Reproducible Research

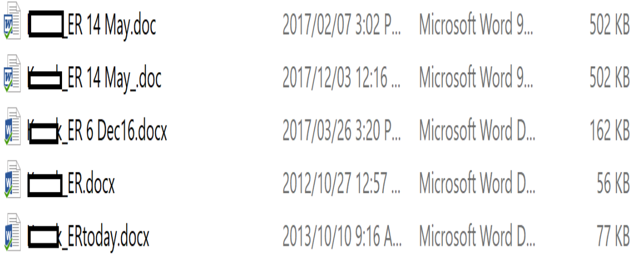
Șebnem Er

University of Cape Town, Statistical Sciences Department

Sebnem.Er@uct.ac.za

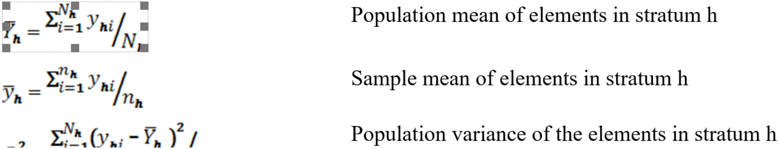
December 05, 2017

Examples from my own experiences: Documentation



[redacted]_ER 14 May.doc	2017/02/07 3:02 P...	Microsoft Word 9...	502 KB
[redacted]_ER 14 May_.doc	2017/12/03 12:16 ...	Microsoft Word 9...	502 KB
[redacted]_ER 6 Dec16.docx	2017/03/26 3:20 P...	Microsoft Word D...	162 KB
[redacted]ER.docx	2012/10/27 12:57 ...	Microsoft Word D...	56 KB
[redacted]_ERtoday.docx	2013/10/10 9:16 A...	Microsoft Word D...	77 KB

Examples from my own experiences: Documentation

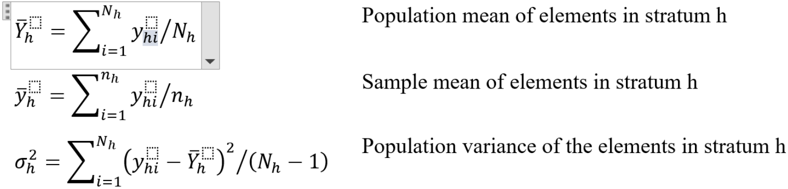


$\bar{Y}_h = \sum_{i=1}^{N_h} y_{hi} \big/ N_h$ Population mean of elements in stratum h

$\bar{y}_h = \sum_{i=1}^{n_h} y_{hi} \big/ n_h$ Sample mean of elements in stratum h

$\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 /$ Population variance of the elements in stratum h

Examples from my own experiences: Documentation



$$\bar{Y}_h^{\square} = \sum_{i=1}^{N_h} y_{hi}^{\square}/N_h$$

Population mean of elements in stratum h

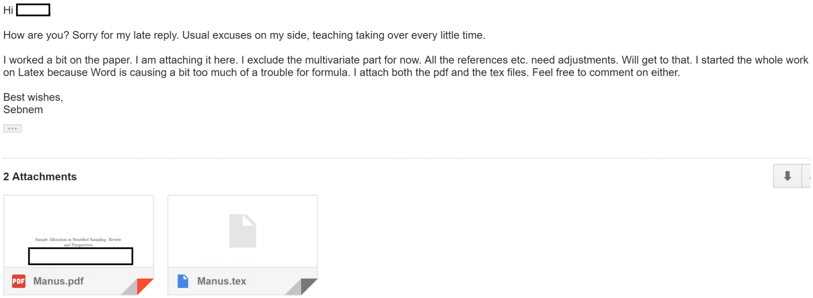
$$\bar{y}_h^{\square} = \sum_{i=1}^{n_h} y_{hi}^{\square}/n_h$$

Sample mean of elements in stratum h

$$\sigma_h^2 = \sum_{i=1}^{N_h} \left(y_{hi}^{\square} - \bar{Y}_h^{\square}\right)^2/(N_h - 1)$$

Population variance of the elements in stratum h

Examples from my own experiences: Documentation



Hi [redacted]

How are you? Sorry for my late reply. Usual excuses on my side, teaching taking over every little time.

I worked a bit on the paper. I am attaching it here. I exclude the multivariate part for now. All the references etc. need adjustments. Will get to that. I started the whole work on Latex because Word is causing a bit too much of a trouble for formula. I attach both the pdf and the tex files. Feel free to comment on either.

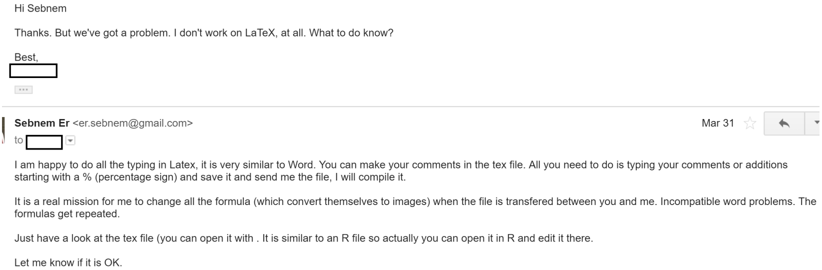
Best wishes,
Sebnem

2 Attachments

Sample Allocation in Stratified Sampling: Review and Perspectives

[redacted]

Examples from my own experiences: Documentation



Hi Sebnem

Thanks. But we've got a problem. I don't work on LaTeX, at all. What to do know?

Best,
[redacted]

Sebnem Er <er.sebnem@gmail.com>
to [redacted]

Mar 31

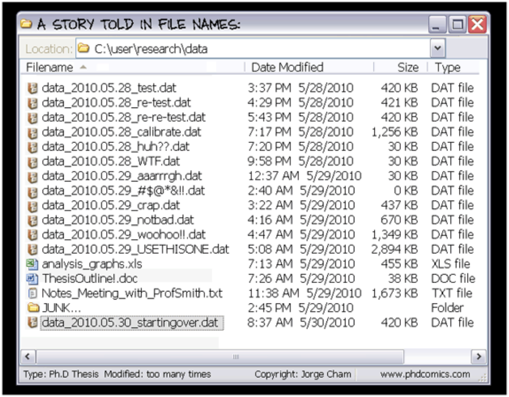
I am happy to do all the typing in Latex, it is very similar to Word. You can make your comments in the tex file. All you need to do is typing your comments or additions starting with a % (percentage sign) and save it and send me the file, I will compile it.

It is a real mission for me to change all the formula (which convert themselves to images) when the file is transfered between you and me. Incompatible word problems. The formulas get repeated.

Just have a look at the tex file (you can open it with . It is similar to an R file so actually you can open it in R and edit it there.

Let me know if it is OK.

Examples from others: Data

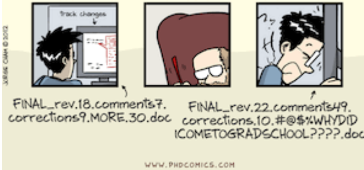


A STORY TOLD IN FILE NAMES:

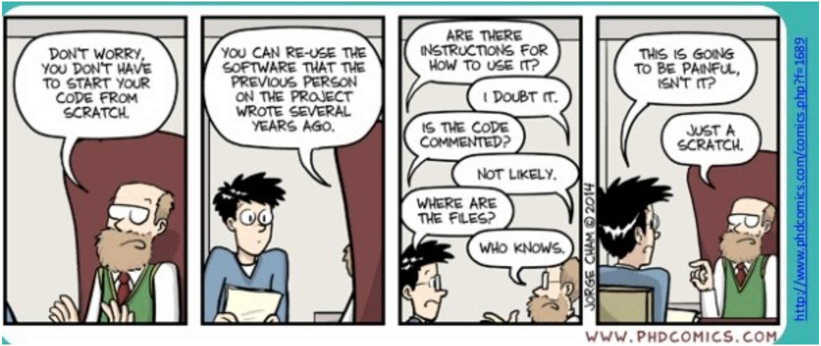
Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

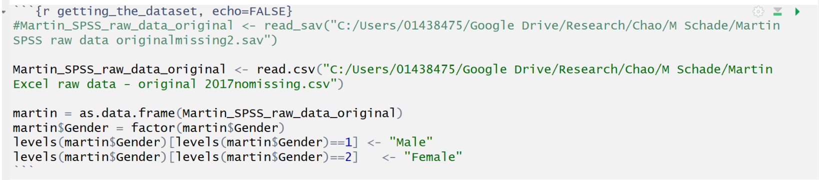


Examples from others: Code


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

http://www.phdcomics.com/comics.php?f=1689

What did I do with the missing observations?: Data

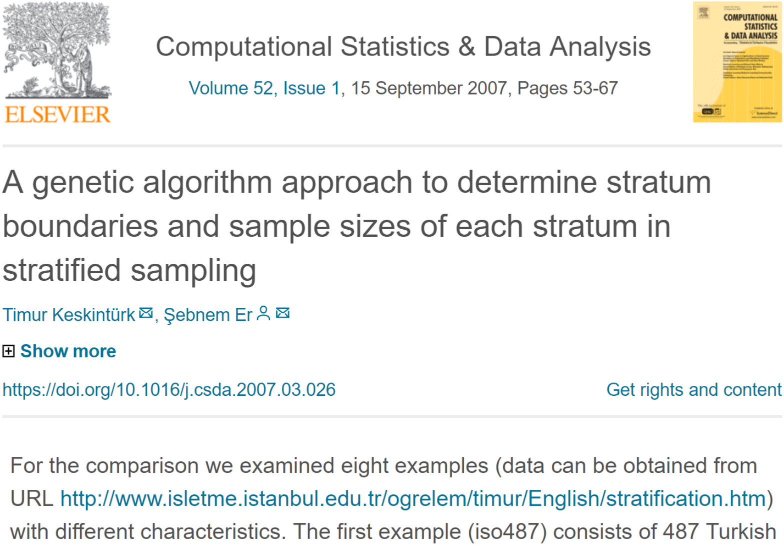


~~~
```{r getting_the_dataset, echo=FALSE}
#Martin_SPSS_raw_data_original <- read_sav("C:/Users/01438475/Google Drive/Research/Chao/M Schade/Martin
SPSS raw data originalmissing2.sav")

Martin_SPSS_raw_data_original <- read.csv("C:/Users/01438475/Google Drive/Research/Chao/M Schade/Martin
Excel raw data - original 2017nomissing.csv")

martin = as.data.frame(Martin_SPSS_raw_data_original)
martin$Gender = factor(martin$Gender)
levels(martin$Gender)[levels(martin$Gender)==1] <- "Male"
levels(martin$Gender)[levels(martin$Gender)==2]   <- "Female"
```
~~~

Examples from my own research: Storage



Computational Statistics & Data Analysis

Volume 52, Issue 1, 15 September 2007, Pages 53-67

A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling

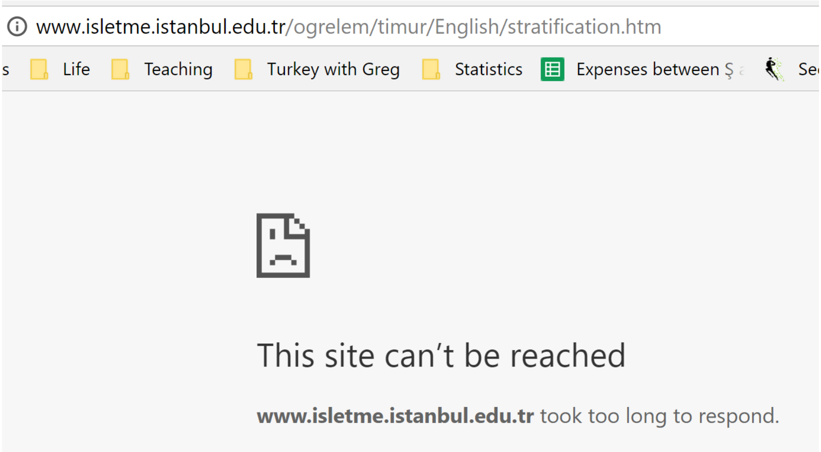
Timur Keskintürk ✉, Şebnem Er 👤 ✉

⊞ **Show more**

https://doi.org/10.1016/j.csda.2007.03.026 Get rights and content

For the comparison we examined eight examples (data can be obtained from URL http://www.isletme.istanbul.edu.tr/ogrelem/timur/English/stratification.htm) with different characteristics. The first example (iso487) consists of 487 Turkish

Examples from my own research: Storage



www.isletme.istanbul.edu.tr/ogrelem/timur/English/stratification.htm

This site can’t be reached

www.isletme.istanbul.edu.tr took too long to respond.

Why Should Research be Reproducible?

In all of these examples, you will see problems from

- Researchers/Collaborators
- Students
- Instructors
- Editors
- Private Sector Researchers
- Others who want to access your data and repeat what you have done

point of view. Everyone is affected.

What is Reproducible Research?

Science, according to the American Physical Society, “is the systematic enterprise of gathering knowledge, organizing and condensing that knowledge into testable laws and theories.”

What is Reproducible Research?

Science, according to the American Physical Society, “is the systematic enterprise of gathering knowledge, organizing and condensing that knowledge into testable laws and theories.”

How do we evaluate scientific claims?

What is Reproducible Research?

Science, according to the American Physical Society, “is the systematic enterprise of gathering knowledge, organizing and condensing that knowledge into testable laws and theories.”

How do we evaluate scientific claims?

Replication

What is Reproducible Research?

Science, according to the American Physical Society, “is the systematic enterprise of gathering knowledge, organizing and condensing that knowledge into testable laws and theories.”

How do we evaluate scientific claims?

Replication

Same results again and again –> findings are relevant

What is Reproducible Research?

Science, according to the American Physical Society, "is the systematic enterprise of gathering knowledge, organizing and condensing that knowledge into testable laws and theories."

How do we evaluate scientific claims?

Replication

Same results again and again –> findings are relevant

However, replication "requires the complete and open exchange of data, procedures, and materials". We cannot replicate many scientific research, due to time, money or uniqueness of the research.

What do we do if we cannot replicate?

Then what do we do? We can have a middle ground and deploy reproducible research instead, which is the calculation of quantitative scientific results by independent scientists using the original datasets and methods.

What do we do if we cannot replicate?

Then what do we do? We can have a middle ground and deploy reproducible research instead, which is the calculation of quantitative scientific results by independent scientists using the original datasets and methods.

- What are the advantages of Reproducible Research?

What do we do if we cannot replicate?

Then what do we do? We can have a middle ground and deploy reproducible research instead, which is the calculation of quantitative scientific results by independent scientists using the original datasets and methods.

- What are the advantages of Reproducible Research?
    - Better work habits
    - Better team work
    - Changes are easier
    - Higher research impact

What do we do if we cannot replicate?

Then what do we do? We can have a middle ground and deploy reproducible research instead, which is the calculation of quantitative scientific results by independent scientists using the original datasets and methods.

- What are the advantages of Reproducible Research?
    - Better work habits
    - Better team work
    - Changes are easier
    - Higher research impact

- What tools can we use for RR?

Tools for Reproducible Research (RR)?

- Before explaining the tools used for RR, what is the life cycle of a research?
  - Data collection
  - Data cleaning
  - Statistical analysis
  - Presentation of results/documentation/editing

For every step of your research, you need a tool for RR. Everything in one script.

Practical Tips for Reproducible Research (RR)

1. Document everything!,
2. Everything in a script file,
3. All files should be human readable: Literate Programming,
4. Explicitly tie your files together: data, codes, pdf output etc.,
5. Have a plan to organize, store, and make your files available.

Using these tips will help make your computational research really reproducible.

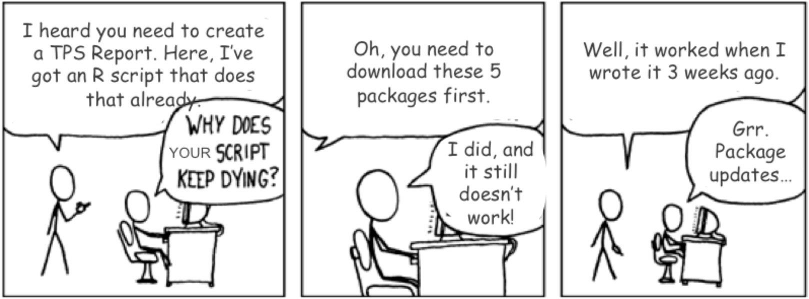
1. Document everything

Ideally, you should tell your readers how you

- gathered your data,
- analyzed it, and
- presented the results.
- a key part of documenting with R is that you should record your session info:

```r
sessionInfo()
```

Session Info


I heard you need to create a TPS Report. Here, I've got an R script that does that already.
WHY DOES YOUR SCRIPT KEEP DYING?
Oh, you need to download these 5 packages first.
I did, and it still doesn't work!
Well, it worked when I wrote it 3 weeks ago.
Grr. Package updates...

2. Everything is a script file

- .R file
- .txt file
- .Rmd file
- .m file

etc.

3. All files should be human readable

Treat all of your research files as if someone who has not worked on the project will, in the future, try to understand them.

3. All files should be human readable

Treat all of your research files as if someone who has not worked on the project will, in the future, try to understand them.

including yourself!

3. All files should be human readable

Treat all of your research files as if someone who has not worked on the project will, in the future, try to understand them.

including yourself!

With this in mind it is a good idea to comment frequently.

3. All files should be human readable

Treat all of your research files as if someone who has not worked on the project will, in the future, try to understand them.

including yourself!

With this in mind it is a good idea to comment frequently.

Commenting Guidelines

- write a comment before a block of code describing what the code does,
- comment on any line of code that is ambiguous

Tools for Reproducible Research (RR)?

- **R**: First step is to prepare your research in an environment where you can type your code and the machine can convert the code into analysis, in between you can leave comments what each code is doing.

Tools for Reproducible Research (RR)?

- **R**: First step is to prepare your research in an environment where you can type your code and the machine can convert the code into analysis, in between you can leave comments what each code is doing.
- **RStudio**: RStudio allows you to do all of the things R can do. It is a happy medium between R's text-based interface and a pure GUI and it can be linked to numerous reproducible research publishing environments such as LaTeX.

Tools for Reproducible Research (RR)?

- **R**: First step is to prepare your research in an environment where you can type your code and the machine can convert the code into analysis, in between you can leave comments what each code is doing.
- **RStudio**: RStudio allows you to do all of the things R can do. It is a happy medium between R's text-based interface and a pure GUI and it can be linked to numerous reproducible research publishing environments such as LaTeX.
- **knitr**: an R package for literate programming, i.e. it allows you to combine your statistical analysis and the presentation of the results into one document. Yihui Xie is the developer. see: https://yihui.name/knitr/

Tools for Reproducible Research (RR)?

- **R Markdown**: One of the document formats that knitr supports, and it is also the simplest one. Markdown is a both easy-to-read and easy-to-write language.

Tools for Reproducible Research (RR)?

- **R Markdown**: One of the document formats that knitr supports, and it is also the simplest one. Markdown is a both easy-to-read and easy-to-write language.
- **Cloud storage & versioning**: Services such as Dropbox and Git/Github that can store data, code, and presentation files, save previous versions of these files, and make this information widely available.

Tools for Reproducible Research (RR)?

- **R Markdown**: One of the document formats that knitr supports, and it is also the simplest one. Markdown is a both easy-to-read and easy-to-write language.
- **Cloud storage & versioning**: Services such as Dropbox and Git/Github that can store data, code, and presentation files, save previous versions of these files, and make this information widely available.
- **Unix-like shell programs**: These tools are useful for working with large research projects.

Knitr and R Markdown

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

```r
plot(pressure)
```

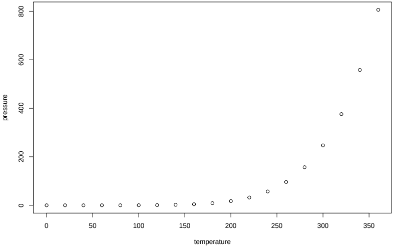


Figure 1: Scatterplot of Pressure vs Temperature

Knitr and R Markdown

```r
# Fit simple linear regression model
M1 <- lm(Examination ~ Education, data = swiss)
```

% latex table generated in R 3.4.2 by xtable 1.8-2 package % Tue Dec 05 18:08:46 2017

Table 1: Linear Regression, Dependent Variable: Exam Score

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.1	1.3	7.9	0.0
Education	0.6	0.1	6.5	0.0

This was created with **xtable** package. For multiple tables, you can use **apsrtable** package in **R**.

Cloud Storage and Versioning

We need to be able to access our files from multiple devices in different locations.

Cloud Storage and Versioning

We need to be able to access our files from multiple devices in different locations.

We often need a way for our collaborators to access and edit research files as well.

Cloud Storage and Versioning

We need to be able to access our files from multiple devices in different locations.

We often need a way for our collaborators to access and edit research files as well.

When working on a collaborative project, one of the authors may accidentally delete something in a file that another author needed.

Cloud Storage and Versioning

We need to be able to access our files from multiple devices in different locations.

We often need a way for our collaborators to access and edit research files as well.

When working on a collaborative project, one of the authors may accidentally delete something in a file that another author needed.

To deal with these issues we need to store our data in a system that has version control.

Cloud Storage and Versioning

We need to be able to access our files from multiple devices in different locations.

We often need a way for our collaborators to access and edit research files as well.

When working on a collaborative project, one of the authors may accidentally delete something in a file that another author needed.

To deal with these issues we need to store our data in a system that has version control.

Version control systems keep track of changes we make to our files and allows us to access previous versions if we want to.

Git/Github

At the heart of GitHub is an open source version control system (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.

Git/Github

At the heart of GitHub is an open source version control system (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.

To use Git on the command line, you'll need to download, install, and configure Git on your computer.

Git/Github

At the heart of GitHub is an open source version control system (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.

To use Git on the command line, you'll need to download, install, and configure Git on your computer.

If you want to work with Git locally, but don't want to use the command line, you can instead download and install the GitHub Desktop client.

Git/Github

At the heart of GitHub is an open source version control system (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.

To use Git on the command line, you'll need to download, install, and configure Git on your computer.

If you want to work with Git locally, but don't want to use the command line, you can instead download and install the GitHub Desktop client.

If you don't need to work with files locally, GitHub lets you complete many Git-related actions directly in the browser, including:

Git/Github

At the heart of GitHub is an open source version control system (VCS) called Git. Git is responsible for everything GitHub-related that happens locally on your computer.

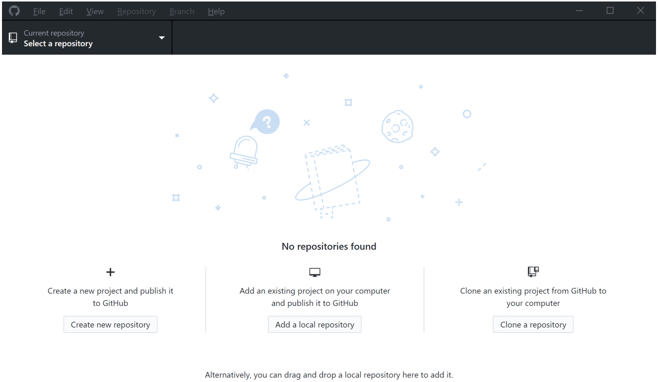
To use Git on the command line, you'll need to download, install, and configure Git on your computer.

If you want to work with Git locally, but don't want to use the command line, you can instead download and install the GitHub Desktop client.

If you don't need to work with files locally, GitHub lets you complete many Git-related actions directly in the browser, including:

- Creating a repository - Forking a repository - Managing files - Being social

Github Desktop Client



No repositories found

Create a new project and publish it to GitHub

Create new repository

Add an existing project on your computer and publish it to GitHub

Add a local repository

Clone an existing project from GitHub to your computer

Clone a repository

Alternatively, you can drag and drop a local repository here to add it.

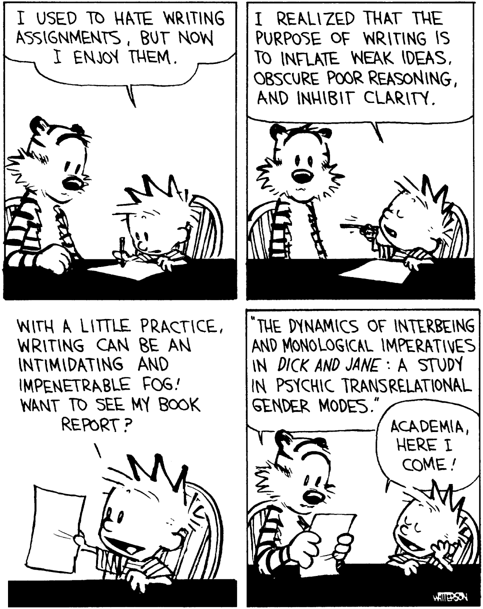
https://desktop.github.com/

Summary

What problems does reproducibility solve?

- Transparency
- Data availability
- Software/methods availability
- Improved transfer of knowledge
- ?Validity / correctness of the analysis

Calvin and Hobbes


I USED TO HATE WRITING ASSIGNMENTS, BUT NOW I ENJOY THEM.
I REALIZED THAT THE PURPOSE OF WRITING IS TO INFLATE WEAK IDEAS, OBSCURE POOR REASONING, AND INHIBIT CLARITY.
WITH A LITTLE PRACTICE, WRITING CAN BE AN INTIMIDATING AND IMPENETRABLE FOG! WANT TO SEE MY BOOK REPORT?
"THE DYNAMICS OF INTERBEING AND MONOLOGICAL IMPERATIVES IN *DICK AND JANE*: A STUDY IN PSYCHIC TRANSRELATIONAL GENDER MODES."
ACADEMIA, HERE I COME!
WATTERSON

References

Christopher Gandrud (2014). Reproducible Research with R and RStudio, CRC Press.

Victoria Stodden, Friedrich Leisch, Roger Peng (2014). Implementing Reproducible Research. CRC Press.

Yihui Xie. Dynamic Documents with R and knitr, CRC Press

You can start looking at this cheat sheet: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf.

For more details on using R Markdown see http://rmarkdown.rstudio.com.

How to setup Github using web: https://dannguyen.github.io/github-for-portfolios/lessons/setup-github/.

For more details on using R Markdown see http://rmarkdown.rstudio.com.

Thank you. . .

https://sebnemer.github.io/