

# Flujo de trabajo reproducible usando R

Nicolas Ballarini

Section for Medical Statistics, CeMSIIS  
Medical University of Vienna

[nicolas.ballarini@meduniwien.ac.at](mailto:nicolas.ballarini@meduniwien.ac.at)

December 5th, 2018

# Introducción

- PhD en Bioestadística @MUW
- Investigo en métodos estadísticos para el análisis de subgrupos en ensayos clínicos
- Colaboración con estadísticos de otras universidades y empresas farmacéuticas
- Colaboración con médicos del hospital general de Viena.  
Análisis de estudios clínicos.

# Breve encuesta

Kahoot!

# 7-8 años atrás...



Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística  
Escuela de Estadística

## TESINA

Propiedades de gráficos de control estadístico de procesos multivariado utilizando Cadenas de Markov

Nicolás Marcelo Ballarini

Licenciatura en Estadística

Directora: Dra. Marta B. Quagliino

Rosario - 2011

ARL3 4 5 (Copia conflictiva de Nicolas Ballarini 2011-11-28).sas	✓
presentacion tesina-1.pptx	✓
presentación tesina.ppsx	✓
presentación tesina.pptx	✓
Ballarini_Tesina.pdf	✓
TESINA.pdf	✓
3102009a.pdf	✓
Aplicacion.pdf	✓
ARL.sas	✓
ARL3 4 5.sas	✓
raices.sas	✓
Recorte.shs	✓
ARLs.rtf	✓
Compilación-final	►
Compilación-3	►
Compilación-4	►
Compilación-5	►
Figuras	✓
previos	✓

# Reproducibilidad

Cualidad de un experimento de poder ser repetido por otros

Estudio sobre convulsiones retractado luego que los autores admiten que los datos se “mezclaron terriblemente”

### Low Dose Lidocaine for Refractory Seizures in Preterm Neonates:

“The article has been retracted at the request of the authors. After carefully re-examining the data presented in the article, they identified that data of two different hospitals got terribly mixed. The published results cannot be reproduced in accordance with scientific and clinical correctness.”

Source: <http://retractionwatch.com/2013/02/01/seizure-study-retracted-after-authors-realize-data-got-terribly-mixed/>

# Para pensar

Alguna vez trataste de reproducir un análisis de datos de un colega/compañero?

Alguna vez trataste de reproducir tu propio trabajo?

Qué fue lo mas difícil? Sería fácil extender el análisis?

# Casos de colaboracion

3 meses despues de haber entregado el “analysis\_final.pdf”:

*“Hacemos un analisis de sensibilidad imputando valores faltantes?”*

*“Habia un error en el paciente ID3104, se puede hacer de nuevo el analisis?”*

*“Por favor cambia los colores del gráfico para que sea legible en blanco y negro”*

*“Para la publicacion necesito los gráficos en formato tiff en lugar de png”*

# Data scientist

*"A surprising amount of doing data science really well is just being good at managing and organizing all of these files so that they are:*

- *Easy to find*
- *Easy to share*
- *Easy to understand*
- *Easy to update"*

(Chromebook data science: Material for Organizing Data Science Projects)

# 6 meses atrás...



## Subgroup identification in clinical trials via the predicted individual treatment effect

Nicolás M. Ballarini<sup>1</sup>, Gerd K. Rosenkranz<sup>1</sup>, Thomas Jakl<sup>2</sup>, Franz König<sup>1</sup>, Martin Posch<sup>1\*</sup>

**1** Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria,

**2** Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.

\* martin.posch@meduniwien.ac.at

### Abstract

Identifying subgroups of treatment responders through the different phases of clinical trials has the potential to increase success in drug development. Recent developments in

**Revision: "I want to try the methods, can the author provide the code?"**

# 6 meses atrás...

The screenshot shows a GitHub repository page. At the top, there's a navigation bar with links for "Why GitHub?", "Business", "Explore", "Marketplace", "Pricing", "Search", "Sign in", and "Sign up". Below the navigation is the repository header for "nicoballarini / PLOS-2018-PITE". The repository title is "Code repository for: Subgroup Identification Clinical Trials via the Predicted Individual Treatment Effect". It shows 2 commits, 1 branch, 0 releases, and 1 contributor. A "Find file" and "Clone or download" button are present. The main area displays a list of files and their upload history:

File	Action	Date
nicoballarini/uploaded code	upload code	4 months ago
R	upload code	4 months ago
man	upload code	4 months ago
paper	upload code	4 months ago
aim	upload code	4 months ago
Rbuildignore	upload code	4 months ago
Rhistory	upload code	4 months ago
DESCRIPTION	upload code	4 months ago
LICENSE	upload code	4 months ago
Makefile	upload code	4 months ago
NAMESPACE	upload code	4 months ago
PITE.Rproj	upload code	4 months ago
README.md	first commit	4 months ago

At the bottom, it says "This project contains the article and code for:".

The screenshot shows a PLOS ONE research article page. The header features the PLOS ONE logo. The article title is "Subgroup identification in clinical trials via the predicted individual treatment effect". It is categorized as a "RESEARCH ARTICLE". The authors listed are Nicolas M. Ballarini<sup>1</sup>, Gerd K. Rosenkranz<sup>2</sup>, Thomas Jaki<sup>1</sup>, Franz König<sup>3</sup>, Martin Posch<sup>1\*</sup>. The institutions involved are the Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria; 2 Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom; and the Institute of Biostatistics and Medical Information Science, Medical University of Vienna, Vienna, Austria. The DOI is 10.1371/journal.pone.0205971, and the URL is <https://doi.org/10.1371/journal.pone.0205971>.

**Abstract**

Identifying subgroups of treatment responders through the different phases of clinical trials has the potential to increase success in drug development. Recent developments in subgroup analysis consider subgroups that are defined in terms of the predicted individual treatment effect, i.e. the difference between the predicted outcome under treatment and the predicted outcome under control for each individual, which in turn may depend on multiple biomarkers. In this work, we study the properties of different modeling strategies to estimate the predicted individual treatment effect. We explore linear models and compare different estimation methods, such as maximum likelihood and the Lasso with and without randomized response. For the latter, we implement confidence intervals based on the selective inference framework to account for the model selection stage. We illustrate the methods in a dataset of a treatment for Alzheimer disease (normal response) and in a dataset of a treatment for prostate cancer (survived outcome). We also evaluate via simulations the performance of using the predicted individual treatment effect to identify subgroups where a novel treatment leads to better outcomes compared to a control treatment.



# Reproducibilidad en revistas científicas

The screenshot shows the 'Information for Authors' section of the Biostatistics journal website. At the top, there is a navigation bar with links for 'Issues', 'Advance articles', 'Submit', 'Purchase', 'Alerts', and 'About'. Below the navigation bar, the title 'Biostatistics' is displayed. The main content area is titled 'Information for Authors' and contains two sections: 'Reproducible Research' and 'Code Availability'. The 'Reproducible Research' section states that the journal's policy is for papers to be kite-marked D if the data on which they are based are freely available, C if the authors' code is freely available, and R if both data and code are available, and that the Associate Editor for Reproducibility is able to use these to reproduce the results in the paper. Data and code are published electronically on the journal's website as Supplementary Materials. The 'Code Availability' section encourages authors to submit code supporting their publications, suggesting GitHub or Figshare/Zenodo as submission options.

The screenshot shows the 'Guidelines for Code and Data Submission' document for the Biometrical Journal. At the top, the journal logo is shown with the text 'Biometrical Journal'. Below the logo, the title 'Guidelines for Code and Data Submission' and 'Specific Guidance on Reproducible Research (RR)' are displayed. The document is signed by Benjamin Hofner and Fabian Scheipl (RR Editors, Biometrical Journal) with the email address fabian.scheipl@stat.uni-muenchen.de. The document version is 1.7 (2016/10/28). The table of contents includes:

<b>1</b>	<b>Submission material</b>	<b>2</b>
1.1	Code and data . . . . .	2
1.2	README . . . . .	4
1.3	Structure of code submission (ZIP-folder) . . . . .	5
<b>2</b>	<b>Submission process</b>	<b>6</b>
<b>3</b>	<b>Reference the code</b>	<b>6</b>
<b>4</b>	<b>Example</b>	<b>7</b>
<b>5</b>	<b>Possible improvements</b>	<b>7</b>
<b>A</b>	<b>Using relative paths in R</b>	<b>8</b>

[https://academic.oup.com/biostatistics/pages/General\\_Instructions](https://academic.oup.com/biostatistics/pages/General_Instructions)

<https://onlinelibrary.wiley.com/page/journal/15214036/homepage/forauthors.html>

# Reproducibility checklist: Documentación (1/2)

- Hay un archivo “README” que indica el propósito del proyecto, a quien contactar, un mapa de la estructura del directorio, y qué software/hardware se necesita para reproducir tu flujo de trabajo?
- Hay un archivo “README” en cada sub-carpetas que describe su contenido?
- Hay un archivo “CITATION” que indica como citar el proyecto, los datos y el código?
- Hay instrucciones de cómo obtener el conjunto de datos original?
- Hay una lista de las dependencias del proyecto con el número de versión que se necesita para cada una? (e.g. R packages)
- Están indicadas las fechas en las que fueron accedidos los sitios

## Reproducibility checklist: Documentación (2/2)

- Hay un archivo “LICENSE” que especifica la licencia con la cual se distribuye el proyecto? Fue editada para incluir la información específica de tu proyecto? Fueron tenidas en cuenta las licencias de otro contenido incluido en el proyecto?
- Para análisis que requieren estudios de simulación o generación de números aleatorios, se indica el ‘seed’ que se utilizó?
- Está el código bien documentado?
- Usas comentarios a lo largo del script?
- Se incluye en cada script un encabezado detallando inputs, outputs y dependencias?
- Está documentado como se relacionan los archivos entre si?

# Reproducibility checklist: Organización (1/2)

- Estan todos los conjuntos de datos, el codigo y la documentacion en un mismo proyecto bajo una estructura adecuada?
- Se uso 'version control'?
- Hay un repositorio publicamente disponible para el proyecto? Se puede asegurar que este repositorio estara disponible a largo plazo?
- Es posible ejecutar el codigo desde la carpeta 'home' y se producen los resultados necesarios?

## Reproducibility checklist: Organización (2/2)

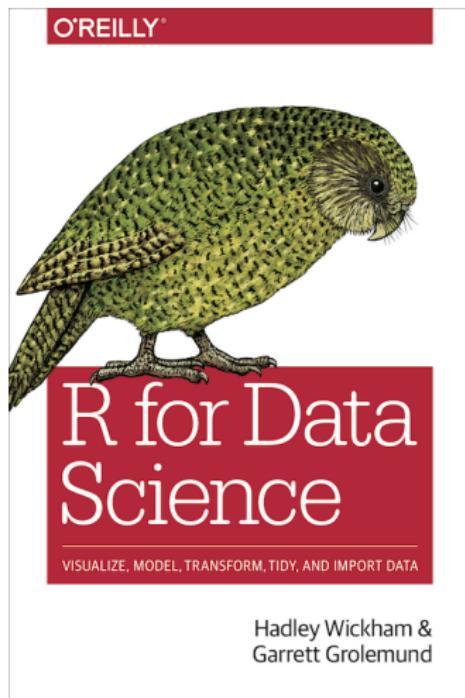
- Están separados los conjuntos de datos originales y los manipulados para el análisis?
- Se mantuvieron los conjuntos de datos originales? (sin ninguna manipulación)
- Usaste un sistema consistente en los nombres de los archivos?
- Hay un mecanismo para guardar grandes conjuntos de datos?

## Reproducibility checklist: Otros:

- Que tan sensibles son los resultados en cuanto al sistema operativo que se utilice?
- Se incluyen tests para confirmar que el codigo realmente hace lo que intenta hacer?
- Los reportes son generados automaticamente a partir de los resultados obtenidos?

# Reproducibilidad en práctica

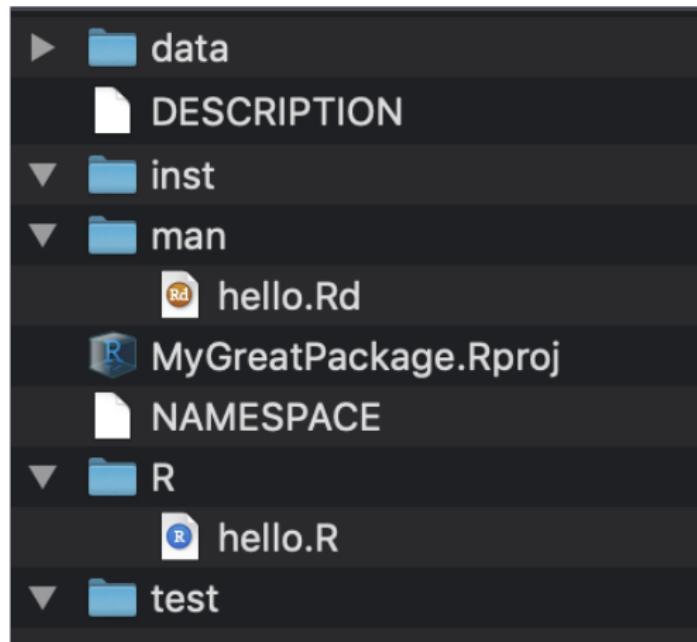
## Funciones en R



*"You should consider writing a function whenever you've copied and pasted a block of code more than twice (i.e. you now have three copies of the same code)."*

# Reproducibilidad en práctica

## R packages: Estructura básica de un paquete de R



# Reproducibilidad en práctica

Estructura recomendada para un proyecto de investigación:

## Reproducible Research Project Initialization

Research project initialization and organization following reproducible research guidelines.

### Overview

```
project
|- doc/           # documentation for the study
|  +- paper/      # manuscript(s), whether generated or not
|
|- data          # raw and primary data, are not changed once created
|  |- raw/         # raw data, will not be altered
|  +- clean/       # cleaned data, will not be altered once created
|
|- code/          # any programmatic code
|- results        # all output from workflows and analyses
|  |- figures/    # graphs, likely designated for manuscript figures
|  +- pictures/   # diagrams, images, and other non-graph graphics
|
|- scratch/       # temporary files that can be safely deleted or lost
|- README          # the top level description of content
|- study.Rmd      # executable Rmarkdown for this study, if applicable
|- Makefile        # executable Makefile for this study, if applicable
|- study.Rproj     # RStudio project for this study, if applicable
|- datapackage.json # metadata for the (input and output) data files
```

<https://github.com/Reproducible-Science-Curriculum/rr-init>

# Reproducibilidad en práctica

Proyecto de investigación: proponer un método para calcular el tamaño muestral para un estudio clínico.

Hipótesis de interés  $H_0: \mu_1 = \mu_2$ .

Luego de meses de investigación y numerosas discusiones con colegas, se llega a la conclusión de que el tamaño de muestra debe ser calculado con la siguiente forma:

$$n = \frac{2\sigma^2(z_{\alpha/2} + z_{\beta})^2}{\Delta^2}$$

# Reproducibilidad en práctica

Paso 1: Creá tu propio paquete de R!

file > new project > new directory > R package

Paso 2: Adaptando el paquete de R para proveer el método propuesto: Editar el archivo DESCRIPTION.

# Reproducibilidad en práctica

Paso 3: Crear un archivo en la carpeta R/ que llamado sample\_size.R. En primer lugar, calcular el tamaño de muestra si  $\alpha = 0.05$ ,  $\beta = 0.2$ ,  $\sigma^2 = 1$  y  $\Delta = 0.1$ . Luego, escribir una función get\_sample\_size que tome los cuatro parámetros como argumentos y calcule el tamaño muestral.

# Reproducibilidad en práctica

Paso 4: Documentar la función usando como plantilla el archivo en la carpeta man.

Nota: Roxygen permite hacer esto mucho más fácil.

(<https://support.rstudio.com/hc/en-us/articles/200532317-Writing-Package-Documentation>)

# Reproducibilidad en práctica

## Paso 6: 'Build'

Paso 7: Crear un archivo “001\_sample\_size.r” en la carpeta inst/code en donde se use la función con valores de  $\Delta$  desde 0.05 a 0.50 (step = 0.01). Crear un data.frame con los resultados (dos columnas: delta y N) y guardarlo en un archivo 001\_sample\_size.rda en la misma carpeta.

Paso 8: Crear un archivo “002\_sample\_size\_plot.r” en la carpeta inst/code en donde se use el data.frame creado en el paso anterior para hacer un gráfico de delta vs. N. Guardar el resultado en un archivo “002\_sample\_size\_plot.pdf”

# Reproducibilidad en práctica

Paso 9: Crear un archivo “01\_reporte.Rmd” en la carpeta `inst/reports/` en el que se carguen el `data.frame` y se muestre como una tabla, junto con el grafico.

Paso 10: Crear un archivo “`make.r`” en la carpeta principal del proyecto. Este script ejecuta cada uno de los scripts en la capera `inst/code/` y compila el reporte.

# Reproducibilidad en práctica

Proyectos de análisis de datos tal vez no requieran de construir un paquete, ni funciones 'custom-made'. Sin embargo, seguir una estructura similar puede ser muy útil

# Recursos adicionales/Tips

(Parte del material de esta charla proviene de las siguientes fuentes)

- data carpentry reproducible research course:  
<https://datacarpentry.org/rr-workshop/> -  
<https://github.com/Reproducible-Science-Curriculum>
- Hadley Wickham's Creating R packages: <http://r-pkgs.had.co.nz/>
- Karl Browman's tutorials: minimal make.  
[http://kbroman.org/minimal\\_make/](http://kbroman.org/minimal_make/) - steps for RR  
<https://kbroman.org/steps2rr/>
- Chromebook data science course:  
<https://jhudatascience.org/chromebookdatascience/>
- Check styling guides (e.g. <http://adv-r.had.co.nz/Style.html>)
- github student account (repositorios privados)  
<https://education.github.com/pack>

# Gracias!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567.



contact: [nicolas.ballarini@meduniwien.ac.at](mailto:nicolas.ballarini@meduniwien.ac.at)