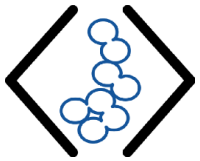
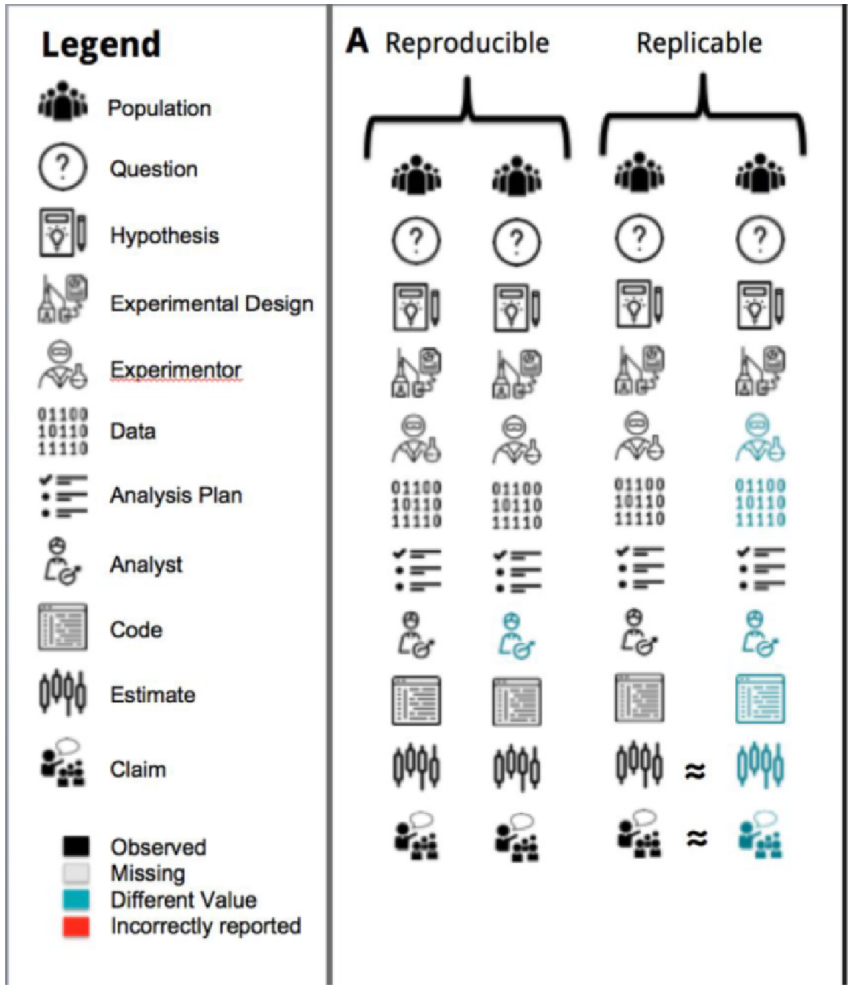


P-hacking

Miguel A. Martinez-Beneito - Área de Desigualdades en Salud, FISABIO-Salud Pública.

8 de junio de 2017

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/826860

"If you **torture** the data long enough, it will **confess**" (https://en.wikiquote.org/wiki/Ronald._Coa

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

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas 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**.

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), $F(1, 17) = 4.92$, $p = .040$.

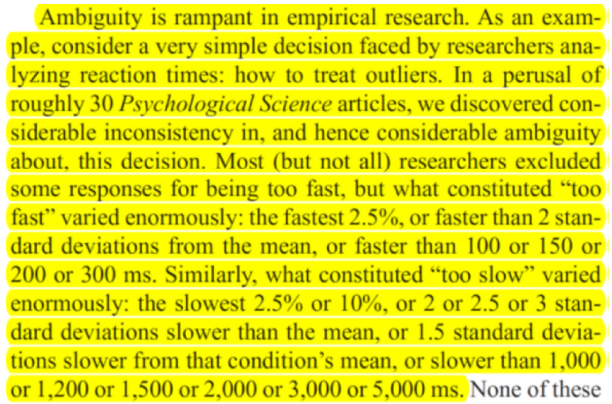
Individuos **SON más jóvenes** tras escuchar "When I'm sixty four" (The Beatles).

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

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.

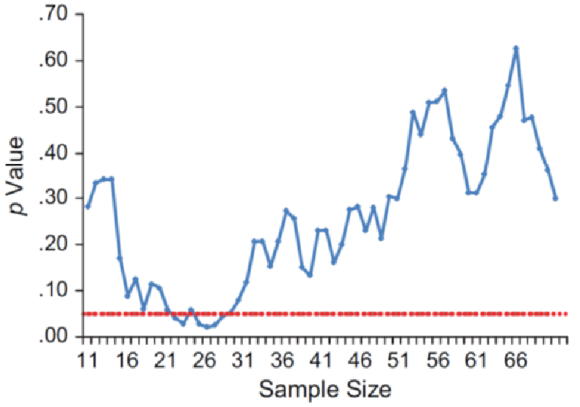


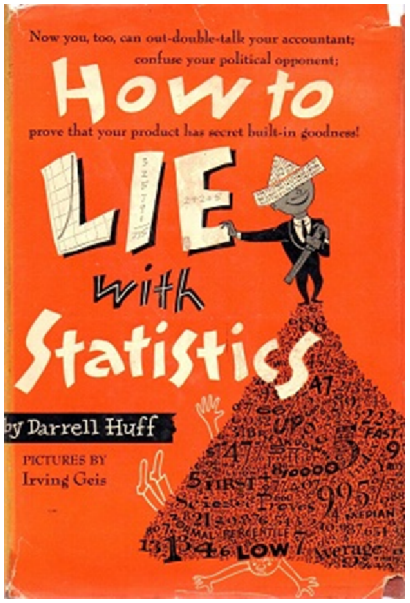
Ambiguity is rampant in empirical research. As an example, consider a very simple decision faced by researchers analyzing reaction times: how to treat outliers. In a perusal of roughly 30 *Psychological Science* articles, we discovered considerable inconsistency in, and hence considerable ambiguity about, 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

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, …).
- …

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 ingles).

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 × 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).

El uso de **grados de libertad** en los dos estudios anteriores es un ejemplo de **P-hacking** que conduce a encontrar **relaciones** significativos cuando **no existen** realmente.

P-hacking y picos de fertilidad

"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)

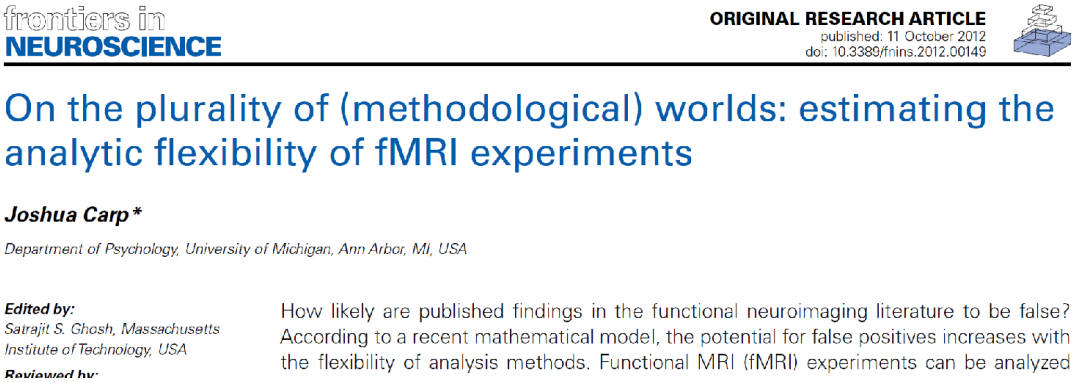
El artículo en breve generó controversia (http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psycl por los "**grados de libertad**" del estudio:

- 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

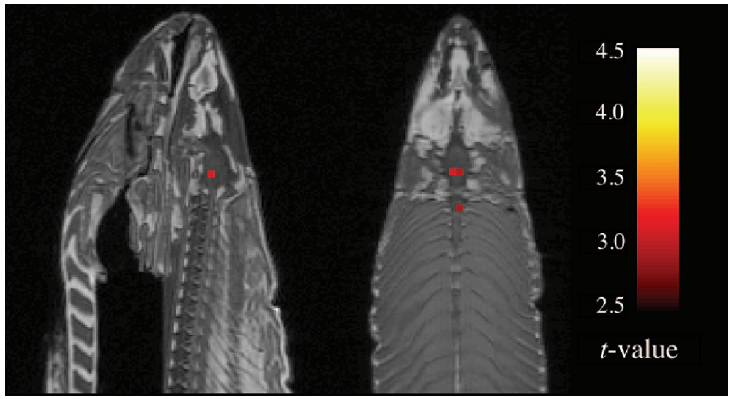
"Premio" **IGnobel** 2012 en neurociencias (Poster original) (http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf)

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/)