generate dynamic presentation slides
 - ▶ generate dynamic R help files and package vignette
- ▶ Analysis documentation/interpretation is written within *Rmd* files

Who can use `rmarkdown`?

1. Students - as early as introductory statistics courses - can use `rmarkdown` to actively take note inside RStudio
2. University lecturers who teach statistics using R, can use `rmarkdown` to generate PDF slides, educational materials
3. Statisticians can use `rmarkdown` for creating dynamic analysis reports
4. Researchers can use it to create publication-ready Microsoft Word or LaTeX documents



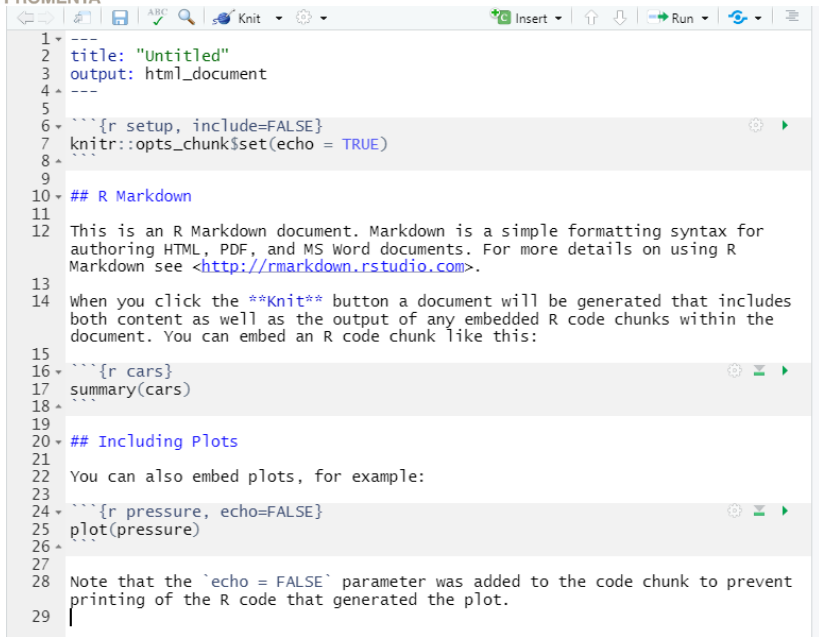
Supported Markup Languages

- ▶ `rmarkdown` supports several languages such as **Markdown**, **LaTeX**, and **HTML**
- ▶ In this lecture we will focus on Markdown, which is the simplest. The following links, from its developer's site, can provide a good background about Markdown:
 - ▶ <https://daringfireball.net/projects/markdown/>
 - ▶ <https://daringfireball.net/projects/markdown/syntax>
 - ▶ <https://daringfireball.net/projects/markdown/dingus>
- ▶ Markdown is:
 - ▶ minimalistic and clean
 - ▶ simple to read and write
 - ▶ helps to focus on the content
 - ▶ can be converted to many formats

Markdown syntax	Result
Heading 1 =====	Heading 1
Heading 2 -----	Heading 2
###Heading 3	Heading 3
####Heading 4	Heading 4
plain text paragraph	plain text paragraph
> text	block quote
Bold or __Bold__ text	Bold or Bold text
<i>*Italic*</i> or <i>_Italic_</i> text	<i>Italic</i> or <i>Italic</i> text
`monospace` text	monospace text
superscript ²	superscript ²
---	horizontal rule
1. Ordered item1 A. Sublist 1 a. Subsublist 1 2. Ordered item2	1. Ordered item1 A. Sublist 1 a. Subsublist 1 2. Ordered item2
* Unordered item1 * Sublist 1 * Subsublist 1 * Unordered item2	• Unordered item1 - Sublist 1 * Subsublist 1 • Unordered item2
![Text] (filename)	Insert an image with description
[Text] (http://url)	Insert a hyperlink

Planning the analysis report

- ▶ **R** script files have an `.r` extension, `rmarkdown` files have `.rmd` extension
- ▶ Human language and computer languages are separated from one another
- ▶ You can nest the analysis into multiple Rmd files
- ▶ You have **full control** about what to include or exclude in your document
 - ▶ dynamic text
 - ▶ R code
 - ▶ R output
 - ▶ graphs
 - ▶ tables
 - ▶ mathematical notations



```
1 ---
2 title: "Untitled"
3 output: html_document
4 ---
5
6 ```{r setup, include=FALSE}
7 knitr::opts_chunk$set(echo = TRUE)
8 ```
9
10 ## R Markdown
11
12 This is an R Markdown document. Markdown is a simple formatting syntax for
13 authoring HTML, PDF, and MS Word documents. For more details on using R
14 Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the Knit button a document will be generated that includes
17 both content as well as the output of any embedded R code chunks within the
18 document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
31
32 Note that the `echo = FALSE` parameter was added to the code chunk to prevent
33 printing of the R code that generated the plot.
```

Figure 14: Rmarkdown example

```

1 ---
2 title: "Title of the project"
3 author: "E. F. Haghish"
4 date: "February 28, 2021"
5 output:
6   html_document:
7     toc: yes
8   word_document:
9     toc: yes
10 ---
11
12 The data analysis is organized in several Rmd files. Each of these files can
13 be executed **independently**, following their numeric order (i.e. `01_`, `02_`,
14 `03_`, etc). However, the **MAIN.Rmd** file execute them all within a single
15 document.
16
17 ### Software versions
18
19 - The analysis is caried out in `R version x.x.x` on `Mac OSX version x.x`
20 - The following R packages are also required:
21   + `package name version x.x.x.`
22   + `package name version x.x.x.`
23   + ...
24
25 ---|
26
27 {r child = '01_preparation.Rmd'}
28
29
30 {r child = '02_descriptive.Rmd'}
31
32
33 {r child = '03_analysis.Rmd'}
34
35
36

```

Dynamic tables

- ▶ **markdown** language offers syntax for creating tables, but it is tedious <https://pandoc.org/MANUAL.html#Tables>
- ▶ we can create tables by constructing a string matrix and converting it to a table
- ▶ consider the output document before designing your table:
 - ▶ for MS Word use **markdown** only
 - ▶ for LaTeX and HTML you have plenty of option
- ▶ For general purpose, I recommend **pander**
 - ▶ for different *classess*, **pander** offers automated table designs

References

- ▶ Garfield, J. (1995). How students learn statistics. *International Statistical Review / Revue Internationale de Statistique*, 63 (1), 25-34. Retrieved from <http://www.jstor.org/stable/1403775>
- ▶ Baloglu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences*, 34 (5), 855-865.
- ▶ Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assessment & Evaluation in Higher Education*, 29 (1), 3-19.
- ▶ Loscalzo, J. (2012). Irreproducible experimental results: Causes, (mis)interpretations, and consequences. *Circulation*, 125 (10), 1211-1214. Retrieved from <http://circ.ahajournals.org/content/125/10/1211.short> doi: 10.1161/CIRCULATIONAHA.112.098244
- ▶ Baggerly, K. A., & Berry, D. A. (2009). Reproducible research.
- ▶ Peng, R. D. (2011). Reproducible research in computational science. *Science* (New York, Ny), 334 (6060), 1226.
- ▶ Stodden, V., Leisch, F., & Peng, R. D. (2014). Implementing reproducible research. CRC Press.
- ▶ Gentleman, R., & Lang, D. T. (2012). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*.

- ▶ Knuth, D. E. 1983. The WEB system of structured documentation. Technical Report STAN-CS-83-980, Department of Computer Science, Stanford University. <http://infolab.stanford.edu/pub/cstr/reports/cs/tr/83/980/CS-TR-83-980.pdf>
- ▶ Xie, Yihui, Joseph J. Allaire, and Garrett Grolemond (2018). R markdown: The definitive guide. CRC Press.
- ▶ Haghish E. F. (2016). Markdoc: Literate Programming in Stata. The Stata Journal, 16(4):964-988. doi:10.1177/1536867X1601600409
- ▶ Haghish, E. F. (2020). Software documentation with markdoc 5.0. The Stata Journal, 20(2), 336-362.
- ▶ Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).
- ▶ Iso-Ahola, S. E. (2017). Reproducibility in psychological science: When do psychological phenomena exist?. Frontiers in Psychology, 8, 879.

- ▶ Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.
- ▶ Zunger, J. (2018). Computer science faces an ethics crisis. The Cambridge Analytica scandal proves it. Boston Globe, 22.