

Análisis de datos reproducible con R:

Herramientas y ejemplos para ciencias de la salud

Felipe Ortega

Data Science Lab.
Universidad Rey Juan Carlos. Madrid.

9 de junio de 2021



- 1 Introducción: Reproducibilidad
- 2 La base: R + RStudio
- 3 Autoría de documentos: R Markdown
- 4 Herramientas: `validate`, `validatetools`
- 5 Conclusiones

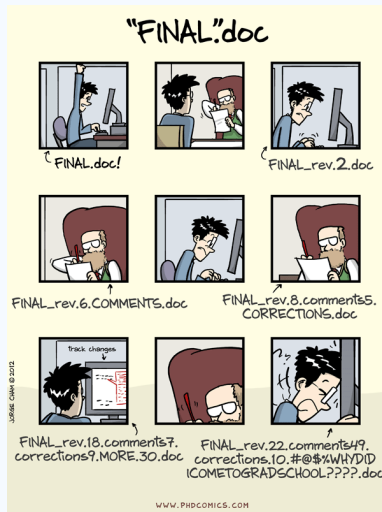
Sección 1

Introducción: Reproducibilidad

- Durante muchos años, el método científico se ha basado en la publicación de investigaciones, describiendo el resultado de análisis de datos.
- Ejemplos:
 - Eficacia de un nuevo medicamento para el tratamiento de pacientes...
 - Comparación de la capacidad de aprendizaje de alumnos en diferentes colegios de...
 - Modelado de la evolución epidémica de una enfermedad a lo largo del tiempo en una zona geográfica...
 - Beneficios de un nuevo método quirúrgico para intervenciones de pacientes, respecto a los métodos previamente aplicados...
- En todos estos casos, es importante confiar en las condiciones, los datos, el método y las herramientas que los autores de la publicación han empleado para el análisis.

- Sin embargo, los avances en las herramientas y métodos de análisis de datos hacen que ahora sea más sencillo comprobar el resultado de estos análisis...
- ... y aquí empiezan los problemas:
 - **Oncología** [3]: Dpto. Biotecnología de la firma Amgen (Thousand Oaks) sólo confirmó 6 de un total de 53 artículos emblemáticos. Bayer HealthCare (Alemania) pudo validar un 25 % de estudios.
 - **Psicología** [4]: De un total de 249 artículos de la APA, el 73 % de los autores no respondieron sobre sus datos en 6 meses.
 - **Economía y finanzas** [5]: Diferentes paquetes software producen resultados muy distintos con técnicas estadísticas directas aplicadas sobre datos idénticos a los originales.
- De hecho, han llegado a aparecer artículos que sugieren que buena parte de los resultados publicados en áreas como Medicina pueden no ser fiables (Ioannidis, 2005) [6]. Esto ha generado gran polémica y una crisis de confianza ¹.

¹<https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/>



<http://phdcomics.com/comics/archive.php?comiciid=1531>

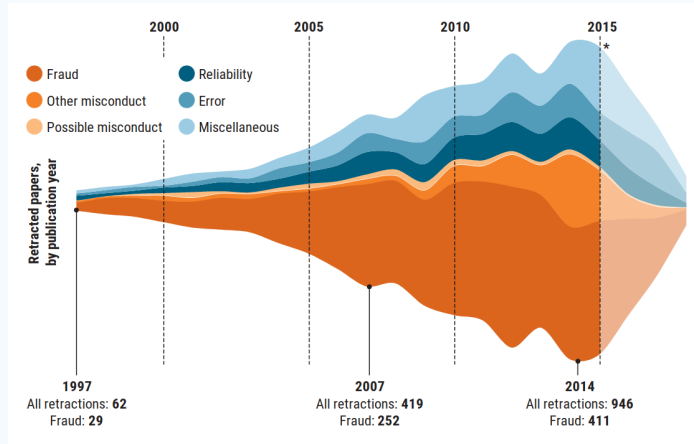


Figura: Evolución de fraude detectado en investigación, 1997-2014 ² [8].

²<https://www.sciencemag.org/news/2018/10/about-these-data>

Se habla con frecuencia de *reproducir* y de *replicar* un análisis [2]:

- Podemos definir la **reproducibilidad** como la capacidad para recomputar los resultados de un análisis, con los mismos datos que se emplearon en el análisis original, y conociendo los detalles de la secuencia (*pipeline*) de análisis.
 - Si uso las mismas herramientas (e.g. R, un listado de paquetes, mismas versiones) y el mismo código (sentencias en R) sobre los mismos datos, los resultados y conclusiones han de ser consistentes con el análisis original.
 - Los autores originales deben proporcionar todos los elementos (datos, código y procedimiento empleado) para permitir que el análisis sea reproducible [7].
- Se define la **replicabilidad** de un estudio como la capacidad de realizar un experimento independiente, que aborde el mismo objetivo pero con un conjunto de datos diferente del original. Si los resultados no son consistentes, es necesario realizar más réplicas y armonizar las conclusiones, por ejemplo mediante **meta-análisis**.

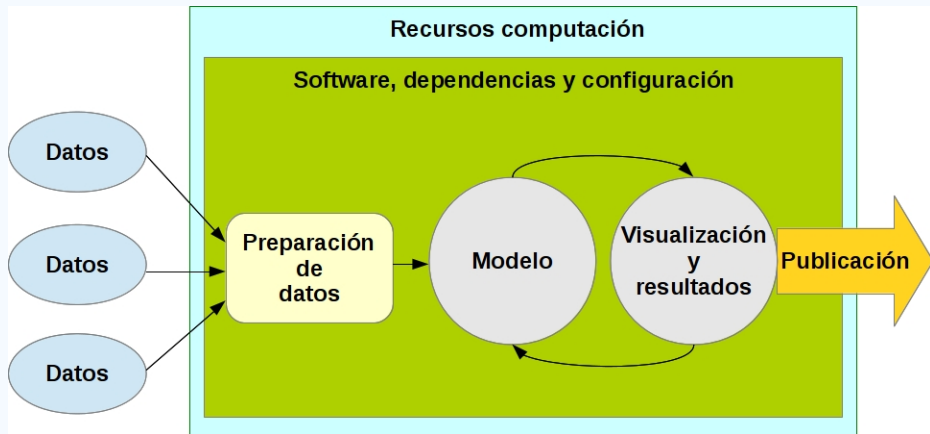


Figura: Flujo de trabajo en un análisis típico de datos.

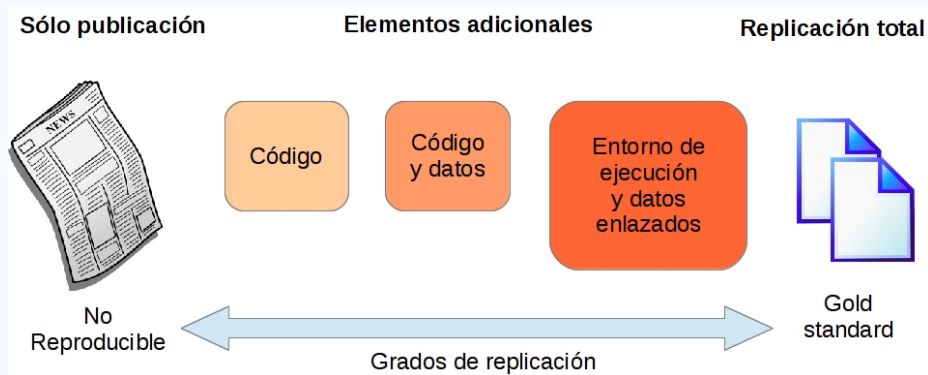


Figura: Niveles de replicabilidad de un análisis de datos, según (Peng, 2011) [1].

- **Conjuntos de datos** que se han utilizado.
- **Infraestructura** equivalente (recursos computacionales).
- **Software:**
 - **Código** para llevar a cabo el análisis.
 - **Dependencias** satisfechas (otros programas, bibliotecas, S.O., etc.).
 - **Configuración** original para el análisis.
- **Metodología.**
 - Explicación detallada del **proceso** (limpieza y preparación de datos, análisis, resultados, conclusiones).

- En este seminario, vamos a mostrar algunos elementos básicos para facilitar que nuestros análisis sean replicables, utilizando R.
- Plan de trabajo:
 1. Familiarizarnos con R y RStudio como entorno de trabajo.
 2. Introducción a R Markdown, paquete que permite integrar fácilmente descripción y contenidos (código, figuras, resultados numéricos) en un solo producto (HTML, documento PDF o MS Word).
 3. Otras herramientas interesantes para análisis de datos reproducibles, como paquetes que automatizan protocolos de validación de datos (`validate` y `validatetools`).
- Los materiales facilitados incluyen muchos enlaces y referencias a documentación adicional, para ampliar conocimientos y seguir practicando.

Sección 2

La base: R + RStudio

- Repaso rápido de la interfaz de RStudio, el entorno de trabajo con R más popular.
- Referencias introducción a R y RStudio:
 - E. López Cano y J. Martínez Moguerza: *R desde el principio: curso ceRo de R*. Ediciones del Orto, Madrid (2015). Última versión: 1 de febrero de 2018. [[Enlace](#)].
- Referencias adicionales sobre RStudio:
 - Webinars “RStudio Essentials”: <https://resources.rstudio.com/>.
 - RStudio IDE *cheat sheet*:
<https://resources.rstudio.com/rstudio-cheatsheets/rstudio-ide-cheat-sheet>.
 - Data import *cheat sheet*:
<https://resources.rstudio.com/rstudio-cheatsheets/data-import-cheat-sheet>.

Sección 3

Autoría de documentos: R Markdown

- Introducimos las características y funcionalidades principales de R Markdown.
- Objetivos:
 - Entender los principales casos de aplicación de R Markdown.
 - Flujo de trabajo básico para autoría de documentos usando R Markdown.
- R Markdown facilita la inclusión de la explicación del análisis, las dependencias necesarias (paquetes R), el código para llevarlo a cabo y los resultados (numéricos y gráficos) dentro de un mismo documento.
- Es sencillo generar a partir de un documento varios formatos de salida: HTML, PDF (LaTeX) o MS Word.
 - Nos centraremos en HTML, también LaTeX (si la configuración lo permite).

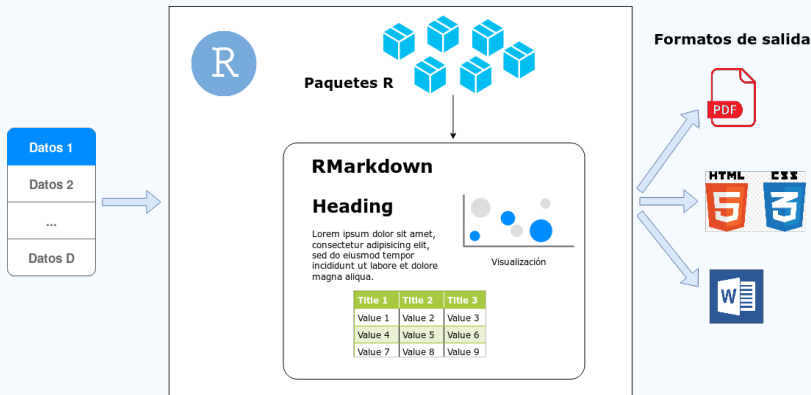


Figura: Esquema de trabajo con R Markdown. Un documento `.Rmd` centraliza la explicación, dependencias software, código de preparación y análisis de datos. Se integran resultados numéricos y de visualización. Podemos elegir tres formatos de salida: HTML, PDF o MS Word.

- Con R Markdown también podemos generar otros tipos de documentos:
 - **Presentaciones con diapositivas**, en formato HTML o PDF (Beamer). Existen diferentes estilos de presentación, con plantillas.
 - **Artículos científicos**, con plantillas para las principales publicaciones de organizaciones, revistas científicas, congresos, etc.
 - **Blogs**, mediante el paquete adicional blogdown.
 - **Sitios web** completos. Existen paquetes para automatizar la creación y estilo profesional del acabado, como [distill](#). Se basa en servicios y herramientas externas como GitHub, Hugo y Netlify.
 - **Libros completos**, maquetados para la web (HTML, barra lateral de navegación, cuadro de búsqueda, etc.), o bien en formato PDF (LaTeX). Se necesita el paquete adicional bookdown.

1. En el menú principal de RStudio: `File` → `New file` → `R Markdown...`
2. Rellenamos los datos: Título, autor. Seleccionamos formato de salida. Pulsamos `OK`.
3. Para compilar el documento resultante, pulsamos el botón `Knitr`.

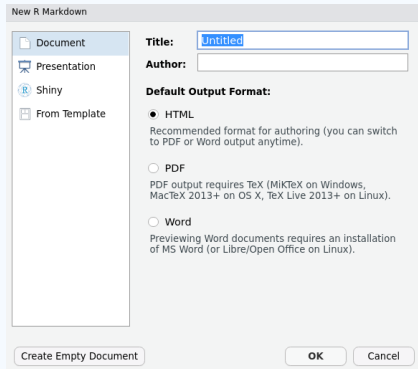


Figura: Ventana de diálogo para crear un nuevo documento R Markdown.

Incluidos en la carpeta `examples` de material del seminario.

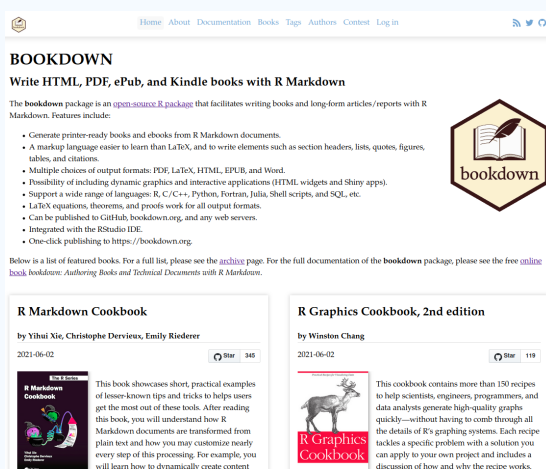
- `0-Mi-primer-documento.Rmd`
 - Un documento muy básico pero que incluye un tema personalizado y selección de resaltado de sintaxis.
- `1-Elementos-y-opciones-RMarkdown.RMD`
 - Incluye varias opciones más para personalizar el estilo del documento.
- `2-Limpieza-datos.Rmd`
 - Un documento real para una clase, que muestra todo tipo de contenido y presentación avanzada de datos, resultados, gráficos, etc.

■ Referencias adicionales sobre R Markdown:

- Getting started: <https://rmarkdown.rstudio.com/lesson-1.html>.
- H. Wickham, G. Grolemund. *R for Data Science*, 2nd Ed. O'Reilly Media, 2019. Chap. 27. [Enlace].
- Y. Xie, J.J. Allaire, G. Grolemund. *R Markdown: The Definitive Guide*. CRC Press, 2018. [Enlace].
- Y. Xie, A. Presmanes Hill, A. Thomas. *blogdown: Creating Websites with R Markdown*. CRC Press, 2017. [Enlace].
- Y. Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*. CRC Press, 2016. [Enlace].

■ Hojas y guías de referencia (RStudio):

- R Markdown cheat sheet (v. 2):
<https://resources.rstudio.com/rstudio-cheatsheets/rmarkdown-2-0-cheat-sheet>.
- R Markdown reference guide (2014):
<https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>.



The screenshot shows the homepage of the bookdown website. At the top, there is a navigation bar with links: Home, About, Documentation, Books, Tags, Authors, Contest, and Log in. The main heading is "BOOKDOWN" followed by the subtitle "Write HTML, PDF, ePub, and Kindle books with R Markdown". Below this, a paragraph states: "The bookdown package is an [open-source R package](#) that facilitates writing books and long-form articles/reports with R Markdown. Features include:" followed by a bulleted list of features. To the right of the text is a logo for bookdown, which is a hexagon containing a quill pen and the word "bookdown". Below the features list, a paragraph says: "Below is a list of featured books. For a full list, please see the [archive](#) page. For the full documentation of the bookdown package, please see the free [online book](#), bookdown: Authoring Books and Technical Documents with R Markdown." Below this, there are two book cards. The first card is for "R Markdown Cookbook" by Yihui Xie, Christophe Dervieux, and Emily Riederer, dated 2021-06-02, with 345 stars. The second card is for "R Graphics Cookbook, 2nd edition" by Winston Chang, dated 2021-06-02, with 119 stars. Each card includes a small image of the book cover and a brief description.

BOOKDOWN
Write HTML, PDF, ePub, and Kindle books with R Markdown

The **bookdown** package is an [open-source R package](#) that facilitates writing books and long-form articles/reports with R Markdown. Features include:

- Generate printer-ready books and ebooks from R Markdown documents.
- A markup language easier to learn than LaTeX, and to write elements such as section headers, lists, quotes, figures, tables, and citations.
- Multiple choices of output formats: PDF, LaTeX, HTML, EPUB, and Word.
- Possibility of including dynamic graphics and interactive applications (HTML widgets and Shiny apps).
- Support a wide range of languages: R, C/C++, Python, Fortran, Julia, Shell scripts, and SQL, etc.
- LaTeX equations, theorems, and proofs work for all output formats.
- Can be published to GitHub, bookdown.org, and any web servers.
- Integrated with the RStudio IDE.
- One-click publishing to <https://bookdown.org>.

Below is a list of featured books. For a full list, please see the [archive](#) page. For the full documentation of the **bookdown** package, please see the free [online book](#), bookdown: Authoring Books and Technical Documents with R Markdown.

R Markdown Cookbook
by Yihui Xie, Christophe Dervieux, Emily Riederer
2021-06-02 Star 345

This book showcases short, practical examples of lesser-known tips and tricks to help users get the most out of these tools. After reading this book, you will understand how R Markdown documents are transformed from plain text and how you may customize nearly every step of this processing. For example, you will learn how to dynamically create content

R Graphics Cookbook, 2nd edition
by Winston Chang
2021-06-02 Star 119

This cookbook contains more than 150 recipes to help scientists, engineers, programmers, and data analysts generate high-quality graphs quickly—without having to comb through all the details of R's graphing systems. Each recipe tackles a specific problem with a solution you can apply to your own project and includes a discussion of how and why the recipe works.

Figura: La web <https://bookdown.org/> reúne libros publicados en abierto creados mediante bookdown. El código de la mayoría se encuentra disponible en GitHub u otros sitios.

Sección 4

Herramientas: `validate`, `validatetools`

- Otro aspecto importante para automatizar la reproducción y replicación de un estudio es validar los datos empleados.
- Muchas de las variables analizadas tienen que cumplir una serie de reglas, para asegurar que sus valores son conceptualmente correctos:
- Ejemplos:
 - La variable tiene que ser estrictamente positiva (por ejemplo peso, altura...).
 - La variable tiene que estar dentro de un rango de valores (por ejemplo, una probabilidad debe estar en el intervalo $[0, 1]$).
 - Comprobar si una variable tiene valores faltantes.
 - Etc.

■ Ejemplo:

```
library(validate)
cf <- check_that(women, height > 0, weight > 0, height/weight > 0.5)
summary(cf)
```

##	name	items	passes	fails	nNA	error	warning	expression
## 1	V1	15	15	0	0	FALSE	FALSE	height > 0
## 2	V2	15	15	0	0	FALSE	FALSE	weight > 0
## 3	V3	15	2	13	0	FALSE	FALSE	height/weight > 0.5

- `validate` y `validatetools`.
- Referencias adicionales:
 - Introducción al paquete `validate`: <https://cran.r-project.org/web/packages/validate/vignettes/introduction.html>.
 - Centralizar reglas de validación en archivos de texto: https://cran.r-project.org/web/packages/validate/vignettes/rule_files.html.
 - Indicadores de calidad de los datos: <https://cran.r-project.org/web/packages/validate/vignettes/indicators.html>.

Sección 5

Conclusiones

- La **reproducibilidad** ayuda, pero **no es el fin de nuestros problemas**.
 - Facilita que **se detecten problemas**. Pero si se detectan, la conclusión es que el análisis es erróneo y sus resultados y conclusiones no son fiables. Hay que retirarlo y volver a empezar.
 - Importancia de los **pre-prints**, y de que revisen más personas que los editores/supervisores y revisores habituales (en una revista, congreso o en una organización).
- Necesidad acuciante de **mejorar la formación**, a todos los niveles, de investigadores y profesionales involucrados en el análisis de datos.
 - Elegir el **método** adecuado para realizar un análisis.
 - Ser consciente de las **limitaciones** de las herramientas y métodos empleados.
 - Garantizar que se cumplen las **condiciones previas** requeridas para aplicar un método o un modelo.
 - **Diagnosticar** los resultados de un **modelo** para evaluar sus resultados y las conclusiones derivadas del mismo.

"80 % of my time was spent cleaning the data."

***Better data** will always beat better models".*



Thomson Nguyen

<http://supportvectorhumans.com/>.

1. Peng, R.D. 2011. Reproducible research in computational science. *Science (New York)* (6060):1226–1227. [Link](#).
2. Leek J. T., Peng R. D. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences* 112, 1645–1646 (2015). [[PMC link](#)].
3. Begley, C. Glenn, and Lee M. Ellis. "Drug development: Raise standards for preclinical cancer research." *Nature* 483.7391 (2012): 531-533.
4. Wicherts, Jelte M., et al. "The poor availability of psychological research data for reanalysis.." *American Psychologist* 61.7 (2006): 726.
5. Burman, Leonard E., W. Robert Reed, and James Alm. ".^A call for replication studies." *Public Finance Review* 38.6 (2010): 787-793.

6. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. [[PMC Free article](#)].
7. Barba, L. A. (2018). Terminologies for reproducible research. arXiv preprint arXiv:1802.03311. [[Abstract & PDF](#)]
8. Brainard, J. (2018). What a massive database of retracted papers reveals about science publishing's 'death penalty'. Science. [[Link](#)].