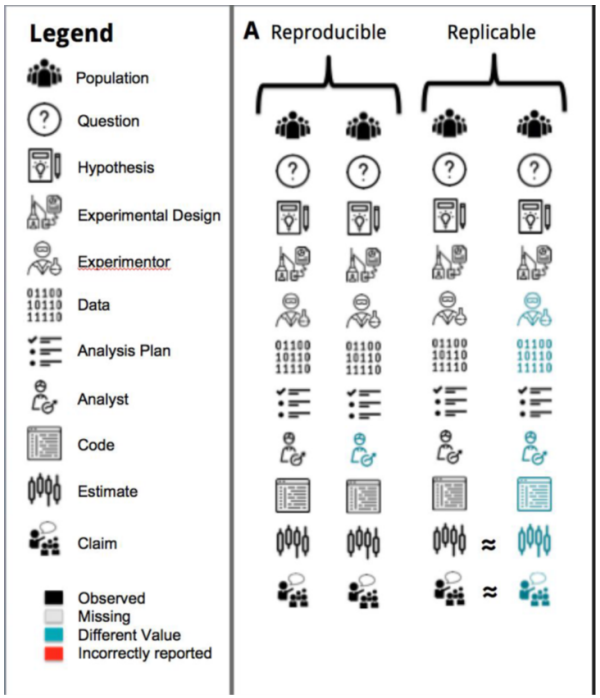
Replicabilidad en la investigación científica

Sesiones de estadística DAP-Cat

Miguel A. Martinez-Beneito
miguel.a.martinez@uv.es

Reproducibilidad: El problema

Reproducibilidad y replicabilidad



- **Reproducibilidad**: **Mismos datos, distintos experimentadores**, resultados (presumiblemente) iguales -> **Transparencia** de procesos.
- **Replicabilidad**: **Distintos datos, similares condiciones experimentales**, resultados similares -> **Generabilidad** de los resultados.

- **Reproducibilidad** no implica **validez** de los resultados, sólo transpariencia. "Results that are not reproducible are **hard to verify** and results that do not replicate in new studies are **hard to trust**."

¿También crisis de replicabilidad?

Replication crisis

From Wikipedia, the free encyclopedia

The **replication crisis** (or **replicability crisis**) refers to a methodological crisis in science in which scientists have found that the results of many scientific experiments are difficult or impossible to replicate on subsequent investigation, either by independent researchers or by the original researchers themselves.[1] While the crisis has long-standing roots, the phrase was coined in the early 2010s as part of a growing awareness of the problem.

SCIENCE NEWS | Wed Mar 28, 2012 | 7:09pm BST

In cancer science, many "discoveries" don't hold up

Cancer Research Is Broken

There's a replication crisis in biomedicine—and no one even knows how deep it runs.

2.1k 696 83

By Daniel Engber

- Las crisis de reproducibilidad y replicabilidad han dado lugar a un campo de investigación emergente, **metaciencia**, que se encarga del **estudio científico de la ciencia** en sí misma y los factores que influyen en la validez de sus resultados.

Replicabilidad en la literatura médica

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most

- Ioannidis (JAMA, 2005) lleva a cabo un **estudio de replicación** de trabajos publicados entre 1990 y 2003 en revistas médicas con IF>7 y con más de 1000 citas.
- **Busqueda bibliográfica** de estudios similares, con criterios de calidad (tamaño muestral, diseño, . . .) **similares o superiores**, que pudieran **corroborar o refutar** dichos estudios.

- De los **34** artículos **elegibles**:
  - En **7** de ellos el efecto original no se ha podido replicar.
  - En otros **7** el efecto original o su duración se reduce a menos de la mitad.
  - En **20** ocasiones el efecto original ha sido corroborado.
- Así, sólo el **58.8%**(=20/34) de los estudios testados son **corroborados** por estudios de similares características.
- Además, los estudios **más antiguos** tienen **más** probabilidad de haber sido **refutados** por lo que los resultados podrían ser peores.

Replicabilidad en Psicología

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original

- En 2015 (Science), el Center for Open Science publica los resultados de un **estudio colaborativo de replicación** en psicología. "One of the top 10 scientific breakthroughs of the year (Science)"
- Distintos **grupos** de forma independiente **replican 100 estudios** influyentes publicados en la literatura.
- Sólo **39 de los 100** estudios corroboran los resultados originales.

El "Decline Effect"

- Pero ni siquiera la publicación de **replicas positivas** de trabajos es **garantía** de nada.
- **Facciones asimétricas** se consideran signo de **mutaciones genéticas**.

Nature **357**, 238-240 (21 May 1992) | doi:10.1038/357238a0; Accepted 17 March 1992

Female swallow preference for symmetrical male sexual ornaments

Anders Pape Møller*

ARTICLE

Send to

Export c

Export r

- Moller (Nature, 1991) descubre que **hembras de golondrinas prefieren** machos con **plumaje más simétrico**.
  - **Simetría** de plumaje sería indicador indirecto de **calidad genética**.
  - Hembras **aplican** este criterio (mecanismo de selección genética) de forma **inconsciente**.

- En los 3 años siguientes **9 (de 10)** artículos corroboran la teoría anterior.
- Dicha teoría se estudia en **humanos**:
  - Mujeres prefieren el **olor** de hombres con facciones **simétricas**, aunque sólo durante su periodo **fértil**.
  - Mujeres tenían más **orgasmos** con hombres **simétricos**.
  - **Bailarines** simétricos son consistentemente **evaluados** como mejores.
- Pero:
  - En 1994 sólo **8 de 14** artículos corroboran el efecto original.
  - En 1995, **4 de 8** artículos lo corroboran.
  - En 1998, **4 de 12** artículos lo corroboran.
  - De 1992 a 1997, el **efecto** originalmente encontrado **decae en un 80%**.

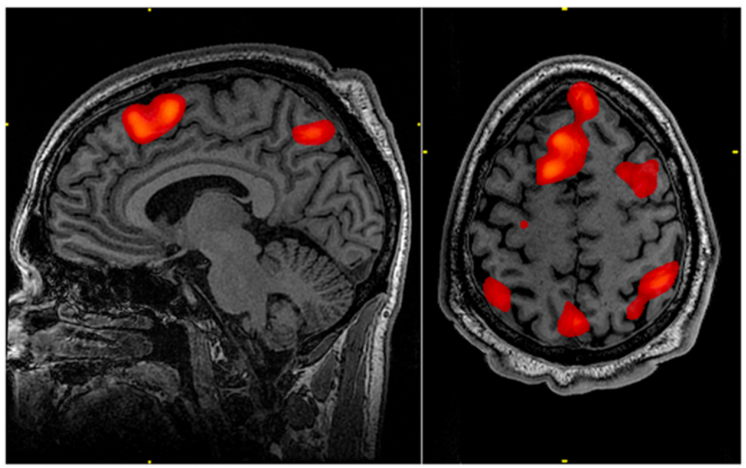
Decline effect, estructura general:

- Tras proponerse un paradigma, los **procesos de publicación científica tienden a corroborarlo** (sesgo de publicación). **Años después** los incentivos editoriales cambian, difundiendo aquellos **resultados que desaprueban** el paradigma establecido.
- Decline effect ha sido documentado en muchas **más situaciones**:
  - **Efectividad** de fármacos **antipsicóticos** de segunda generación.
  - **Percepción extra-sensorial**: Capacidad de predecir hechos futuros.
  - ...
- Decline effect es una **expresión** más **de la crisis** de replicabilidad y de la repercusión que puede tener el **sesgo de publicación** en este problema.

Replicabilidad: Algunas causas de la crisis.

fMRI

- **fMRI** ha sido la herramienta principal para estudiar la **funcionalidad** de cada región **cerebral**.



An FMRI scan during working memory tasks.

- Habitualmente, **individuos** se someten a **estímulos** y, mediante resonancia magnética, se determinan las **regiones del cerebro** con mayor consumo de hemoglobina tras dicho estímulo.
- Dichas **áreas** serían las **encargadas de procesar** el estímulo.

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

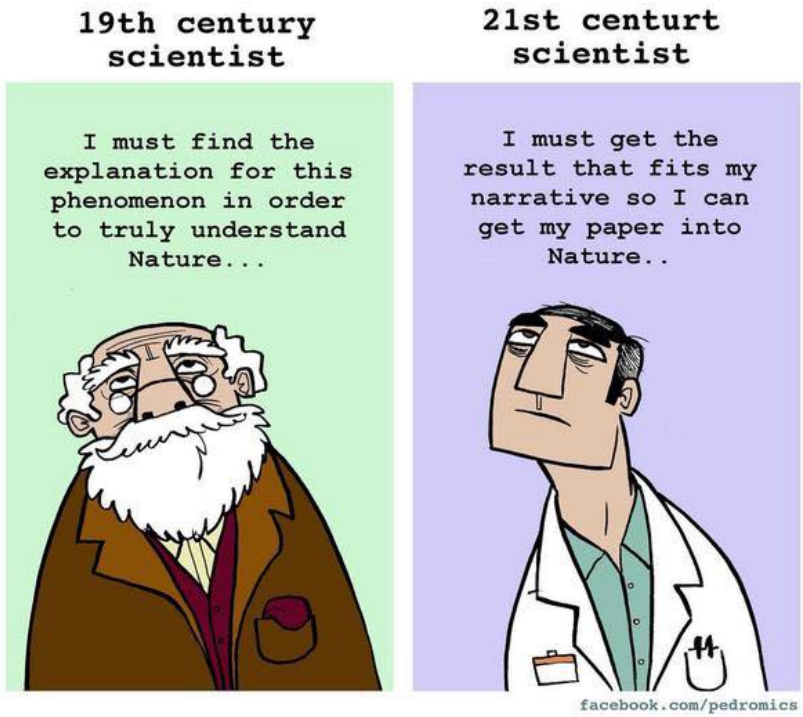
PNAS, 12/07/2016

- "We used **resting-state fMRI data** from 499 healthy controls ... **Using this null data we estimate the incidence of significant results**. In theory, **we should find 5% false positives**, but instead **we found** that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in **false-positive rates of up to 70%**. These results question the validity of a number of fMRI studies and may have a large impact on the interpretation of weakly significant neuroimaging results."
- La alta tasa de **falsos positivos** se debe a **deficiencias en los procedimientos** de análisis estadístico.
- Unos **3500 artículos** podrían estar **afectados** (alrededor del 9% de la literatura del campo).

2.- Conflictos de intereses

- Conflictos de intereses **distorsionan la literatura** científica, sesgándola en **direcciones interesadas**.

- No se pública en función de la **evidencia** sino de la **conveniencia**.
- Las **revistas top** demandan resultados **sorprendentes**, más que sólidos, para publicar un artículo.



- La **presión** de los científicos **por publicar** supone un claro conflicto de interés (avidez por encontrar resultados significativos).
- Pero aún hay **más** . . .

Intereses comerciales, políticos y sociales.

- **Investigación científica** ofrece un estupendo **argumento a decisiones o intereses** arbitrarios que de otra forma serían difíciles de justificar.
- La **aureola de veracidad** de la ciencia **legitima decisiones** políticas y sociales ("evidence based policy") o intereses comerciales.
- Grupos de presión, asociaciones, lobbies ... buscan **sustento** en **literatura científica**.
- La ciencia, a menudo, **no** se usa para **guiar** decisiones sino **para justificarlas**.

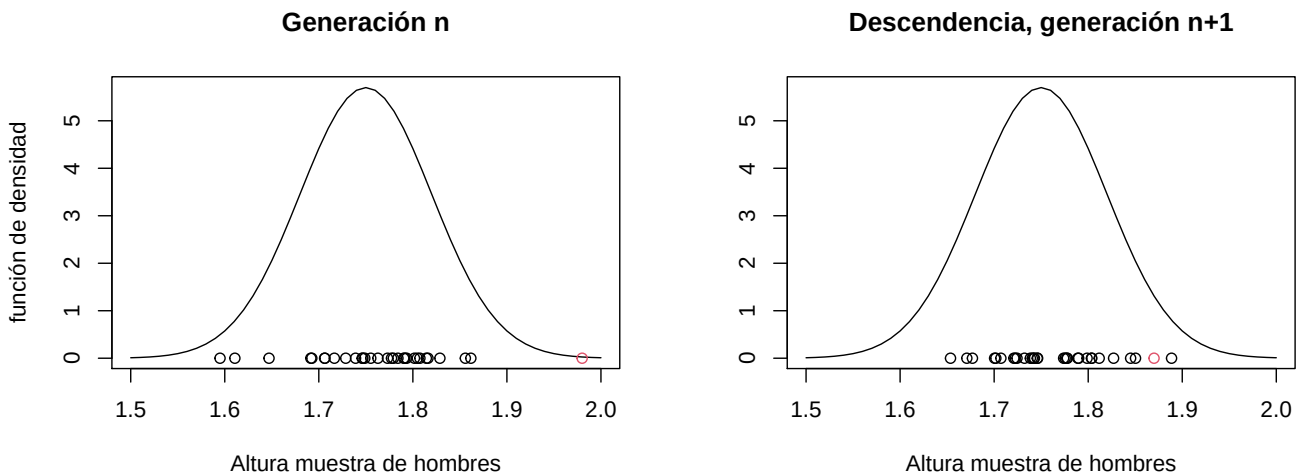
El caso de **Actimel**

- **Actimel**: uno de los productos estrellas de Danone que reporta un **25%** de la facturación del grupo.
- **Alimento funcional** con supuestas propiedades beneficiosas para la salud (**"mejora tus defensas"**).
- Su principal propiedad, reducción de diarreas, se atribuye a la presencia de una cepa patentada de **Lactobacilus Casei ~~Imunitass~~**.
- Sus **efectos positivos** se sustentan (supuestamente) en **literatura** científica generada al efecto. De ahí la **importancia comercial** de disponer de literatura científica que permita aducir estas propiedades.

- La European Food Safety Administration (**EFSA**) emite en 2010 un **informe** sobre los **efectos** de este producto sobre la salud.
- El informe "**rechazó** los más de **20 trabajos** de investigación que pretendían avalar los beneficios del Actimel".
- "The Panel concludes that the **evidence** provided is **insufficient to establish a cause and effect relationship** between the consumption of Actimel and a reduction of the risk of C. difficile diarrhoea"
- La **EFSA destaca errores** repetidos de procedimiento en los estudios evaluados.

3.- Regresión a la media

- Es el fenómeno estadístico por el que los **individuos con observaciones extremas** tenderán a estar **más cerca a la media** de la población en posteriores observaciones.

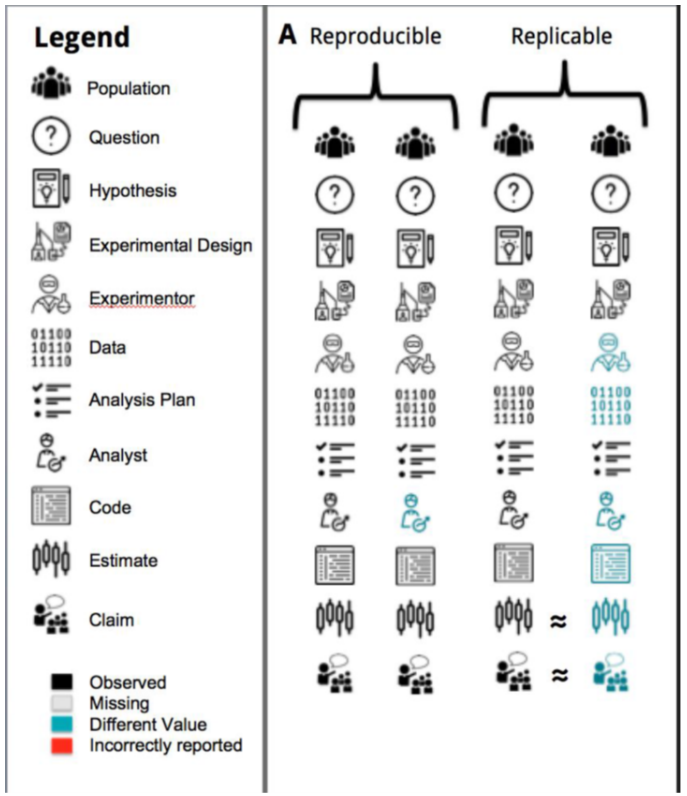


- Este fenómeno de "regresar" hacia la media **dio nombre** originalmente a los modelos de regresión.

- Estudios son **muestras de tamaño 1** del universo de posibles estudios. Sacamos **conclusiones a partir** de una muestra de un **único individuo**.
- Tomando una **muestra** lo **suficientemente extrema** podremos demostrar virtualmente **cualquier resultado**. Un 5% de las muestras posibles, simplemente por azar, nos deberían dar un resultado significativo (aunque no existiera efecto).
- Así, muchos **hallazgos "significativos"** pueden ser simplemente **outliers estadísticos** (muestras suficientemente anómalas) que al intentar ser replicados **pierden su excepcionalidad**.

4.- P-hacking

- "Given a population, hypothesis, experimental design, experimenter, data, analysis plan and analyst the **code changes to match** a desired experiment"
- "If the data can **speak** for themselves they can also **lie** for themselves"
- "If you **torture** the data long enough, it will **confess**"



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

Grados de libertad en la investigación

- Dentro de **cualquier investigación** científica hay un buen número de **decisiones** más o menos **arbitrarias** (grados de libertad) que hemos de tomar.
- La **combinación** de todas estas decisiones produce un **gran número** de análisis estadísticos posibles, posiblemente **alguno de ellos** pueda conducir a **resultados significativos**, exista o no efecto subyacente.

- Este tipo de prácticas hace relativamente **fácil encontrar efectos** "significativos", existan éstos o no.

> 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

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.

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
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).

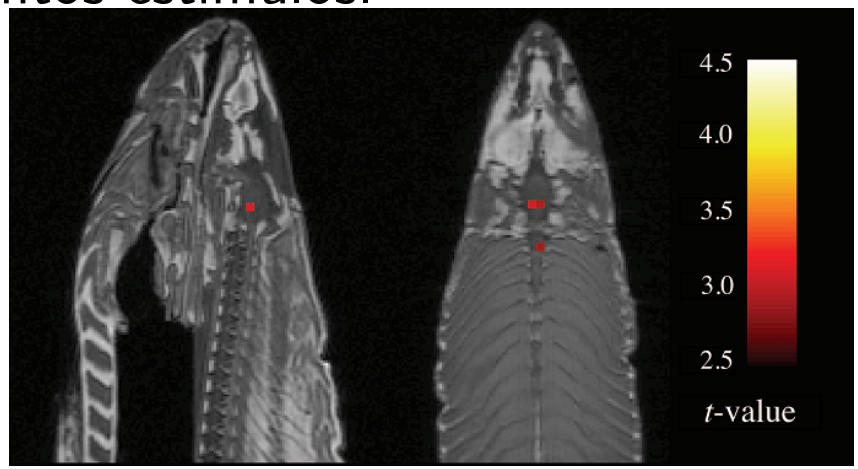
P-hacking y fMRI.

- "Premio" **IGnobel** 2012 en neurociencias (Poster original)
- 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%.

En resumen (mensajes para llevarnos a casa)

- La literatura científica, a día de hoy, se enfrenta a al menos dos crisis, una **crisis de reproducibilidad** y una segunda de **replicabilidad**.
- La **crisis de reproducibilidad** nos invita a ser **escépticos** con la corrección de **estudios concretos**, volver a los orígenes del método científico.
- La **crisis de replicabilidad** nos invita a ser **escépticos más allá** de la corrección procedimental de **ciertos trabajo**.
- !!**No creas todo** lo que lees (en la literatura científica), cultiva el espíritu crítico, la validez de la **Ciencia** nos **va en ello**!!