



Algunas reflexiones/herramientas sobre reproducibilidad y replicabilidad para estadísticos

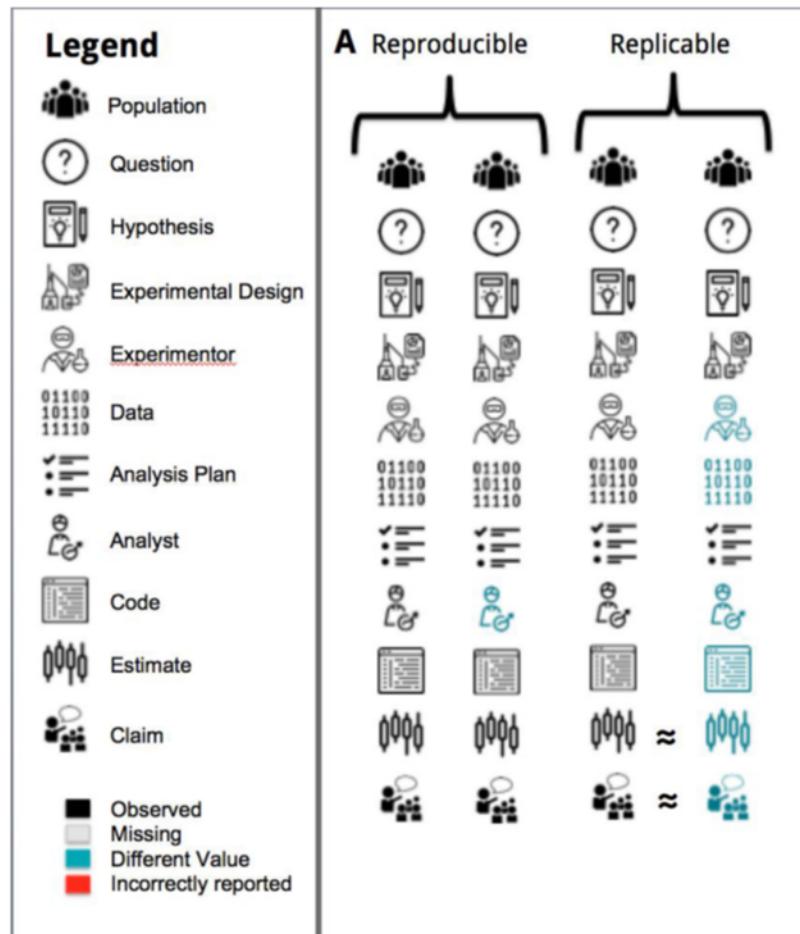
Miguel A. Martínez-Beneito - Grupo de Investigación Bayensians. Fundación FISABIO (Valencia).
25 de noviembre de 2019

Estructura de la sesión:

- **P-hacking** una explicación (estadística) más al problema de la reproducibilidad.
- **Algunas herramientas** estadísticas (del entorno R) para acometer el problema de la reproducibilidad.

P-hacking una explicación (estadística) más al problema de la reproducibilidad.

P-hacking: una definición



<http://...> (<http://tinyurl.com/cea6krt>)

“Given a population, hypothesis, experimental design, experimenter, data, analysis plan and analyst the **code changes to match** a desired experiment”
(<http://dx.doi.org/10.1101/066803>)

“If the data can **speak** for themselves they can also **lie** for themselves”
(<https://twitter.com/ImperialSpark/826860755>)

“If you **torture** the data long enough, it will **confess**”
(https://en.wikiquote.org/wiki/Ronald._Coase)

También conocido como **data dredging** (dragado de datos), **data fishing** o **fishing expedition**.

Ilustración de P-hacking

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
<http://pss.sagepub.com>


Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

<http://...> (<http://tinyurl.com/cea6krt>)

Este trabajo ilustra lo **sencillo** que resulta obtener resultados **significativos** en estudios estadísticos, **exista o no** efecto subyacente.

In this article, we show that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses make false positives vastly more likely. In fact, it is unacceptably easy to publish “statistically significant” evidence consistent with *any* hypothesis.

This exploratory behavior is not the by-product of malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions and (b) the researcher’s desire to find a statistically significant result. A large literature

Dos experimentos:

Study 1: musical contrast and subjective age

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older. In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song by Mr. Scruff that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated survey: They answered the question "How old do you feel right now?" by choosing among five options (*very young, young, neither young nor old, old, and very old*). They also reported their father's age, allowing us to control for variation in baseline age across participants.

An analysis of covariance (ANCOVA) revealed the predicted effect: People felt older after listening to "Hot Potato" (adjusted $M = 2.54$ years) than after listening to the control song (adjusted $M = 2.06$ years), $F(1, 27) = 5.06, p = .033$.

Individuos se **sienten más mayores** tras escuchar una **canción infantil**.

¿Cómo se puede haber llegado a **estas conclusiones** (sobre todo la segunda)?

Study 2: musical contrast and chronological rejuvenation

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted $M = 20.1$ years) rather than to "Kalimba" (adjusted $M = 21.5$ years), $F(1, 17) = 4.92, p = .040$.

Individuos **SON más jóvenes** tras escuchar "When I'm sixty four" (The Beatles).

Grados de libertad del investigador

En general existen innumerables **factores** de análisis que se eligen de forma **arbitraria**. Estos factores son lo que los autores llaman **grados de libertad** del investigador.

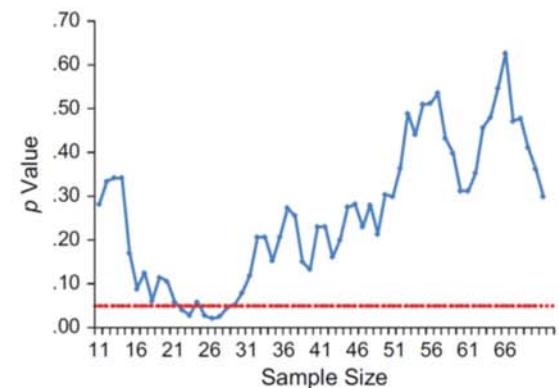
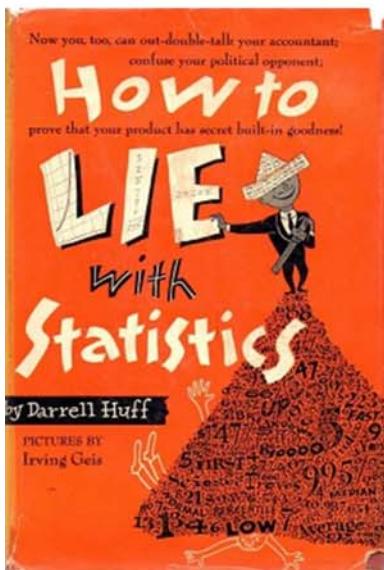
Por **cada grado** de libertad que tenemos podemos hacer **un análisis distinto** y su **combinación multiplica** el número de análisis que podemos hacer.

Grados de libertad:

- Distintas **variables respuesta** (incidencia, prevalencia, supervivencia, mortalidad, ...).
- Distintas **covariables** y sus combinaciones (2^p posibles modelos).
- Selección de **parte de la muestra** (eliminación de outliers, valores perdidos, ...).
- ...

Ambiguity is rampant in empirical research. As an example, consider a very simple decision faced by researchers analyzing reaction times: how to treat outliers. In a perusal of roughly 30 *Psychological Science* articles, we discovered considerable inconsistency in, and hence considerable ambiguity about, this decision. Most (but not all) researchers excluded some responses for being too fast, but what constituted "too fast" varied enormously: the fastest 2.5%, or faster than 2 standard deviations from the mean, or faster than 100 or 150 or 200 or 300 ms. Similarly, what constituted "too slow" varied enormously: the slowest 2.5% or 10%, or 2 or 2.5 or 3 standard deviations slower than the mean, or 1.5 standard deviations slower from that condition's mean, or slower than 1,000 or 1,200 or 1,500 or 2,000 or 3,000 or 5,000 ms. None of these

Otro grado de libertad ampliamente utilizado es la posibilidad de **modular el tamaño muestral** del estudio a conveniencia de los resultados buscados



El uso de los grados de libertad es una **herramienta** de primer orden para encontrar (las haya o no) **asociaciones en los datos**.

Según **Wikipedia** uno de los libros con mayor éxito de la historia de la estadística (1.5 millones de copias vendidas sólo en su edición en inglés).

Un ejemplo con datos simulados

15000 bancos de datos, respuesta independiente de la covariable.

Grados de libertad:

- 2 variables **respuesta**.
- Incremento del **tamaño muestral** si no significativo.
- Uso de **covariable adicional** y su interacción con la original.
- Considerar una variable categórica (**3 grupos**) y hacer análisis 2 a 2 de los grupos.

Researcher degrees of freedom	Significance level		
	p < .1	p < .05	p < .01
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three t tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one t test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a t test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender \times Condition interaction was significant. Results for Situation D were obtained by conducting t tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = -1, medium = 0, high = 1).

La combinación masiva de **grados de libertad** es una técnica de **P-hacking** que conduce a encontrar **relaciones** significativos cuando éstas **no existen** necesariamente.

P-hacking y picos de fertilidad

The screenshot shows a dark red header with the text "Psychological Science". Below it is a light gray navigation bar with five items: "Home", "Browse", "Submit Paper", "About", and "Subscribe".

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall, Jessica L. Tracy

First Published July 10, 2013

“Building on evidence that **men are sexually attracted** to women wearing or surrounded **by red**, we **tested** whether women show a behavioral tendency toward wearing reddish clothing when at peak fertility. ... Women at **high conception risk** were more than **three times more likely to wear a red or pink shirt** than were women at low conception risk. ... Our results thus suggest that red and pink adornment in women is reliably associated with fertility and that **female ovulation**, long assumed to be hidden, is **associated with a salient visual cue**.”

P-hacking y picos de fertilidad (II)

El artículo en breve generó controversia

(http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_in_fertility_research.html) por los “**grados de libertad**” del estudio:

- 9 **colores** (no sólo rojo o rosa) -> múltiples posibilidades y muchas combinaciones.
- Definición de **pico fertil**: entre 6 y 14 días desde el inicio de la menstruación.
- **Otras prendas**, no sólo camisas.
- ...

Bastantes otros resultados podrían dar lugar a “**bonitas historias**”: ¿Mujeres en periodo fertil evitan colores oscuros? ¿Mujeres en periodo fertil usan más tangas?

Estos resultados serían seguramente **espúreos** pero podrían ser **publicados con facilidad**.

P-hacking y fMRI

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 October 2012
doi: 10.3389/fnins.2012.00149



On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments

Joshua Carp *

Department of Psychology, University of Michigan, Ann Arbor, MI, USA

Edited by:
Satrajit S. Ghosh, Massachusetts
Institute of Technology, USA
Reviewed by:

How likely are published findings in the functional neuroimaging literature to be false?
According to a recent mathematical model, the potential for false positives increases with
the flexibility of analysis methods. Functional MRI (fMRI) experiments can be analyzed

Carp (2012) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3468892/pdf/fnins-06-00149.pdf>) enumera los **factores** que intervienen en **análisis estadísticos** de fMRI.

Carp identifica **10 factores** (analysis steps) en la literatura con entre 2 y 4 posibles elecciones, sumando un total de **6912 combinaciones** posibles.

90.3% de los voxels dieron resultados **significativos** para alguna de las 6912 combinaciones de los parámetros evaluadas.

Básicamente, **cualquier voxel** que quisiéramos podría ser catalogado como **significativo**.

P-hacking y + fMRI

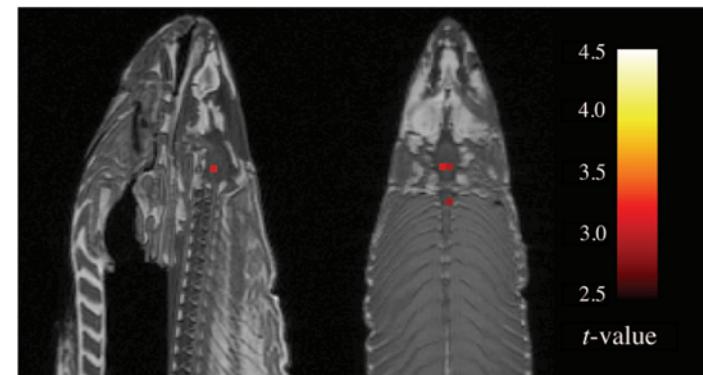
“Premio” **IGnobel** 2012 en neurociencias (Poster original)
(<http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>)

Someten a un **salmón muerto a fMRI** para ver qué regiones cerebrales se activan ante distintos estímulos.

METHODS

Subject. One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.



Si no implementan métodos de corrección de errores adecuado “detectan” regiones cerebrales que se **activan**.

En la fecha en la que el póster original fue presentado, **25-40% de los estudios de fMRI no implementaban corrección de error**. Cuando ganó IGnobel esta cifra había disminuido al 10% (<https://blogs.scientificamerican.com/scicurious-brain/ignobel-prize-in-neuroscience-the-dead-salmon-study/>)

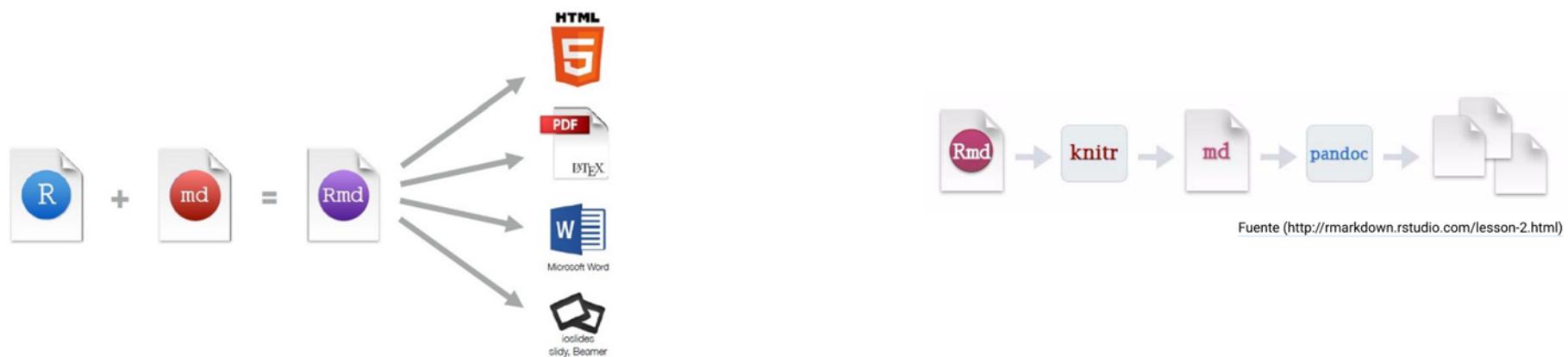
Algunas herramientas estadísticas (del entorno R) para acometer el problema de la reproducibilidad.

R-studio

- R-studio es un **entorno de desarrollo** integrado para R.
- Proporciona un **entorno centralizado** y bien organizado donde **hacer casi todo** lo que quieras en R.
- R-studio participa en la creación y **desarrollo de paquetes** importantes (ggplot2, dplyr, tidyr, lubridate, stringr,...) y herramientas como RMarkdown o Shiny.
- R-Studio está especialmente **bien integrado** con **herramientas** útiles para hacer **investigación reproducible**:
 - Un **software estadístico** para analizar datos (**R**).
 - Lenguajes para crear documentos, presentaciones, libros, artículos o webpages (**LaTeX, Markdown, Shiny**), capaces de integrar código de R.

R-markdown

- R-markdown integra R en **Markdown**, un procesador de texto (versión simplificada de latex).
- Su objetivo será distribuir documentos y el análisis estadístico asociado en un **único documento integrado**.



R Markdown

```
R Markdown rocks
=====
This is an R Markdown document.

```{r}
x <- rnorm(1000)
head(x)
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the `r length(x)` random normal variates we just generated is `r round(mean(x), 3)`. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.

```{r}
plot(density(x))
```

Note that all the previously demonstrated math typesetting still works. You don't have to choose between having math cred and being web-friendly!

Inline equations, such as ... the average is computed as $\\frac{1}{n} \\sum_{i=1}^n x_i$. Or display equations like this:

$$
\\begin{equation*}
|x| =
\\begin{cases} x & \\text{if } x \\geq 0, \\\\ -x & \\text{if } x \\leq 0. \\end{cases}
\\end{equation*}
$$
```

Markdown

```
R Markdown rocks
=====
This is an R Markdown document.

```{r}
x <- rnorm(1000)
head(x)
```

```
[1] -1.3007 0.7715 0.5585 -1.2854 1.1973
2.4157
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the 1000 random normal variates we just generated is -0.081. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.

```{r}
plot(density(x))
```

!{plot of chunk unnamed-chunk-2}(figure/unnamed-chunk-2.png)

...  
...
```

- Podrás encontrar una breve, aunque detallada, **introducción a R-markdown** en este link
https://raw.githubusercontent.com/fisabio/material_publico_cur_inv_rep/master/Sesio

Análisis estadístico **reproducible** con R-markdown:

Example 4.2

Disease mapping: from foundations to multidimensional modeling
Martínez-Benito M.A. and Botella-Rocamora P.

This document reproduces the analysis made at Example 4.2 of the book: "Disease mapping: from foundations to multidimensional modeling" by Martínez-Benito M.A. and Botella-Rocamora P., published by CRC press in 2019. You can watch the analysis made with full detail at this pdf document, or even execute it if you want with the material available at <https://github.com/MiguelBenito/DMBook>. Anyway, this pdf file should be enough for following most of the details of the analysis made for this example.

The statistical analysis below has been run in R, by additionally using the library `ReMarkdown`, so be sure that you have this software installed if you want to reproduce by yourself the content of this document. In that case we advise you to download first the annex material at <https://github.com/MiguelBenito/DMBook>, open with `RStudio` the corresponding `.Rproj` file that you will find at the folder corresponding to this example and compile the corresponding `.Rmd` document. This will allow you to reproduce the whole statistical analysis below.

This document has been executed with real data that are not provided in order to preserve their confidentiality. Slightly modified data are provided instead, as described in Chapter 1 of the book. Thus, when reproducing this document you will not obtain exactly the same results, although they should be very close to those shown here.

Libraries and data loading

```
# Libraries loading
#-----
if (!require(RColorBrewer)) {
  install.packages("RColorBrewer")
  library(RColorBrewer)
}
if (!require(rgdal)) {
  install.packages("rgdal")
  library(rgdal)
}

# Data loading
#-----  
#----- reproducing the document, the following line should be changed to
# load("../Data/ObseRval-m2.Rdata") since that file contains the
# modified data making it possible to reproduce this document.
load("../Data/ObseRval.Rdata")
load("../Data/Expval.Rdata")
load("../Data/Population.Rdata")
load("../Data/VR.Rdata")
```

Choropleth maps

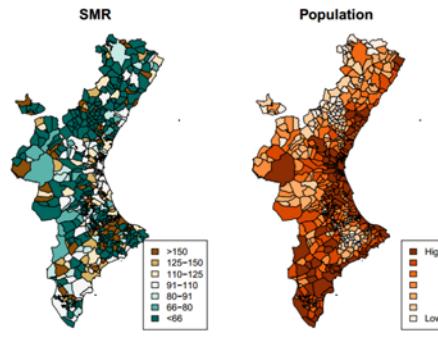
```
# SMRs
SMR.muni <- 100 * Obs.muni/Exp.muni
```

```
# Population
Pop.muni <- apply(PopM, 1, sum)/25

# Choropleth maps
par(mfrow = c(1, 2))
par(mar = c(0, 0, 1, 0) + 0.1)

colors <- brewer.pal(7, "BrBG")[7:1]
plot(VR.cart, col = colors(as.numeric(cut(SMR.muni, 100 * c(-0.1, 1/1.5,
  1/1.1, 1/1.25, 1.5, 100)))))
title("SMR", cex = 0.75)
legend(x = "bottomright", fill = colors[7:1], legend = c(">150", "125-150",
  "110-125", "91-110", "80-91", "66-80", "<66"), cex = 0.75, inset = 0.03)

colors2 <- brewer.pal(7, "Oranges")
plot(VR.cart, col = colors2(as.numeric(cut(Pop.muni,
  (1/6)/7), 1e+07)))
title("Population", cex = 0.75)
legend(x = "bottomright", fill = colors2[7:1], legend = c("High", "", "", "", "", "Low"), cex = 0.75, inset = 0.03)
```



Population in the municipalities of high and low risks

Cualquier **otro estadístico** con los conocimientos adecuados debería ser capaz de **reproducir de forma fidedigna el análisis estadístico** del estudio correspondiente.

Este es el **cambio** en la forma de trabajar **del que hablábamos** en la presentación de reproducibilidad.

R-markdown como referencia de reproducibilidad:



Ben Marwick
@benmarwick

❤️ this editorial just appearing in @bmj_latest by @bengoldacre @dr_c_morton & @NDevito1: "Why researchers should share their analytic code", recommending #rstats markdown as best practice 🏆
Every journal needs an editorial like this 👍👍👍
doi.org/10.1136/bmj.l6...

Traducir Tweet

The screenshot shows the full text of the BMJ editorial. The title is "Why researchers should share their analytic code". It discusses the importance of transparency in research and how sharing code can increase reproducibility. It also mentions the use of R Markdown as a best practice. The text is in two columns, with some sections highlighted in blue.

6:50 a. m. · 23 nov. 2019 · Twitter for Android



Daniël Lakens ✅
@lakens

The question is not whether we will move to a system where data and code will be expected to be made available by default, but when the journals you submit to will start expecting this from you. Train yourself in computational reproducibility. RMarkdown is a great place to start.

Traducir Tweet



Ben Marwick @benmarwick · 12h

❤️ this editorial just appearing in @bmj_latest by @bengoldacre @dr_c_morton & @NDevito1: "Why researchers should share their analytic code", recommending #rstats markdown as best practice 🏆 Every journal needs an editorial like this 👍👍
doi.org/10.1136/bmj.l6...

The screenshot shows the full text of the BMJ editorial. The title is "Why researchers should share their analytic code". It discusses the importance of transparency in research and how sharing code can increase reproducibility. It also mentions the use of R Markdown as a best practice. The text is in two columns, with some sections highlighted in blue.

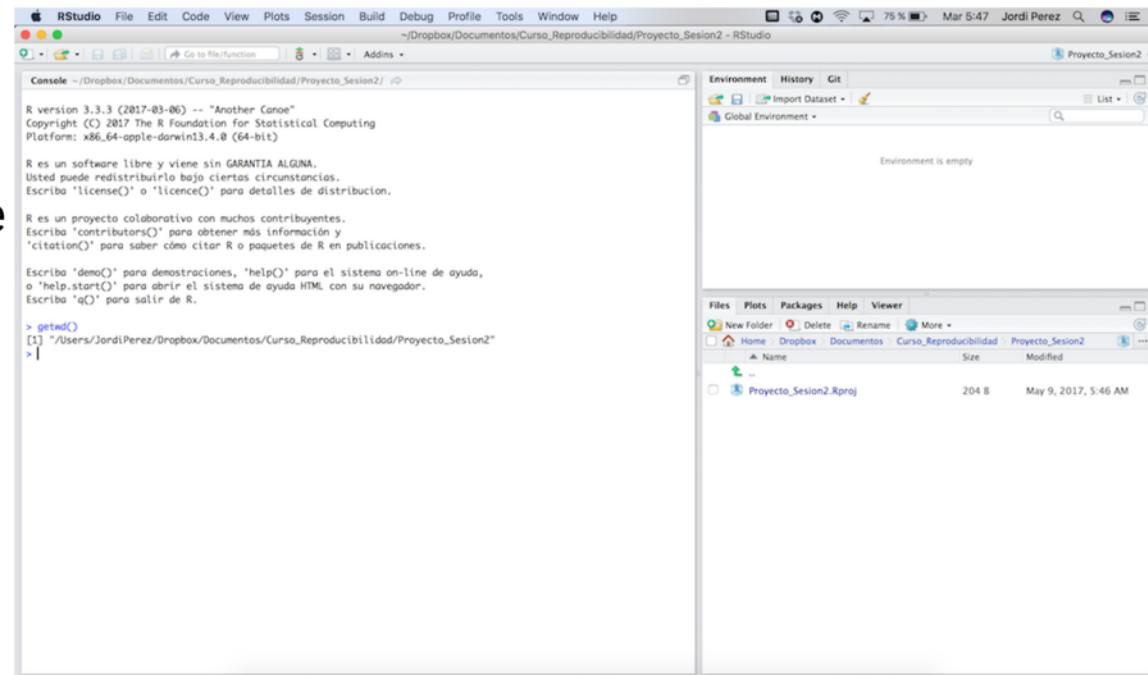
7:05 a. m. · 23 nov. 2019 · Twitter for Android

Proyectos de R.

- Otra de las **utilidades de R-studio**.
- **Ventajas** del uso de proyectos.
 - Recuperación de la **configuración** previa de la sesión de R para dicho proyecto y de la **última sesión de dicho proyecto** (archivos abiertos tal y como los dejamos).
 - Establece como **working directory** aquel que contiene el archivo del proyecto.
 - Guardado automático del **historial** de la sesión, y si se quiere del **entorno de trabajo**.
 - Unido al uso de **rutas relativas** permite que la **ejecución** de las rutinas del proyecto funcionen **independientemente del ordenador y ruta** que contenga al proyecto.

¿Cómo trabajar con proyectos en R-studio?

- **Creación de proyectos:**
 - R-studio>File>New project.
 - Rellenar las **opciones** en función de qué se pretenda (crear proyecto en nuevo directorio, crear proyecto en directorio ya existente, ...).
- El proceso anterior **crea un archivo** de proyecto en el directorio determinado y **inicia** en R-studio la **sesión** de dicho proyecto.
- Haciendo **doble click** en el archivo de proyecto generado recuperaremos la **sesión tal cual la dejamos**.



Proyectos de R-studio y reproducibilidad.

- Resulta evidente que los **proyectos** de R-studio son una opción muy **conveniente** desde el punto de vista **práctico**.
- Sin embargo, los **proyectos** presentan una ventaja fundamental que los convierten en una importante **herramienta** de mejora de la **reproducibilidad**.
- Si **guardamos los archivos** que integran el estudio en el **mismo directorio que el** archivo de **proyecto** podremos **trasladar** el material a un nuevo directorio o mandárselo a un nuevo usuario y todo **debería seguir funcionando**, tal cual.
- Únicamente tendremos que llevar cuidado de usar **rutas relativas** al llamar a nuestros archivos.
- Por tanto, el uso de **proyectos** hace los **análisis estadísticos independientes de la máquina y el directorio** en el que sean ejecutados, favoreciendo su reproducibilidad.

Uso de rutas relativas

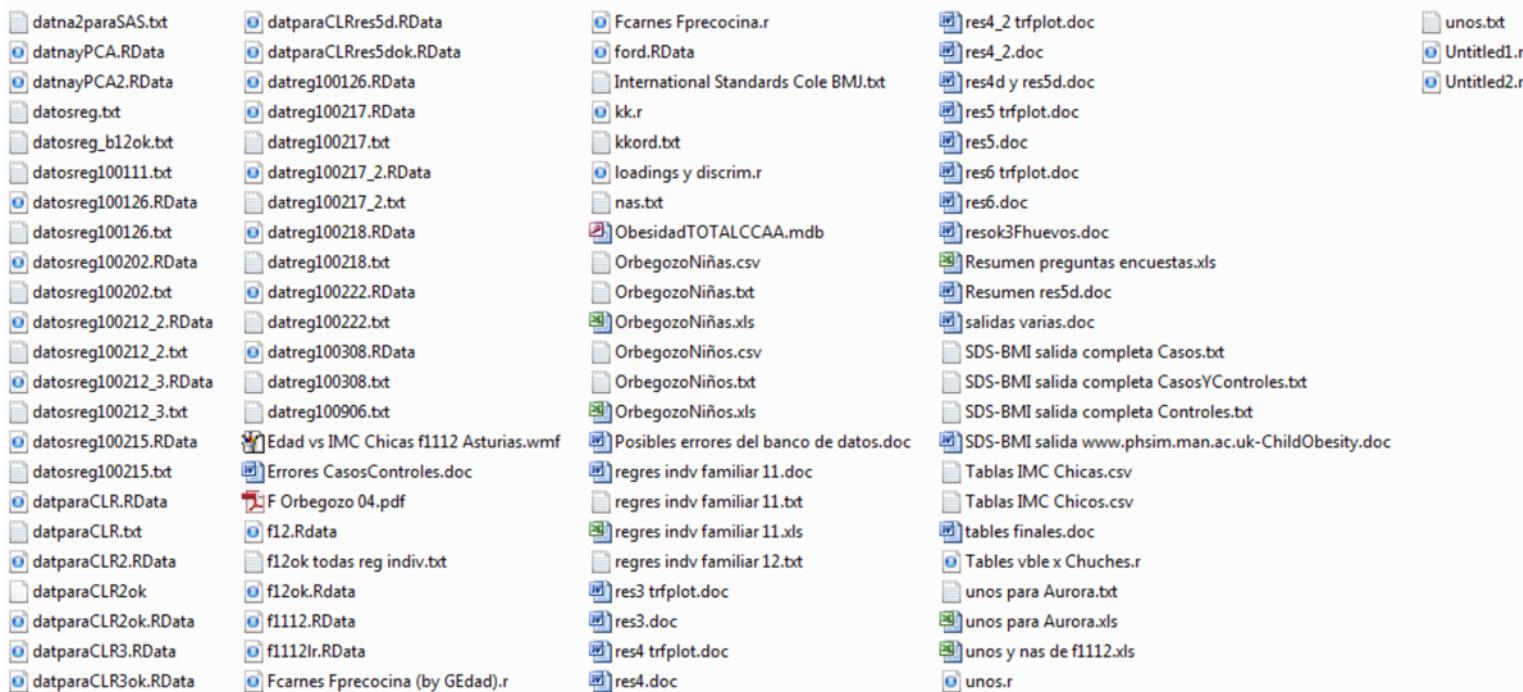
- **Ruta absoluta:** c:/trabajo/analisis/resul.pdf
- **Ruta relativa:** ./analisis/resul.pdf
- El uso de **rutas relativas** es siempre mucho más adecuado ya que hace las **rutinas independientes del directorio de trabajo** donde estén ubicadas. Su uso **se aconseja en TODOS** los análisis siempre que resulte posible.
- **Caracteres especiales:**
 - “.” directorio **actual**.
 - “..” sube al **nivel superior** de un directorio.
 - “~” **directorio home** del usuario (en Windows: C:/users/USUARIO/Documents).
- Un problema para la reproducibilidad de estudios es la sintaxis de rutas en Windows (datos\datos_brutos) y Linux/Mac (datos/datos_brutos). La función file.path construye rutas adecuadas independientemente del sistema (file.path("datos","datos_brutos")) .

Normas de estilo para redacción de sintaxis

- Cuando se **colabora** en un proyecto, un código tiene **muchos autores y lectores**. En estos casos es muy útil tener un **estilo claro** y bien definido, y a ser posible, **común**.
- “Good coding style is like correct punctuation: you can manage without it, but it’s sure to make things easier to read”.
- Existen **distintas guías de estilo** ([Hadley Wickham \(style.tidyverse.org\)](https://style.tidyverse.org) o [Google \(https://google.github.io/styleguide/Rguide.xml\)](https://google.github.io/styleguide/Rguide.xml)), que proponen normas de redacción de sintaxis de R. **Cualquiera puede ser buena**, lo importante es adoptar una concreta y acostumbrarse a desarrollar sintaxis con ella.
- **Cuestiones consideradas** en guías de estilo: Nombres de objetos, funciones y archivos; espaciado (operadores matemáticos, funciones if/for, ...); tabulación; longitud de las líneas de código; llaves/paréntesis; ...
- El paquete [formatR \(https://cran.r-project.org/web/packages/formatR/index.html\)](https://cran.r-project.org/web/packages/formatR/index.html) de R implementa de **forma automática** un conjunto de normas sobre el archivo de sintaxis que se deseé.

Organiza tu trabajo

- Los **directorios** de trabajo de distintos análisis estadístico habitualmente no tienen ninguna **estructura más o menos lógica**.
 - Una **estructura** que podría parecer **lógica en un principio** se puede **convertir** fácilmente con el tiempo en algo **ineficiente**.



- La **organización** del material en el directorio de trabajo permite **entenderse** con los miembros de tu grupo de trabajo incluso !!contigo mismo!! (pasado un tiempo), además su **carenica** supone una potencial **fuente de errores**.
- Organiza tu trabajo de forma **sistemática y consistente**.
- El uso constante de la **misma estructura** de directorios hará que **encuentres tus archivos** con facilidad, y la **programación** de tus análisis **más facil e intuitiva**.
- Prácticamente **cualquier propuesta** razonable de estructura de directorios **puede ser buena**, lo importante es que siempre utilices **la misma**.

The screenshot shows the RStudio interface. In the top-left pane, there is an 'Untitled1*' script with the following R code:

```

1 # Vamos a crear la estructura de directorios inicial del proyecto
2 dir.create("r")
3 dir.create("figuras")
4 dir.create("informes")
5 dir.create("datos")
6 dir.create("datos/brutos")
7 dir.create("datos/procesados")
8
9
10

```

In the bottom-left pane, the R Console displays:

```

Console <-/Dropbox/Documentos/Curso_Reproducibilidad/Proyecto_Sesion2/ >
Escriba 'contributors()' para obtener mas información y
'citation()' para saber como citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

[Workspace loaded from ~/Dropbox/Documentos/Curso_Reproducibilidad/Proyecto_Sesion2/.RData]

> # Vamos a crear la estructura de directorios inicial del proyecto
> dir.create("r")
> dir.create("figuras")
> dir.create("informes")
> dir.create("datos")
> dir.create("datos/brutos")
> dir.create("datos/procesados")
>

```

The right side of the interface includes the Environment pane (empty), the Global Environment pane, and the Files pane showing the directory structure:

| Name | Size | Modified |
|------------------------|--------|-----------------------|
| .. | | |
| .gitignore | 14 B | May 9, 2017, 6:05 AM |
| RData | 2,5 KB | May 9, 2017, 6:31 AM |
| Rhistory | 146 B | May 16, 2017, 6:42 AM |
| Rprofile | 117 B | May 9, 2017, 6:05 AM |
| packrat | | |
| Proyecto_Sesion2.Rproj | 192 B | May 9, 2017, 6:05 AM |
| datos | | |
| figuras | | |
| informes | | |
| r | | |

Paquetes de R.

- Un **paquete** de R es un conjunto de **funciones y datos**, junto con **documentación** para su uso.
- Normalmente **pensamos en paquetes** de R como **proyectos de gran envergadura** (depositados en CRAN), posiblemente por encima de nuestras intenciones y capacidades.
- Pero:
 - El uso de paquetes es una gran herramienta para **compartir información/datos** con tu **grupo de trabajo**, o un círculo más amplio de potencialmente interesados.
 - Su **sistema de versiones** hace que todo el mundo pueda acceder de forma sencilla a **versiones actualizadas** del material y, a su vez, todos compartamos un **mismo material**.
 - La obligación de **documentar el contenido** del paquete hace que su contenido esté **ordenado y explicado**, lo que supone en una gran ventaja.

Paquetes de R como herramientas de reproducibilidad: MEDEA3.

- El proyecto **MEDEA3** es un proyecto de investigación **colaborativo** que reune a 14 grupos de investigación nacionales.
- Su objetivo es el estudio de la **distribución geográfica de la mortalidad en grandes ciudades** españolas de distintas comunidades autónomas.
- MEDEA3 ha sido **coordinado** desde la Fundación **FISABIO** en Valencia.
- MEDEA3 dispone de un **repositorio de datos común** a todos los grupos del proyecto, así como unas **rutinas de análisis** que todos los grupos habrían de seguir.
- El **repositorio común** de funciones/datos/utilidades del proyecto ha sido estructurado en formato de **paquete de R**, accesible públicamente en el repositorio **fisabio/medear** de GitHub (instalable ejecutando en R:
`remotes::install_github("fisabio/medear")`).

- La **ayuda** del paquete supone una **excelente documentación** del proyecto.

Documentation for package ‘medear’ version 0.7.10

- [DESCRIPTION file](#).

Help Pages

| | |
|--|--|
| medear-package | Poblaciones y cartograf a por secci n censal del INE (proyecto MEDEA3) |
| bboxm3 | L mites ('bbox') para los Mapas en MEDEA3 |
| cambios_seccion | Cambios Temporales de Seccionaldo para todas las Ciudades MEDEA3 (per odo 1996-2015) |
| carga_datos | Carga los datos privados del proyecto MEDEA3 |
| cartografia | Cartograf a por Secci n Censal para las Ciudades MEDEA3 (INE 2011) |
| carto_medea3 | Cartograf a por Secci n Censal para las Ciudades MEDEA3 (1996-2015) |
| causas_defuncion | Agrupa las causas de mortalidad |
| censo | Censos de 2001 y 2011 |
| codigos_ine | Nombres de Municipios y Provincias seg n Terminolog a Oficial del INE |
| comprueba_datos | Comprobaciones de la clase de los datos |
| comprueba_geocodificado | Comprobaci n de la asignaci n de distintas coordenadas a una misma direcci n |
| comprueba_punto_poligono | Comprobaci n de inclusi n de una coordenada dentro de un pol gono |
| crea_cubo_mortalidad | Crear la matriz 5-dimensional de mortalidad |
| crea_cubo_poblacion | Crear la matriz 4-dimensional de poblaciones |
| descarga_cartografia | Descarga la cartograf a con el seccionaldo del INE para 2011 |
| descarga_poblaciones | Descarga poblaciones del INE por secci n censal, sexo, edad y a o |
| descarga_trameros | Descarga los callejeros del INE |
| detecta_cambios | Funci n para detectar cambios de seccionaldo en los callejeros INE |
| detecta_cluster | Detecci n de agrupaciones de mortalidad a revisar manualmente |
| elimina_cambios | Eliminar cambios manualmente de la base de datos cambios_seccion |

- El paquete contiene **utilidades cartográficas** (`cartografia`, `bboxm3`, ...), **datos** públicos (`censo`, ...) y privados encriptados (`carga_datos`, ...), **funciones** útiles para la ejecución del proyecto (`detecta_cambios`, `crea_cubo_mortalidad`, ...) ...

- Además, el paquete `medear` dispone de **viñetas** (vignettes) con la **descripción** detallada del **comando de R** a ejecutar en cada fase de análisis del proyecto.
- Por ejemplo:
 - Viñeta para dar **formato a los datos** de cada ciudad del proyecto (`medear-formato-datos.html`)
 - Viñeta para el **proceso de geocodificación** de los datos del proyecto (`medear-geocodificacion.html`)
- De esta manera resulta evidente cómo el proyecto **MEDEA3** es (o al menos pretende ser) **reproducible**, con la salvedad de la confidencialidad de los datos.
- Más allá del proyecto MEDEA3, los **paquetes de R** son una excelente herramienta para **ordenar, documentar** ... funciones y datos que utilicemos habitualmente. Los paquetes de R son una manera de **ser ordenado** y por tanto **minimizar las fuentes de error** de nuestros análisis.

Para concluir

- La **estadística** está en el **nucleo de la investigación científica**, concretamente en toda la literatura cuantitativa.
- Los estadísticos jugamos un **papel clave** en la resolución de los problemas de **reproducibilidad y replicabilidad** que estamos viviendo.
 - **Reproducibilidad:** Haciendo uso de **procedimientos y herramientas** que hagan nuestra investigación **reproducible**.
 - **Replicabilidad:** Introduciendo una **componente ética** en nuestra investigación que permitan evitar las **principales causas** de este problema, como el **P-hacking**.

!!Seamos responsables, comprométamonos, la validez de la literatura científica nos va en ello!!

