

# ¿Están los p-valores condenados al destierro?

Mathieu Kessler

Universidad Politécnica de Cartagena

Almería, 25 de Noviembre de 2022



Marc Chagall, 1961, "Adam y Eva expulsados del paraíso"

En febrero 2015:

---

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1-2, 2015  
Copyright © Taylor & Francis Group, LLC  
ISSN: 0197-3533 print/1532-4834 online  
DOI: 10.1080/01973533.2015.1012991



## Editorial

David Trafimow and Michael Marks

*New Mexico State University*

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that

## Editorial

David Trafimow and Michael Marks

*New Mexico State University*

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that

En febrero 2015:

---

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1-2, 2015  
Copyright © Taylor & Francis Group, LLC  
ISSN: 0197-3533 print/1532-4834 online  
DOI: 10.1080/01973533.2015.1012991



## Editorial

David Trafimow and Michael Marks

*New Mexico State University*

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that

En febrero 2015:

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1-2, 2015  
Copyright © Taylor & Francis Group, LLC  
ISSN: 0197-3533 print/1532-4834 online  
DOI: 10.1080/01973533.2015.1012991



## Editorial

David Trafimow and Michael Marks

*New Mexico State University*

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that

Con Elías Moreno, un artículo de opinión en BEIO:

*"It is forbidden to use p-values!"*, BEIO, 31:2, Junio 2015.

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?
- Porque sigue siendo lo que usa la comunidad científica y los editores de revistas

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?
- Porque sigue siendo lo que usa la comunidad científica y los editores de revistas
- ¿Por qué sigue tanto gente usando  $p < 0,05$ ?

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?
- Porque sigue siendo lo que usa la comunidad científica y los editores de revistas
- ¿Por qué sigue tanto gente usando  $p < 0,05$ ?
- Porque es lo que se enseña en grados y másteres.

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?
- Porque sigue siendo lo que usa la comunidad científica y los editores de revistas
- ¿Por qué sigue tanto gente usando  $p < 0,05$ ?
- Porque es lo que se enseña en grados y másteres.

citado en *R. Wasserman A. & N. Lazar (2016), The American Statistician, 70:2 129-133*

# ¿Por qué enseñamos p-valores?

En febrero 2014, Prof. George Cobb en un foro de discusión de la ASA:

- ¿Por qué se enseña  $p < 0,05$  en tantos grados y másteres?
- Porque sigue siendo lo que usa la comunidad científica y los editores de revistas
- ¿Por qué sigue tanto gente usando  $p < 0,05$ ?
- Porque es lo que se enseña en grados y másteres.

citado en *R. Wasserman A. & N. Lazar (2016), The American Statistician, 70:2 129-133*

- Una reflexión necesaria en nuestra práctica docente.
- Surge también en nuestras colaboraciones con científicos aplicados.

¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance. Such nonsense is taught usually by professors who have had minimal contact with the applications of statistical methods to scientific problems. As a result the number of scientific papers which use statistical methods for window dressing is increasing. It appears that the  $P$  value next to a contingency table is beginning to mean what the "Seal of Good Housekeeping" means to the housewife.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance. Such nonsense is taught usually by professors who have had minimal contact with the applications of statistical methods to scientific problems. As a result the number of scientific papers which use statistical methods for window dressing is increasing. It appears that the  $P$  value next to a contingency table is beginning to mean what the "Seal of Good Housekeeping" means to the housewife.

# ¿Un debate de ahora?

En un seminario de Biométrica, junio 1965:

## FORMAL DISCUSSION

*Dr. Zelen:\**

I would like to congratulate the three speakers on their very interesting presentations. One of the common threads running through all the talks is that drawing a scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance. Such nonsense is taught usually by professors who have had minimal contact with the applications of statistical methods to scientific problems. As a result the number of scientific papers which use statistical methods for window dressing is increasing. It appears that the *P* value next to a contingency table is beginning to mean what the "Seal of Good Housekeeping" means to the housewife.

*Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. (1966).*

*"The Role of Hypothesis Testing in Clinical Trials," J Chron Disease, 19, 857-882*

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

- *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.*

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

- *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ziliak, S.T., and McCloskey, D.N. (2008)

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

- *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ziliak, S.T., and McCloskey, D.N. (2008)
- *Significance tests as sorcery: Science is empirical—significance tests are not.*

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

- *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ziliak, S.T., and McCloskey, D.N. (2008)
- *Significance tests as sorcery: Science is empirical—significance tests are not.* Lambdin (2012).

En realidad desde 1930, Ha dado lugar a mucho debate y artículos, algunos con títulos llamativos:

- *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ziliak, S.T., and McCloskey, D.N. (2008)
- *Significance tests as sorcery: Science is empirical—significance tests are not.* Lambdin (2012).

En el debate, intervienen estadísticos muy relevantes.

Se ha avivado en los últimos años. Unos hitos:

Se ha avivado en los últimos años. Unos hitos:

- El destierro de los p-valores de Basic Applied Psychology. 2015

Se ha avivado en los últimos años. Unos hitos:

- El destierro de los p-valores de Basic Applied Psychology. 2015
- Un artículo de opinión en Nature:

The screenshot shows the homepage of the journal 'nature'. At the top, there is a search bar with a 'Go' button and a link to 'Advanced search'. Below the header, a navigation menu includes 'Home', 'News & Comment', 'Research', 'Careers & Jobs', 'Current Issue', 'Archive', 'Audio & Video', and 'For Authors'. A breadcrumb trail indicates the article's path: 'Archive > Volume 506 > Issue 7487 > News Feature > Article'. The main content features a large image of a brain, followed by the title 'Scientific method: Statistical errors' in bold. Below the title, a subtitle reads: 'P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.' The author's name, 'Regina Nuzzo', is listed, along with the publication date, '12 February 2014'. On the right side of the article page, there are social sharing icons for E-alert, RSS, Facebook, and Twitter.

Se ha avivado en los últimos años. Unos hitos:

- El destierro de los p-valores de Basic Applied Psychology. 2015
- Un artículo de opinión en Nature:

The screenshot shows the homepage of the journal 'nature'. At the top, there is a search bar with a 'Go' button and a link to 'Advanced search'. Below the header, a navigation menu includes 'Home', 'News & Comment', 'Research', 'Careers & Jobs', 'Current Issue', 'Archive', 'Audio & Video', and 'For Authors'. A breadcrumb trail indicates the article's path: 'Archive > Volume 506 > Issue 7487 > News Feature > Article'. The main content features a large image of a brain, followed by the title 'Scientific method: Statistical errors' in bold. Below the title, a subtitle reads: 'P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.' The author's name, 'Regina Nuzzo', is listed, along with the date '12 February 2014'. To the right of the article, there are social sharing icons for E-alert, RSS, Facebook, and Twitter.

Uno de las más vistos en Nature de siempre.

Ha llevado la American Statistical Association a emitir en 2016:

Ha llevado la American Statistical Association a emitir en 2016:  
**ASA Statement on Statistical Significance and *P*-Values**

Ha llevado la American Statistical Association a emitir en 2016:  
**ASA Statement on Statistical Significance and P-Values**

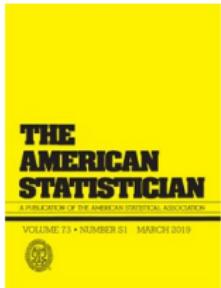


Ha llevado la American Statistical Association a emitir en 2016:  
**ASA Statement on Statistical Significance and P-Values**



- Movimiento casi sin precedente en la historia de la ASA.

Ha llevado la American Statistical Association a emitir en 2016:  
**ASA Statement on Statistical Significance and P-Values**



- Movimiento casi sin precedente en la historia de la ASA.
- Formaron un grupo de más de 20 expertos,
  - discusiones de varios meses,
  - un meeting de 2 días,
  - varios borradores hasta la aprobación del texto final.

Un debate de ahora.

A continuación organizaron en 2017

Un debate de ahora.

A continuación organizaron en 2017



ASA SYMPOSIUM ON  
STATISTICAL  
**INFERENCE**  
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond  $p < 0.05$

Un debate de ahora.

A continuación organizaron en 2017



ASA SYMPOSIUM ON  
STATISTICAL  
**INFERENCE**  
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond  $p < 0.05$

Asociado a este symposium un número especial de The American Statistician:

Un debate de ahora.

A continuación organizaron en 2017



ASA SYMPOSIUM ON  
STATISTICAL  
**INFERENCE**  
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond  $p < 0.05$

Asociado a este symposium un número especial de The American Statistician:



Statistical Inference in the 21st Century: A  
World Beyond  $p < 0.05$

<https://www.tandfonline.com/toc/utas20/73/sup1>

Un debate de ahora.

A continuación organizaron en 2017



ASA SYMPOSIUM ON  
STATISTICAL  
**INFERENCE**  
OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond  $p < 0.05$

Asociado a este symposium un número especial de The American Statistician:



Statistical Inference in the 21st Century: A  
World Beyond  $p < 0.05$

<https://www.tandfonline.com/toc/utas20/73/sup1>

Publicación online: 20 de marzo de 2019

Un debate de ahora.

A la misma vez, en Nature:

Un debate de ahora.

A la misma vez, en Nature:



COMMENT • 20 MARCH 2019

## Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein, Sander Greenland & Blake McShane



Mathieu Kessler

¿Están los p-valores condenados al destierro?

Un debate de ahora.

En este podcast recomendable:

Un debate de ahora.

En este podcast recomendable:

The image shows a screenshot of a podcast website for "NOT SO STANDARD DEVIATIONS". The top banner features the title "NOT SO STANDARD DEVIATIONS" in large white letters on a pink background, with a portrait of two hosts, Roger Peng and Hilary Parker, in the center. To the right, the text "with Roger Peng and Hilary Parker" is displayed. Below the banner is a black navigation bar with social media icons for Twitter, Email, and RSS feed. A navigation bar at the bottom of the page shows "All Episodes / 77 - Back to Statistics". The main content area displays the episode title "77 - Back to Statistics" in blue. Below it is a player interface with a play button, episode details ("NOT SO STANDARD DEVIATIONS", "77 - Back to Statistics"), and playback controls (30%, 00:00:00, 30%, volume, and file download). The date "Mar 28, 2019" is also visible. The bottom of the page contains a brief description of the episode content: "Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice."

NOT SO  
STANDARD  
DEVIATIONS

with  
Roger Peng  
and  
Hilary Parker

All Episodes / 77 - Back to Statistics

77 - Back to Statistics

NOT SO STANDARD DEVIATIONS  
77 - Back to Statistics

30% 00:00:00 30%

Mar 28, 2019

Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice.

Un debate de ahora.

En este podcast recomendable:

The screenshot shows a podcast player interface for 'NOT SO STANDARD DEVIATIONS' with the episode '77 - Back to Statistics'. The top banner features the show's name in large white letters on a pink background, and portraits of hosts Roger Peng and Hilary Parker. To the right, the text 'with Roger Peng and Hilary Parker' is displayed. Below the banner is a black navigation bar with social media icons for Twitter, Email, and RSS. The main content area shows the episode title '77 - Back to Statistics' and a play button. The episode details include the host names and the date 'Mar 28, 2019'. A summary text at the bottom states: 'Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice.'

NOT SO  
STANDARD  
DEVIATIONS

with  
Roger Peng  
and  
Hilary Parker

All Episodes / 77 - Back to Statistics

77 - Back to Statistics

NOT SO STANDARD DEVIATIONS  
77 - Back to Statistics

30s 00:00:00 30s

Mar 28, 2019

Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice.

Un debate de ahora.

En este podcast recomendable:

The screenshot shows the podcast player interface for episode 77 of "NOT SO STANDARD DEVIATIONS". The title "NOT SO STANDARD DEVIATIONS" is on the left, and the hosts' names "with Roger Peng and Hilary Parker" are on the right. Below the title is a black bar with social media icons for Twitter, Email, and RSS. A navigation bar at the bottom shows "All Episodes / 77 - Back to Statistics". The episode title "77 - Back to Statistics" is centered above the player. The player itself has a play button, the title "NOT SO STANDARD DEVIATIONS 77 - Back to Statistics", a progress bar showing 00:00:00, and various control icons. At the bottom left is the date "Mar 28, 2019". A summary text below the player states: "Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice."

NOT SO  
STANDARD  
DEVIATIONS

with  
Roger Peng  
and  
Hilary Parker

All Episodes / 77 - Back to Statistics

77 - Back to Statistics

NOT SO STANDARD DEVIATIONS  
77 - Back to Statistics

30s 00:00:00 30s

Mar 28, 2019

Hilary and Roger discuss the significance of statistical significance, the elements and principles of data analysis, and how to enforce ethical data science practice.

Un debate de ahora.

Y además, en cuanto a enseñanza

Un debate de ahora.

Y además, en cuanto a enseñanza

## United States Conference on Teaching Statistics 2019

Workshops begin May 14<sup>th</sup>, Conference May 16<sup>th</sup> - 18<sup>th</sup>, 2019  
Early-bird Registration prices end April 1



# ¿Por qué tanto ataque a los p-valores?

# ¿Por qué tanto ataque a los p-valores?

Se enmarca en la llamada

Crisis de la reproducibilidad

Se enmarca en la llamada

## Crisis de la reproducibilidad

A veces se distinguen dos términos:

- **Replicabilidad:** relacionado a las posibilidades de que un experimento independiente que busque contestar a la misma pregunta científica confirme los resultados.
- **Reproducibilidad:** capacidad de reproducir el análisis de datos a partir de los datos originales, y el conocimiento de flujo de manipulación y análisis.

# La crisis de la reproducibilidad

## Algunos ejemplos con impacto

2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

## Algunos ejemplos con impacto

2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

- Gran repercusión

## Algunos ejemplos con impacto

### 2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

- Gran repercusión
- Datos públicos, dos estadísticos Baggerly & Coombes intentan aplicar los algoritmos

## Algunos ejemplos con impacto

### 2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

- Gran repercusión
- Datos públicos, dos estadísticos Baggerly & Coombes intentan aplicar los algoritmos
- Encontraron muchos errores, desde triviales a muy graves

## Algunos ejemplos con impacto

### 2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

- Gran repercusión
- Datos públicos, dos estadísticos Baggerly & Coombes intentan aplicar los algoritmos
- Encontraron muchos errores, desde triviales a muy graves
- Fueron incluso capaces de reproducir el análisis erróneo

## Algunos ejemplos con impacto

### 2006: *Nature Medicine*

El equipo de A. Potti en Duke publican un algoritmo que predice, a partir de datos de chips ADN, qué pacientes de cáncer responderán a la quimioterapia.

- Gran repercusión
- Datos públicos, dos estadísticos Baggerly & Coombes intentan aplicar los algoritmos
- Encontraron muchos errores, desde triviales a muy graves
- Fueron incluso capaces de reproducir el análisis erróneo
- Fue retirado de la revista en 2011.

# La crisis de la reproducibilidad

The screenshot shows the top navigation bar of the Science magazine website. It includes links for Home, News, Journals, Topics, and Careers. Below the navigation is a red banner with the text "Science | AAAS Love science? We deliver." and images of Science magazine covers. To the right of the banner is an offer for "Get 50 issues of Science and a free T-shirt when you join." At the bottom right of the header are "Log in" and "My account" links.

## SHARE



Anil Potti, DUKE UNIVERSITY

## Potti found guilty of research misconduct

By Jocelyn Kaiser | Nov. 9, 2015, 12:30 PM

Eight years after questions were first raised about the work of Duke University cancer researcher Anil Potti, federal officials have found Potti guilty of research misconduct. The findings bring to a close one of the most egregious U.S. scientific misconduct cases in recent years.

In 2006 Potti's team published several papers in high-profile journals reporting that certain gene expression signatures predicted a patient's response to chemotherapy. Two outside biostatisticians soon raised concerns about the studies. In 2010, Duke put Potti on administrative leave and **suspended three clinical trials based on his work** after *The Cancer Letter*, a newsletter in Washington, D.C., reported that Potti had padded his resume. Potti resigned a few months later.

Many of Potti's papers were later retracted, and Duke faced a **lawsuit filed by patients in the clinical trials**. His troubles also led to an Institute of Medicine report that faulted Duke's oversight and found **broad problems in the cancer field with using gene signatures and other biomarkers to**

<https://www.sciencemag.org/news/2015/11/potti-found-guilty-research-misconduct>

# La crisis de la reproducibilidad

Science Home News Journals Topics Careers

Science | AAAS Love science? We deliver.

Get 50 issues of Science and a free T-shirt when you join.

Log in | My account

## SHARE



29



Anil Potti, DUKE UNIVERSITY

## Potti found guilty of research misconduct

By Jocelyn Kaiser | Nov. 9, 2015, 12:30 PM

Eight years after questions were first raised about the work of Duke University cancer researcher Anil Potti, **federal officials have found Potti guilty of research misconduct**. The findings bring to a close one of the most egregious U.S. scientific misconduct cases in recent years.

In 2006 Potti's team published several papers in high-profile journals reporting that certain gene expression signatures predicted a patient's response to chemotherapy. Two outside biostatisticians soon raised concerns about the studies. In 2010, Duke put Potti on administrative leave and **suspended three clinical trials based on his work** after *The Cancer Letter*, a newsletter in Washington, D.C., reported that Potti had padded his resume. Potti resigned a few months later.

Many of Potti's papers were later retracted, and Duke faced a lawsuit filed by patients in the clinical trials. His troubles also led to an Institute of Medicine report that faulted Duke's oversight and found broad problems in the cancer field with using gene signatures and other biomarkers to

<https://www.sciencemag.org/news/2015/11/potti-found-guilty-research-misconduct>

# La crisis de la reproducibilidad

Science Home News Journals Topics Careers

Science | AAAS Love science? We deliver.

Get 50 issues of Science and a free T-shirt when you join.

Log in | My account

SHARE



29



Anil Potti, DUKE UNIVERSITY

## Potti found guilty of research misconduct

By Jocelyn Kaiser | Nov. 9, 2015, 12:30 PM

federal officials have found Potti guilty of research misconduct.

use one of the most egregious that additional information about the study.

In 2006 Potti's team published several papers in high-profile journals reporting that certain gene expression signatures predicted a patient's response to chemotherapy. Two outside biostatisticians soon raised concerns about the studies. In 2010, Duke put Potti on administrative leave and suspended three clinical trials based on his work after *The Cancer Letter*, a newsletter in Washington, D.C., reported that Potti had padded his resume. Potti resigned a few months later.

Many of Potti's papers were later retracted, and Duke faced a lawsuit filed by patients in the clinical trials. His troubles also led to an Institute of Medicine report that faulted Duke's oversight and found broad problems in the cancer field with using gene signatures and other biomarkers to

<https://www.sciencemag.org/news/2015/11/potti-found-guilty-research-misconduct>

# La crisis de la reproducibilidad

Science Home News Journals Topics Careers

Science | AAAS Love science? We deliver.

Get 50 issues of Science and a free T-shirt when you join.

Log in | My account

SHARE



29



Anil Potti, DUKE UNIVERSITY

## Potti found guilty of research misconduct

By Jocelyn Kaiser | Nov. 9, 2015, 12:30 PM

Eight years after questions were first raised about the work of Duke University cancer researcher Anil Potti, **federal officials have found Potti guilty of research misconduct**. The findings bring to a close one of the most egregious U.S. scientific misconduct cases in recent years.

In 2006 Potti's team published several papers in high-profile journals reporting that certain gene expression signatures predicted a patient's response to chemotherapy. Two outside biostatisticians soon raised concerns about the studies. In 2010, Duke put Potti on administrative leave and **suspended three clinical trials based on his work** after *The Cancer Letter*, a newsletter in Washington, D.C., reported that Potti had padded his resume. Potti resigned a few months later.

**Many of Potti's papers were later retracted,** by patients in the with whom he worked. The U.S. Department of Justice has been investigating, and enforcement actions were initiated. Duke's oversight and found broad problems in the cancer field with using gene signatures and other biomarkers to

<https://www.sciencemag.org/news/2015/11/potti-found-guilty-research-misconduct>

# La crisis de la reproducibilidad

Science Home News Journals Topics Careers

Science | AAAS Love science? We deliver.

Get 50 issues of Science and a free T-shirt when you join.

Log in | My account

SHARE



29



Anil Potti, DUKE UNIVERSITY

## Potti found guilty of research misconduct

By Jocelyn Kaiser | Nov. 9, 2015, 12:30 PM

Eight years after questions were first raised about the work of Duke University cancer researcher Anil Potti, **federal officials have found Potti guilty of research misconduct**. The findings bring to a close one of the most egregious U.S. scientific misconduct cases in recent years.

In 2006 Potti's team published several papers in high-profile journals reporting that certain gene expression signatures predicted a patient's response to chemotherapy. Two outside biostatisticians soon raised concerns about the studies. In 2010, Duke put Potti on administrative leave and **suspended three clinical trials based on his work** after *The Cancer Letter*, a newsletter in Washington, D.C., reported that Potti had padded his resume. Potti resigned a few months later.

**Duke faced a lawsuit filed by patients in the clinical trials.** In the **September issue**, *Cancer Research* published an **internal report** that faulted Duke's oversight and found broad problems in the cancer field with using gene signatures and other biomarkers to

<https://www.sciencemag.org/news/2015/11/potti-found-guilty-research-misconduct>

COLUMNAS

## *La depresión del Excel*

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

COLUMNAS

## *La depresión del Excel*

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

*2010, the National Bureau of Economic Research*

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”.

COLUMNAS

## La depresión del Excel

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

### 2010, the National Bureau of Economic Research

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”. Aseguraron que países con *Deuda/PIB* altos sufrían bajo crecimiento, identificando incluso el umbral 90 %.

COLUMNAS

## La depresión del Excel

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

### 2010, the National Bureau of Economic Research

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”. Aseguraron que países con *Deuda/PIB* altos sufrían bajo crecimiento, identificando incluso el umbral 90 %.

- Inmediatamente críticas: por una parte por llevar a confundir asociación y causalidad.

COLUMNAS

## La depresión del Excel

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

### 2010, the National Bureau of Economic Research

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”. Aseguraron que países con *Deuda/PIB* altos sufrían bajo crecimiento, identificando incluso el umbral 90 %.

- Inmediatamente críticas: por una parte por llevar a confundir asociación y causalidad.
- Porque otros, partiendo de datos similares, no pudieron reproducir los resultados.

COLUMNAS

## La depresión del Excel

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

### 2010, the National Bureau of Economic Research

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”. Aseguraron que países con *Deuda/PIB* altos sufrían bajo crecimiento, identificando incluso el umbral 90 %.

- Inmediatamente críticas: por una parte por llevar a confundir asociación y causalidad.
- Porque otros, partiendo de datos similares, no pudieron reproducir los resultados.
- Por la presión, publicaron el Excel.

COLUMNAS

## La depresión del Excel

¿Puede un error en una hoja de cálculo haber destruido casi por completo la economía de Occidente?



PAUL KRUGMAN

21 ABR 2013 - 00:01 CEST

### 2010, the National Bureau of Economic Research

2010, Carmen Reinhart & Kenneth Rogoff, Harvard, publican “*Growth in a time of debt*”. Aseguraron que países con *Deuda/PIB* altos sufrían bajo crecimiento, identificando incluso el umbral 90 %.

- Inmediatamente críticas: por una parte por llevar a confundir asociación y causalidad.
- Porque otros, partiendo de datos similares, no pudieron reproducir los resultados.
- Por la presión, publicaron el Excel.
- Thomas Herndon, estudiante de grado economía, encontró varios errores: un error fórmula Excel, ponderaciones no estándar.

En *Science*, 2015, 270 autores Open Science Collaboration:

En *Science*, 2015, 270 autores Open Science Collaboration:  
**Estimating the reproducibility of psychological science.**  
Prueban a replicar 100 experimentos en 3 revistas de psicología.

En *Science*, 2015, 270 autores Open Science Collaboration:

## **Estimating the reproducibility of psychological science.**

Prueban a replicar 100 experimentos en 3 revistas de psicología.

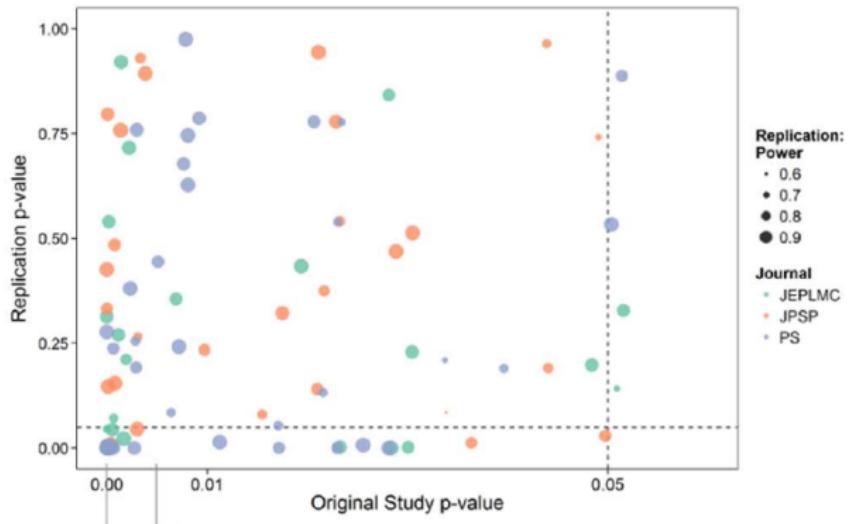
Originals: 97 % estad. significativos. Réplicas: 36 % estad. significativos.

En *Science*, 2015, 270 autores Open Science Collaboration:

## **Estimating the reproducibility of psychological science.**

Prueban a replicar 100 experimentos en 3 revistas de psicología.

Originals: 97 % estad. significativos. Réplicas: 36 % estad. significativos.



## Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem  
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

## Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem  
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

### Extra Sensory Perception (ESP)

Nueve experimentos.

## Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem  
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

### Extra Sensory Perception (ESP)

Nueve experimentos.

E.g el experimento 1:

- los participantes tenían que adivinar si una imagen iba a aparecer a la izquierda o derecha de la pantalla.

## Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem  
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

### Extra Sensory Perception (ESP)

Nueve experimentos.

E.g el experimento 1:

- los participantes tenían que adivinar si una imagen iba a aparecer a la izquierda o derecha de la pantalla.
- Bem encuentra que, si se trata de imágenes eróticas, aciertan con más frecuencia que si fuera al azar.

## Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem  
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

## Correcting the Past: Failures to Replicate Psi

Jeff Galak  
Carnegie Mellon University

Robyn A. LeBoeuf  
University of Florida

Leif D. Nelson  
University of California, Berkeley

Joseph P. Simmons  
University of Pennsylvania

Across 7 experiments ( $N = 3,289$ ), we replicate the procedure of Experiments 8 and 9 from Bem (2011), which had originally demonstrated retroactive facilitation of recall. We failed to replicate that finding.

# La crisis de la reproducibilidad

¿Son casos aislados?

# La crisis de la reproducibilidad

¿Son casos aislados?

The screenshot shows the homepage of the journal *nature*. At the top, the URL "nature.com > nature > special" is visible, along with the journal's logo and the text "a natureresearch journal". The main title "Challenges in irreproducible research" is displayed, with a subtitle "Special | 06 July 2018". Below the main content area, there are navigation links: "Key reads", "Latest", "All articles", and "Research & Reviews". A quote at the bottom states: "Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study."

# La crisis de la reproducibilidad

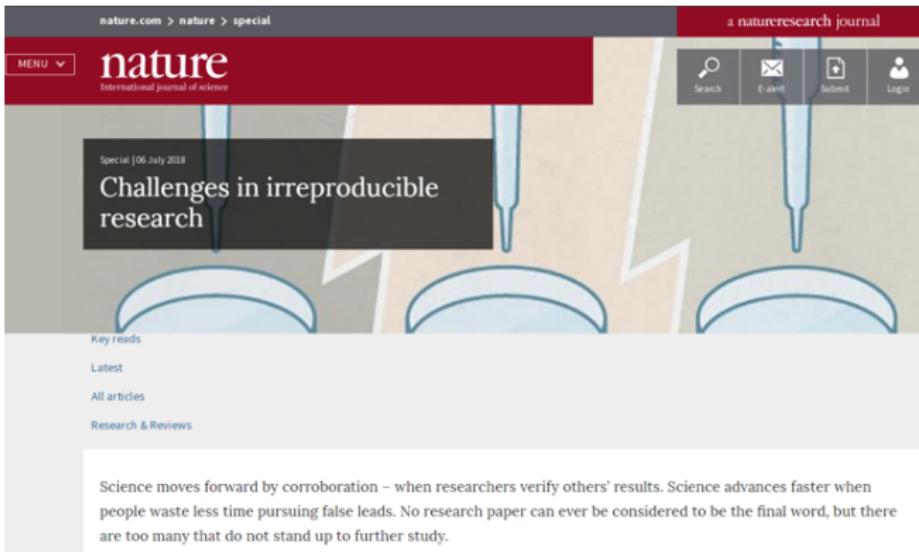
¿Son casos aislados?

The screenshot shows the homepage of the journal 'nature' (international journal of science). At the top, there's a red header bar with the word 'nature' and its subtitle 'international journal of science'. Below this, a dark banner reads 'Special | 06 July 2018 Challenges in irreproducible research'. To the right of the banner, there are icons for 'Search', 'E-alert', 'Submit', and 'Login'. On the left side of the main content area, there are four categories: 'Key reads', 'Latest', 'All articles', and 'Research & Reviews'. A quote at the bottom states: 'Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study.'

Nature realiza una encuesta a 1500 investigadores:

# La crisis de la reproducibilidad

¿Son casos aislados?

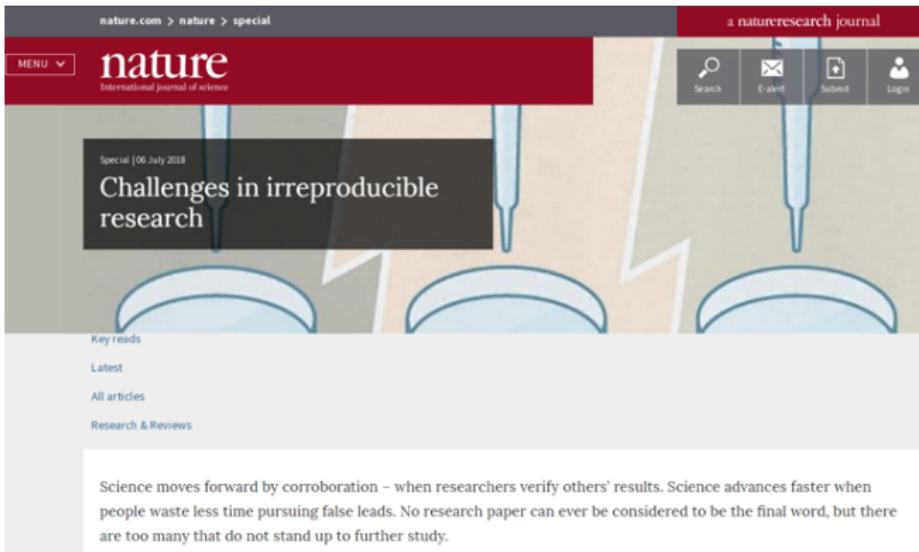


Nature realiza una encuesta a 1500 investigadores:

- El 70 % no consiguió en alguna ocasión replicar experimentos de otros.
- Más del 50 % no consiguió en alguna ocasión replicar sus propios experimentos.

# La crisis de la reproducibilidad

¿Son casos aislados?



Identifican tres ámbitos

Experimental, computacional, estadístico

# La crisis de la reproducibilidad

Una manera poderosa de reducir la irreproducibilidad es fomentar la ciencia abierta

# La crisis de la reproducibilidad

Una manera poderosa de reducir la irreproducibilidad es fomentar la ciencia abierta

# Science as an open enterprise

June 2012



<https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/>

# La crisis de la reproducibilidad

Una manera poderosa de reducir la irreproducibilidad es fomentar la ciencia abierta

# Science as an open enterprise

June 2012



<https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/>

Open Science = Open Data +  
Open Access to scientific publications +  
effective communication.

# La crisis de la reproducibilidad

Commission and its priorities

Policies, information and services



English EN

Search

Search

Home > Research and Innovation > Strategy > Goals of research and innovation policy >

## Open Science

This is the ongoing transition in how research is performed and how knowledge is shared. News, events, publications related to Open Science

Home

Open Access

European Open Science Cloud ▾

Open Science Policy Platform ▾

Groups ▾

Open Science Monitor

**EU could save €10.2 billion per year by using FAIR data.**

**Which funding and business models can make FAIR data sustainable?**

The European Commission has published two reports based on the study "Cost-benefit analysis of FAIR research data", which was conducted for the Commission by PricewaterhouseCoopers.

The report [Cost of not having FAIR research data](#) aims to provide an estimate for the EU economy based on a series of indicators extracted from previous studies and analysed via interviews with subject matter experts. Using quantitative methodology and very conservative assumptions, the analysis shows that the minimum cost for the EU is €10.2 billion per year, which will increase over the years if we do not take action.

Events

Focus on ...

3 April 2019, Ljubljana, Slovenia - [Boosting Innovation for EU industry: Industrial Infrastructures and Open Innovation Ecosystems](#)

Other events

3 April 2019, Ljubljana, Slovenia - [Boosting Innovation for EU industry: Industrial Infrastructures and Open Innovation Ecosystems](#)

8 April 2019, Geneva, Switzerland - [ARCHIVER Open](#)

<https://ec.europa.eu/research/openscience/index.cfm>

# La crisis de la reproducibilidad

Commission and its priorities Policies, information and services

 European Commission | English EN Search Search

Home > Research and Innovation > Strategy > Goals of research and innovation policy >

## Open Science

This is the ongoing transition in how research is performed and how knowledge is shared. News, events, publications related to Open Science

Home Open Access European Open Science Cloud ▾ Open Science Policy Platform ▾ Groups ▾ Open Science Monitor

**EU could save €10.2 billion per year by using FAIR data.  
Which funding and business models can make FAIR data sustainable?**

The European Commission has published two reports based on the study "Cost-benefit analysis of FAIR research data", which was conducted for the Commission by PricewaterhouseCoopers.

The report [Cost of not having FAIR research data](#) aims to provide an estimate for the EU economy based on a series of indicators extracted from previous studies and analysed via interviews with subject matter experts. Using quantitative methodology and very conservative assumptions, the analysis shows that the minimum cost for the EU is €10.2 billion per year, which will increase over the years if we do not take action.

### Events

Focus on ...

3 April 2019, Ljubljana, Slovenia - [Boosting Innovation for EU industry: Industrial Infrastructures and Open Innovation Ecosystems](#)

### Other events

3 April 2019, Ljubljana, Slovenia - [Boosting Innovation for EU industry: Industrial Infrastructures and Open Innovation Ecosystems](#)

8 April 2019, Geneva, Switzerland - [ARCHIVER Open](#)

<https://ec.europa.eu/research/openscience/index.cfm>

**FAIR = Findable + Accesible + Interoperable + Reusable**

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

por el acrónimo...

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

por el acrónimo...

Citado en Nuzzo (2014) [enlace](#)

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

por el acrónimo...

Citado en Nuzzo (2014) [enlace](#)

Se ha establecido

*Statistical Significance*  $\Leftrightarrow$  *Scientific Significance*.

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

por el acrónimo...

Citado en Nuzzo (2014) [enlace](#)

Se ha establecido

*Statistical Significance*  $\Leftrightarrow$  *Scientific Significance*.

*Statistical Significance*  $\Leftrightarrow p < 0,05$

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Qué parte de culpa tienen?

Los *p*-valores llevan muchos años en la piqueta.

Un investigador incluso propuso cambiarle el nombre a

Statistical Hypothesis Inference Testing

por el acrónimo...

Citado en Nuzzo (2014) [enlace](#)

Se ha establecido

*Statistical Significance*  $\Leftrightarrow$  *Scientific Significance*.

*Statistical Significance*  $\Leftrightarrow p < 0,05$

*Scientific Significance*  $\Leftrightarrow p < 0,05$

¿Qué papel juegan los *p*-valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

## ¿Qué papel juegan los $p$ -valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . *Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance*

Fisher 1926b, p. 504, italics supplied.

## ¿Qué papel juegan los *p*-valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

It is convenient to **draw the line** at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . *Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance*

Fisher 1926b, p. 504, italics supplied.

## ¿Qué papel juegan los *p*-valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

It is convenient to **draw the line** at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . **Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.** A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance

Fisher 1926b, p. 504, italics supplied.

## ¿Qué papel juegan los *p*-valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

It is convenient to **draw the line** at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . **Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.**

Fisher 1926b, p. 504, italics supplied.

## ¿Qué papel juegan los *p*-valores en esta crisis?

Por cierto, ¿dónde aparece  $p < 0,05$ ?

It is convenient to **draw the line** at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point) . . . **Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.**

Fisher 1926b, p. 504, italics supplied.

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to *eliminate from further discussion* the greater part of the fluctuations which chance causes have introduced into their experimental results

Fisher 1935 [1960], p. 13, italics supplied.

## ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

## ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.
- 51 % de los abstracts contienen un *p*-valor.

## ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.
- 51 % de los abstracts contienen un *p*-valor.
- Entre los que contienen un *p*-valor, el 96 % tienen  $p < 0,05$ .

## ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.
- 51 % de los abstracts contienen un *p*-valor.
- Entre los que contienen un *p*-valor, el 96 % tienen  $p < 0,05$ .
- Entre 1990 y 2015, la presencia de *p*-valores se ha multiplicado por dos en los abstracts analizados.

# ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.
- 51 % de los abstracts contienen un *p*-valor.
- Entre los que contienen un *p*-valor, el 96 % tienen  $p < 0,05$ .
- Entre 1990 y 2015, la presencia de *p*-valores se ha multiplicado por dos en los abstracts analizados.

*"Too good to be true?"*

## ¿Qué papel juegan los *p*-valores en esta crisis?

¿Están efectivamente muy usados en la investigación científica?

John Ioannidis, Stanford, y colaboradores han estudiado en muchos artículos la presencia de *p*-valores a través de minería de textos en las bases de datos bibliográficas

En su [contribución](#) al volumen especial de *The American Statistician*, 2019:

- Analizan 13 millones de abstracts y 844 000 artículos completos de PubMed Central.
- 51 % de los abstracts contienen un *p*-valor.
- Entre los que contienen un *p*-valor, el 96 % tienen  $p < 0,05$ .
- Entre 1990 y 2015, la presencia de *p*-valores se ha multiplicado por dos en los abstracts analizados.

*"Too good to be true?"*

En 2005: Ioannidis JPA (2005) "Why Most Published Research Findings Are False." *PLOS Medicine* 2(8): e124.

## ¿Qué papel juegan los *p*-valores en esta crisis?

Vamos a hablar de tres problemas, que pueden jugar un papel.



# ¿Qué papel juegan los $p$ -valores en esta crisis?

Vamos a hablar de tres problemas, que pueden jugar un papel.

- ① Las dificultades de interpretación del  $p$ -valor propician la confusión  $p < 0,05 \Leftrightarrow \text{Scientific significance}$ .

# ¿Qué papel juegan los $p$ -valores en esta crisis?

Vamos a hablar de tres problemas, que pueden jugar un papel.

- ① Las dificultades de interpretación del  $p$ -valor propician la confusión  $p < 0,05 \Leftrightarrow$  *Scientific significance*.
- ② La elección del umbral  $p = 0,05$ .

# ¿Qué papel juegan los $p$ -valores en esta crisis?

Vamos a hablar de tres problemas, que pueden jugar un papel.

- ① Las dificultades de interpretación del  $p$ -valor propician la confusión  $p < 0,05 \Leftrightarrow$  *Scientific significance*.
- ② La elección del umbral  $p = 0,05$ .
- ③ La existencia del “ $p$ -hacking”.

Un hecho innegable:

### El *p*-valor se presta a interpretaciones erróneas

En, por ejemplo,

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133. [enlace](#)
- Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)
- Greenland, S., Senn, S.J., Rothman, K.J. et al.(2016) “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations”. *Eur J Epidemiol* 31: 337. [enlace](#).

se enfatiza lo que **NO** es un *p*-valor.

## Primer problema: interpretación del $p$ -valor

El  $p$ -valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

## Primer problema: interpretación del $p$ -valor

El  $p$ -valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El  $p$ -valor sí es una medida de la compatibilidad de los datos observados con el modelo completo

## Primer problema: interpretación del *p*-valor

El *p*-valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El *p*-valor sí es una medida de la compatibilidad de los datos observados con el modelo **completo**

## Primer problema: interpretación del *p*-valor

El *p*-valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El *p*-valor sí es una medida de la compatibilidad de los datos observados con el modelo **completo**

Por lo tanto:

## Primer problema: interpretación del *p*-valor

El *p*-valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El *p*-valor sí es una medida de la compatibilidad de los datos observados con el modelo **completo**

Por lo tanto:

- El *p*-valor no es un indicador sólamente de la evidencia presente en los datos sobre si  $H_0$  es verdadera, sino que depende evidentemente de todas las hipótesis realizadas en la construcción del modelo.

## Primer problema: interpretación del $p$ -valor

El  $p$ -valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El  $p$ -valor sí es una medida de la compatibilidad de los datos observados con el modelo **completo**

Por lo tanto:

- El  $p$ -valor no es un indicador sólamente de la evidencia presente en los datos sobre si  $H_0$  es verdadera, sino que depende evidentemente de todas las hipótesis realizadas en la construcción del modelo.
- En esta compatibilidad de los datos observados con el modelo completo, entra todo.

## Primer problema: interpretación del $p$ -valor

El  $p$ -valor no es la probabilidad de que la hipótesis nula sea cierta basado en mis datos observados!!!!

El  $p$ -valor sí es una medida de la compatibilidad de los datos observados con el modelo **completo**

Por lo tanto:

- El  $p$ -valor no es un indicador sólamente de la evidencia presente en los datos sobre si  $H_0$  es verdadera, sino que depende evidentemente de todas las hipótesis realizadas en la construcción del modelo.
- En esta compatibilidad de los datos observados con el modelo completo, entra todo.
- Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Un hecho innegable:

**El *p*-valor se presta a interpretaciones erróneas**

El artículo pedagógico

- Greenland, S., Senn, S.J., Rothman, K.J. et al.(2016) "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". *Eur J Epidemiol* 31: 337. [enlace](#).

Un hecho innegable:

**El *p*-valor se presta a interpretaciones erróneas**

El artículo pedagógico

- Greenland, S., Senn, S.J., Rothman, K.J. et al.(2016) "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". *Eur J Epidemiol* 31: 337. [enlace](#).

<https://kahoot.it/>

## Primer problema: interpretación del *p*-valor

Cualquier valor del efecto que lleva a un *p*-valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un *p*-valor superior o igual a 0.05.

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$p(\mu_0) = 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0))$$

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$p(\mu_0) = 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0))$$

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$p(\mu_0) = 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0)) \geq 0,05$$

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$\begin{aligned} p(\mu_0) &= 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0)) \geq 0,05 \\ &\Leftrightarrow \mathbb{P}(Z \leq \sqrt{n}|\bar{X} - \mu_0|) \leq 0,975, \end{aligned}$$

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$\begin{aligned} p(\mu_0) &= 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0)) \geq 0,05 \\ &\Leftrightarrow \mathbb{P}(Z \leq \sqrt{n}|\bar{X} - \mu_0|) \leq 0,975, \\ &\Leftrightarrow \sqrt{n}|\bar{X} - \mu_0| \leq z_{0,975}. \end{aligned}$$

## Primer problema: interpretación del $p$ -valor

Cualquier valor del efecto que lleva a un  $p$ -valor mayor que el observado, es más compatible con los datos observados.

Construcción de un intervalo de confianza:

El intervalo de confianza al 95 % es el conjunto de valores del efecto que tienen asociado un  $p$ -valor superior o igual a 0.05.

Ejemplo:  $X_1, X_2 \dots, X_n$  i.i.d  $\mathcal{N}(\mu, 1)$ .

Planteamos

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0, \end{cases}$$

Efecto:  $\mu_0$ . El  $p$ -valor es

$$\begin{aligned} p(\mu_0) &= 2\mathbb{P}(Z \geq \sqrt{n}(\bar{X} - \mu_0)) \geq 0,05 \\ &\Leftrightarrow \mathbb{P}(Z \leq \sqrt{n}|\bar{X} - \mu_0|) \leq 0,975, \\ &\Leftrightarrow \sqrt{n}|\bar{X} - \mu_0| \leq z_{0,975}. \end{aligned}$$

**“Intervalos de compatibilidad”**

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas.

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .
- Es un problema sociológico.

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .
- Es un problema sociológico.

*The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them.*

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .
- Es un problema sociológico.

*The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion.*

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .
- Es un problema sociológico.

*The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the p-value doesn't mean what people think it means; it becomes valuable because of what it buys.*

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Los métodos científicos dentro de las disciplinas (oncología, física nuclear, genética, cardiología, producción animal, etc..) se enseñan y se transmiten dentro de estas mismas disciplinas. ⇒ SILOS .
- Es un problema sociológico.

*The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the p-value doesn't mean what people think it means; it becomes valuable because of what it buys.*

- (...) the use of statistics tests has become obligatory in scientific research. (...) they work mainly as **social technologies**, not as guides to private thinking.

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician*):

- Es una respuesta a un problema de confianza.

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Es una respuesta a un problema de confianza.

*The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations.*

## Primer problema: interpretación del *p*-valor

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Es una respuesta a un problema de confianza.

*The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations.*

- No será fácil cambiarlo

## Primer problema: interpretación del *p*-valor

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Es una respuesta a un problema de confianza.

*The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations.*

- No será fácil cambiarlo

*P-values are part of a rule-based structure that serves as a bulwark against claims of expertise untethered from empirical support. It can be changed, but we must respect the reason why the statistical procedures are there in the first place.*

## Primer problema: interpretación del *p*-valor

¿Qué solución podemos darle?

Formación, concienciación, pensamiento crítico?...

Debemos ser conscientes (Goodman, 2019, *The American Statistician* ):

- Es una respuesta a un problema de confianza.

*The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations.*

- No será fácil cambiarlo

*P-values are part of a rule-based structure that serves as a bulwark against claims of expertise untethered from empirical support. It can be changed, but we must respect the reason why the statistical procedures are there in the first place. This partly explains why there is so much resistance to Bayesian approaches, which are often viewed as a back-door way to reintroduce the subjectivity that conventional statistical methods were introduced to counter.*

Una clara recomendación de

- The American Statistical Association en Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

Una clara recomendación de

- The American Statistical Association en Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

No usemos nunca más el término  
“Estadísticamente significativo”

Una clara recomendación de

- The American Statistical Association en Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

No usemos nunca más el término  
“Estadísticamente significativo”

Y ninguna de sus variantes: “significativamente diferente”, “ $p < 0,05$ ”, “no significativo”, en palabras o con estrellitas en una tabla.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	HIGHLY SIGNIFICANT
0.02	HIGHLY SIGNIFICANT
0.03	HIGHLY SIGNIFICANT
0.04	SIGNIFICANT
0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	ON THE EDGE OF SIGNIFICANCE
0.07	HIGHLY SUGGESTIVE,
0.08	SIGNIFICANT AT THE $p < 0.10$ LEVEL
0.09	SIGNIFICANT AT THE $p < 0.10$ LEVEL
0.099	HEY, LOOK AT THIS INTERESTING
$\geq 0.1$	SUBGROUP ANALYSIS

<https://xkcd.com/1478/>

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Desde inicio siglo XX, los científicos han ido adoptado un híbrido entre el enfoque de Fisher y el de Neyman & Pearson (teoría decisión).

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Desde inicio siglo XX, los científicos han ido adoptado un híbrido entre el enfoque de Fisher y el de Neyman & Pearson (teoría decisión).  
↔ simplicidad del  $p$ -valor y (falsa) seguridad de la teoría de decisión, error, riesgo, etc...

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Desde inicio siglo XX, los científicos han ido adoptado un híbrido entre el enfoque de Fisher y el de Neyman & Pearson (teoría decisión).  
⇒ simplicidad del  $p$ -valor y (falsa) seguridad de la teoría de decisión, error, riesgo, etc...
- Hay disciplinas en las que el umbral estándar es diferente:
  - Genética:  $p < 5 \times 10^{-8}$  (para “genome-wide association studies”)
  - Física de altas energías:  $p < 3 \times 10^{-7}$  (“ $3\sigma$ ”)

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Desde inicio siglo XX, los científicos han ido adoptado un híbrido entre el enfoque de Fisher y el de Neyman & Pearson (teoría decisión).  
⇒ simplicidad del  $p$ -valor y (falsa) seguridad de la teoría de decisión, error, riesgo, etc...
- Hay disciplinas en las que el umbral estándar es diferente:
  - Genética:  $p < 5 \times 10^{-8}$  (para “genome-wide association studies”)
  - Física de altas energías:  $p < 3 \times 10^{-7}$  (“ $3\sigma$ ”)

Es evidente que el valor del umbral debería tener en cuenta el contexto (tamaño muestral etc..)

Segundo problema: es el umbral  $p = 0,05$  adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories  
call for an end to hyped claims and the dismissal of possibly crucial effects.

Segundo problema: es el umbral  $p = 0,05$  adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories  
call for an end to hyped claims and the dismissal of possibly crucial effects.

Dos estudios de análisis de efectos secundarios de medicamentos  
anti-inflamatorios:

- ① El segundo equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [3 %, 48 %], ( $p = 0,091$ ).

Segundo problema: es el umbral  $p = 0,05$  adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories  
call for an end to hyped claims and the dismissal of possibly crucial effects.

Dos estudios de análisis de efectos secundarios de medicamentos  
anti-inflamatorios:

- ① El segundo equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [3 %, 48 %], ( $p = 0,091$ ).

↔ “La toma de estos medicamentos no está asociado con fibrilación  
auricular” .

Segundo problema: es el umbral  $p = 0,05$  adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories  
call for an end to hyped claims and the dismissal of possibly crucial effects.

Dos estudios de análisis de efectos secundarios de medicamentos  
anti-inflamatorios:

- ① El segundo equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [3 %, 48 %], ( $p = 0,091$ ).

↔ “La toma de estos medicamentos no está asociado con fibrilación  
auricular”.

lo que contradice:

Segundo problema: es el umbral  $p = 0,05$  adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories  
call for an end to hyped claims and the dismissal of possibly crucial effects.

Dos estudios de análisis de efectos secundarios de medicamentos anti-inflamatorios:

- ① El segundo equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [3 %, 48 %], ( $p = 0,091$ ).

↔ “La toma de estos medicamentos no está asociado con fibrilación auricular”.

lo que contradice:

- ② El primer equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [9 %, 33 %], ( $p = 0,0003$ ).

## Segundo problema: es el umbral $p = 0,05$ adecuado?

En *Nature*, 567, 305-307, Marzo 2019:

# Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Dos estudios de análisis de efectos secundarios de medicamentos anti-inflamatorios:

- ① El segundo equipo:

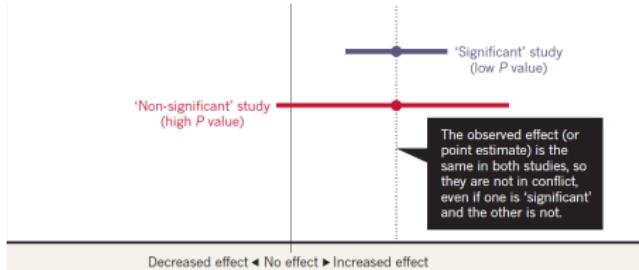
factor de riesgo de 1.2 (20 % más alto), IC: [3 %, 48 %], ( $p = 0,091$ ).

↔ “La toma de estos medicamentos no está asociado con fibrilación auricular”.

lo que contradice:

- ② El primer equipo:

factor de riesgo de 1.2 (20 % más alto), IC: [9 %, 33 %], ( $p = 0,0003$ ).



## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.
- ⇒ hay que huir de la dicotomía basado en valor de  $p$ .

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.
- ⇒ hay que huir de la dicotomía basado en valor de  $p$ .

Es particularmente importante plantearse qué tamaño del efecto es relevante del punto de vista práctico.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.
- ⇒ hay que huir de la dicotomía basado en valor de  $p$ .

Es particularmente importante plantearse qué tamaño del efecto es relevante del punto de vista práctico.

Es decir, hay que tener el contexto en cuenta.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.
- ⇒ hay que huir de la dicotomía basado en valor de  $p$ .

Es particularmente importante plantearse qué tamaño del efecto es relevante del punto de vista práctico.

Es decir, hay que tener el contexto en cuenta.

- Rebecca A. Betensky (2019) The p-Value Requires Context, Not a Threshold, *The American Statistician*, 73:sup1, 115-117, [enlace](#)

## Segundo problema: es el umbral $p = 0,05$ adecuado?

- Es importante no limitarse a basar la decisión sobre significación científica sobre el único  $p$ -valor.
- ⇒ hay que acompañarlo de estimación del efecto, con margen de error.
- ⇒ hay que huir de la dicotomía basado en valor de  $p$ .

Es particularmente importante plantearse qué tamaño del efecto es relevante del punto de vista práctico.

Es decir, hay que tener el contexto en cuenta.

- Rebecca A. Betensky (2019) The p-Value Requires Context, Not a Threshold, The American Statistician, 73:sup1, 115-117, [enlace](#)

El contexto = tamaño relevante del efecto del punto de vista práctico + tamaño muestral (diseño del experimento).

Segundo problema: es el umbral  $p = 0,05$  adecuado?

Esto tampoco es nuevo...

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Esto tampoco es nuevo...

My original question and its modified form. When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery"], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment. This is one of the points on which I should like advice.

Gosset, c. April 1905, in E. S. Pearson 1939, pp. 215-216; italics supplied.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Esto tampoco es nuevo...

My original question and its modified form. When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery"], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment. This is one of the points on which I should like advice.

Gosset, c. April 1905, in E. S. Pearson 1939, pp. 215-216; italics supplied.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Esto tampoco es nuevo...

My original question and its modified form. When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery"], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment. This is one of the points on which I should like advice.

Gosset, c. April 1905, in E. S. Pearson 1939, pp. 215-216; italics supplied.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

### Simulación sencilla

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

### Simulación sencilla

- Contexto: ensayos medicamentosos.  $D_1, D_2, \dots, D_N$  situaciones similares de medicamentos y sus ensayos.

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

### Simulación sencilla

- Contexto: ensayos medicamentosos.  $D_1, D_2, \dots, D_N$  situaciones similares de medicamentos y sus ensayos.
- Se asume  $\mathbb{P}(H_0) = 0,5$ . Se simula para cada  $D_i$  la hipótesis, habiendo escogido el tamaño relevante del efecto si  $H_1$

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

### Simulación sencilla

- Contexto: ensayos medicamentosos.  $D_1, D_2, \dots, D_N$  situaciones similares de medicamentos y sus ensayos.
- Se asume  $\mathbb{P}(H_0) = 0,5$ . Se simula para cada  $D_i$  la hipótesis, habiendo escogido el tamaño relevante del efecto si  $H_1$
- Se simulan ensayos independientes para cada  $D_i$

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Ha habido varios propuestas para calibrar el  $p$ -valor.

Varios introducen elementos Bayesianos.

En particular las probabilidades a priori de las hipótesis y el factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

Aquí aparece  $\mathbb{P}(H_0 | x_{\text{obs}})$ , que es el FPR (False Positive Risk).

T. Sellke; M. J. Bayarri; J. O. Berger (2001), "Calibration of p Values for Testing Precise Null Hypotheses", *The American Statistician*, 55:1, pp. 62-71.

### Simulación sencilla

- Contexto: ensayos medicamentosos.  $D_1, D_2, \dots, D_N$  situaciones similares de medicamentos y sus ensayos.
- Se asume  $\mathbb{P}(H_0) = 0,5$ . Se simula para cada  $D_i$  la hipótesis, habiendo escogido el tamaño relevante del efecto si  $H_1$
- Se simulan ensayos independientes para cada  $D_i$

Entre los  $D_i$  para los cuales  $p \simeq 0,05$ , como mínimo el 23 % no presenta efecto, es decir  $H_0$  cierta.

Segundo problema: es el umbral  $p = 0,05$  adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:

comment

## Redefine statistical significance

We propose to change the default  $P$ -value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers,  
Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini,  
Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber,  
Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster,  
Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald,  
Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka,  
Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler,  
David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell,  
Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa,  
Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau,  
Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt,  
Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young,  
Jonathan Zinman and Valen E. Johnson

Recientemente (2017), en Nature Human Behaviour, 70 autores:

comment

## Redefine statistical significance

We propose to change the default  $P$ -value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

Segundo problema: es el umbral  $p = 0,05$  adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:  
Para justificarlo, consideran factor Bayes:

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:

Para justificarlo, consideran factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

## Segundo problema: es el umbral $p = 0,05$ adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:

Para justificarlo, consideran factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

En un caso sencillo, usando cuotas para el BF:

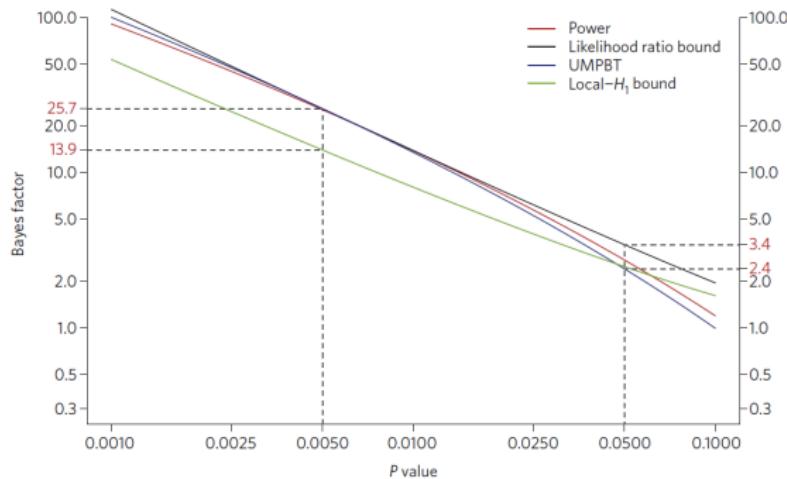
## Segundo problema: es el umbral $p = 0,05$ adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:

Para justificarlo, consideran factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

En un caso sencillo, usando cuotas para el BF:



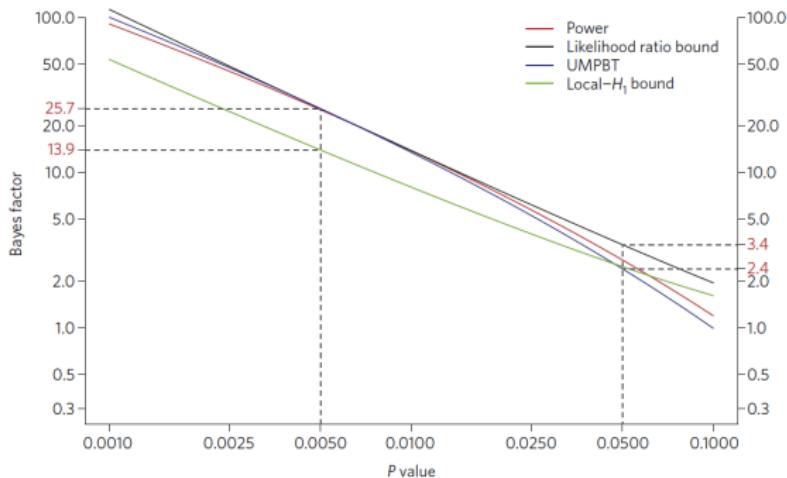
## Segundo problema: es el umbral $p = 0,05$ adecuado?

Recientemente (2017), en Nature Human Behaviour, 70 autores:

Para justificarlo, consideran factor Bayes:

$$\frac{\mathbb{P}(H_1 | x_{\text{obs}})}{\mathbb{P}(H_0 | x_{\text{obs}})} = \frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_0)} \times \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \equiv \text{BF} \times (\text{prior odds})$$

En un caso sencillo, usando cuotas para el BF:



- Para  $p \simeq 0,05$ ,  $2,5 \leq \text{BF} \leq 3,4$  "weak o very weak evidence"
- Para  $p \simeq 0,005$ ,  $14 \leq \text{BF} \leq 26$  "sustancial or strong"

¿Qué es el *p*-hacking?

## Tercer problema: *p*-hacking

¿Qué es el *p*-hacking?

*p*-hacking, popularizado por Uri Simonsohn, psicólogo U. Pensilvania

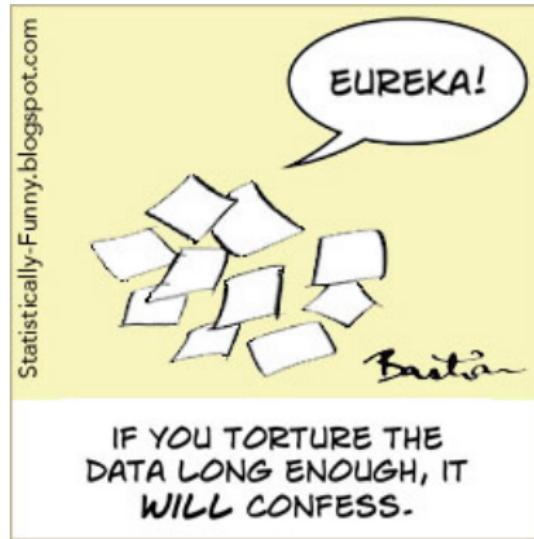
“*p*-hacking is trying multiple things until you get the desired result”

## Tercer problema: *p*-hacking

¿Qué es el *p*-hacking?

*p*-hacking, popularizado por Uri Simonsohn, psicólogo U. Pensilvania

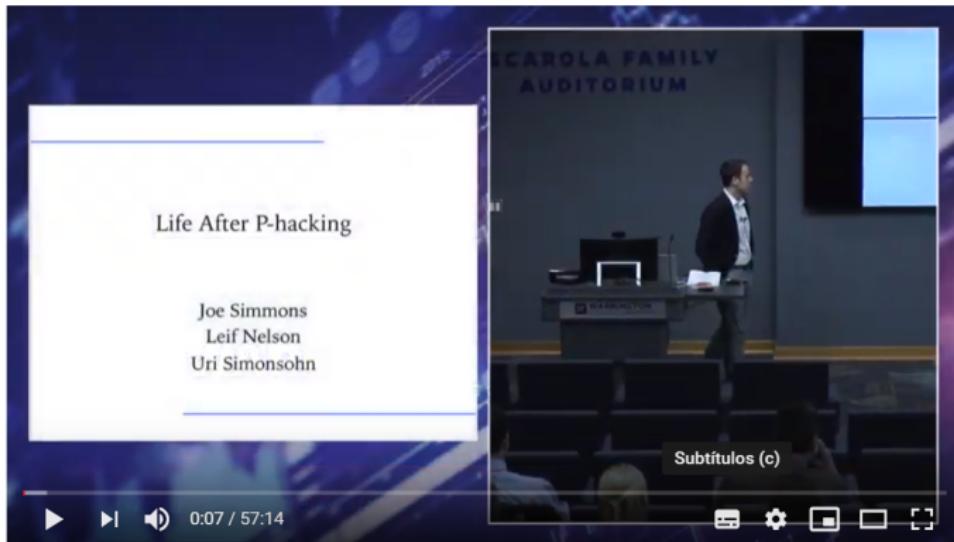
"*p*-hacking is trying multiple things until you get the desired result"



<https://statistically-funny.blogspot.com/2014/>

## Tercer problema: *p*-hacking

Muy recomendable: Seminario, julio 2018 “Reliable Research in Business”, University of Florida, Warrington College of Business [youtube](#)



Life After P-hacking

369 vistas

10    0    COMPARTIR    GUARDAR    ...

## Tercer problema: *p*-hacking

- El *p*-hacking no se realiza con intención de fraude.

- El *p*-hacking no se realiza con intención de fraude.
- Puede ser incluso inconsciente, una tentación natural

- El *p*-hacking no se realiza con intención de fraude.
- Puede ser incluso inconsciente, una tentación natural
- Nace de la presión por publicar.

- El *p*-hacking no se realiza con intención de fraude.
- Puede ser incluso inconsciente, una tentación natural
- Nace de la presión por publicar.

Puede tomar la forma de varios mecanismos, que se pueden resumir todos como “selective reporting”.

- El *p*-hacking no se realiza con intención de fraude.
- Puede ser incluso inconsciente, una tentación natural
- Nace de la presión por publicar.

Puede tomar la forma de varios mecanismos, que se pueden resumir todos como “selective reporting”.

Mencionamos dos de ellos:

- ① File drawering
- ② Cherry picking

### File drawering

Consiste en meter en el cajón los experimentos cuyos resultados salen no significativos.

### File drawering

Consiste en meter en el cajón los experimentos cuyos resultados salen no significativos.

¿Por qué?

### File drawering

Consiste en meter en el cajón los experimentos cuyos resultados salen no significativos.

¿Por qué? Por la presión de publicar.

### File drawering

Consiste en meter en el cajón los experimentos cuyos resultados salen no significativos.

¿Por qué? Por la presión de publicar.

Recordad: Ioannidis (2019), Entre los abstracts que contienen un *p*-valor el 96 % tienen  $p < 0,05$ .

### File drawering

Consiste en meter en el cajón los experimentos cuyos resultados salen no significativos.

¿Por qué? Por la presión de publicar.

Recordad: Ioannidis (2019), Entre los abstracts que contienen un *p*-valor el 96 % tienen  $p < 0,05$ .

Sin embargo, no parece común, por el coste que supone.

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

- Sólo escogiendo variables que presentan efecto significativo.

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

- Sólo escogiendo variables que presentan efecto significativo.
- Sólo me quedo con algunas covariables.

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

- Sólo escogiendo variables que presentan efecto significativo.
- Sólo me quedo con algunas covariables.
- Adaptando el tamaño de la muestra: si no tengo resultados significativos, sigo recogiendo datos.

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

- Sólo escogiendo variables que presentan efecto significativo.
- Sólo me quedo con algunas covariables.
- Adaptando el tamaño de la muestra: si no tengo resultados significativos, sigo recogiendo datos.
- y más...

### Cherry picking

Escoger las cerezas en la tarta, o en la cesta... Escoger sólo los mejores resultados

Se puede hacer de varias maneras.

- Sólo escogiendo variables que presentan efecto significativo.
- Sólo me quedo con algunas covariables.
- Adaptando el tamaño de la muestra: si no tengo resultados significativos, sigo recogiendo datos.
- y más...

Es muy fácil comprobar el efecto del “cherry-picking” por simulaciones, y tiene un gran valor pedagógico.

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

### Planteamiento

- En mi experimento, considero dos variables  $X$  e  $Y$ , posiblemente correladas.
- Quiero probar si presentan efecto, es decir  $\mu_X \neq 0$ ,  $\mu_Y \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - si solo uno de los dos *p*-valores,  $p_X$  o  $p_Y$  es menor de 0.05, sólo publico el resultado sobre esa variable.

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

### Planteamiento

- En mi experimento, considero dos variables  $X$  e  $Y$ , posiblemente correladas.
- Quiero probar si presentan efecto, es decir  $\mu_X \neq 0$ ,  $\mu_Y \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - si solo uno de los dos *p*-valores,  $p_X$  o  $p_Y$  es menor de 0.05, sólo publico el resultado sobre esa variable.

Muestra	$Z_X$	$Z_Y$	$p_X$	$p_Y$
1	-0.918	0.167	0.359	0.867
2	-1.12	0.178	0.261	0.858
3	-0.327	-1.50	0.744	0.134
4	-1.70	-1.63	0.0899	0.103
5	-0.246	0.457	0.806	0.648
:	:	:	:	:
59	-0.345	1.99	0.730	0.0462
:	:	:	:	:
170	2.54	2.20	0.0110	0.0276
:	:	:	:	:

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

### Planteamiento

- En mi experimento, considero dos variables  $X$  e  $Y$ , posiblemente correladas.
- Quiero probar si presentan efecto, es decir  $\mu_X \neq 0$ ,  $\mu_Y \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - si solo uno de los dos *p*-valores,  $p_X$  o  $p_Y$  es menor de 0.05, sólo publico el resultado sobre esa variable.

Muestra	$Z_X$	$Z_Y$	$p_X$	$p_Y$
1	-0.918	0.167	0.359	0.867
2	-1.12	0.178	0.261	0.858
3	-0.327	-1.50	0.744	0.134
4	-1.70	-1.63	0.0899	0.103
5	-0.246	0.457	0.806	0.648
:	:	:	:	:
59	-0.345	1.99	0.730	<b>0.0462</b>
:	:	:	:	:
170	2.54	2.20	0.0110	0.0276
:	:	:	:	:

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

### Planteamiento

- En mi experimento, considero dos variables  $X$  e  $Y$ , posiblemente correladas.
- Quiero probar si presentan efecto, es decir  $\mu_X \neq 0$ ,  $\mu_Y \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - si solo uno de los dos *p*-valores,  $p_X$  o  $p_Y$  es menor de 0.05, sólo publico el resultado sobre esa variable.

Muestra	$Z_X$	$Z_Y$	$p_X$	$p_Y$
1	-0.918	0.167	0.359	0.867
2	-1.12	0.178	0.261	0.858
3	-0.327	-1.50	0.744	0.134
4	-1.70	-1.63	0.0899	0.103
5	-0.246	0.457	0.806	0.648
:	:	:	:	:
59	-0.345	1.99	0.730	0.0462
:	:	:	:	:
170	2.54	2.20	0.0110	0.0276
:	:	:	:	:

# Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

```
library(dplyr)
N <- 1000
n <- 20
rho <- 0.5
simulated_data <- tibble(
  sample = rep(1:N, rep(n, N)),
  x = rnorm(N * n, 0, sd = 1),
  y = x * rho + rnorm(N * n, 0, sd = sqrt(1 - rho^2))
)

## -----
## calculamos los estadisticos para cada muestra
## -----

statistics <- simulated_data %>%
  group_by(sample) %>%
  summarise(
    znx = mean(x) / (1 / sqrt(n())),
    zny = mean(y) / (1 / sqrt(n())),
    px = 2 * pnorm( abs(znx), lower.tail = FALSE),
    py = 2 * pnorm( abs(zny), lower.tail = FALSE),
    Rx = px < 0.05,
    Ry = py < 0.05,
    Rxy = Rx | Ry
  )
## -----
## calculamos los ratios de falsos positivos
## -----

statistics %>%
  ungroup() %>%
  summarise(
    fprx = sum(Rx) / N * 100,
    fpry = sum(Ry) / N * 100,
    fprxy = sum(Rxy) / N * 100
)
```

## Tercer problema: *p*-hacking.

Para simular la elección selectiva de variables:

```
# A tibble: 1 x 3
  fprx   fpry   fprxy
  <dbl>  <dbl>  <dbl>
1     4.8    5.36   9.21
```

## Tercer problema: $p$ -hacking.

Para simular la parada selectiva de recogida de datos:

Para simular la parada selectiva de recogida de datos:

### Planteamiento

- En mi experimento, considero un variable  $X$
- Quiero probar si presenta efecto, es decir  $\mu_X \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - Empiezo por coger una muestra de 20. Si  $p_x \geq 0,05$ , cojo otros 10 y vuelvo a calcular *p*-valor.

## Tercer problema: *p*-hacking.

Para simular la parada selectiva de recogida de datos:

### Planteamiento

- En mi experimento, considero un variable  $X$
- Quiero probar si presenta efecto, es decir  $\mu_X \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - Empiezo por coger una muestra de 20. Si  $p_X \geq 0,05$ , cojo otros 10 y vuelvo a calcular *p*-valor.

Muestra	$Z_X$	$p_X$	$Z_X^{hacked}$	$p_X^{hacked}$
1	0.185	0.853	-0.450	0.653
2	1.31	0.190	1.00	0.316
3	0.493	0.622	1.20	0.230
4	-0.842	0.400	-0.421	0.673
5	1.37	0.170	1.33	0.182
⋮	⋮	⋮	⋮	⋮
26	-2.30	0.0214	-2.30	0.0214
⋮	⋮	⋮	⋮	⋮
43	-1.29	0.197	-2.18	0.0291
⋮	⋮	⋮	⋮	⋮

## Tercer problema: $p$ -hacking.

Para simular la parada selectiva de recogida de datos:

### Planteamiento

- En mi experimento, considero un variable  $X$
- Quiero probar si presenta efecto, es decir  $\mu_X \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - Empiezo por coger una muestra de 20. Si  $p_X \geq 0,05$ , cojo otros 10 y vuelvo a calcular  $p$ -valor.

Muestra	$Z_X$	$p_X$	$Z_X^{hacked}$	$p_X^{hacked}$
1	0.185	0.853	-0.450	0.653
2	1.31	0.190	1.00	0.316
3	0.493	0.622	1.20	0.230
4	-0.842	0.400	-0.421	0.673
5	1.37	0.170	1.33	0.182
⋮	⋮	⋮	⋮	⋮
26	-2.30	0.0214	-2.30	0.0214
⋮	⋮	⋮	⋮	⋮
43	-1.29	0.197	-2.18	0.0291
⋮	⋮	⋮	⋮	⋮

# Tercer problema: $p$ -hacking.

Para simular la parada selectiva de recogida de datos:

## Planteamiento

- En mi experimento, considero un variable  $X$
- Quiero probar si presenta efecto, es decir  $\mu_X \neq 0$ .
- Publico siempre, sea cual sea el resultado del experimento, pero
  - Empiezo por coger una muestra de 20. Si  $p_X \geq 0,05$ , cojo otros 10 y vuelvo a calcular  $p$ -valor.

Muestra	$Z_X$	$p_X$	$Z_X^{hacked}$	$p_X^{hacked}$
1	0.185	0.853	-0.450	0.653
2	1.31	0.190	1.00	0.316
3	0.493	0.622	1.20	0.230
4	-0.842	0.400	-0.421	0.673
5	1.37	0.170	1.33	0.182
⋮	⋮	⋮	⋮	⋮
26	-2.30	0.0214	-2.30	0.0214
⋮	⋮	⋮	⋮	⋮
43	-1.29	0.197	-2.18	0.0291
⋮	⋮	⋮	⋮	⋮

## Tercer problema: *p*-hacking.

Para simular la parada selectiva de recogida de datos:

```
N <- 1000
n <- 20
simulated_data <- tibble(
  sample = rep(1:N, rep(n, N)),
  x = rnorm(N * n, 0, sd = 1),
)
statistics <- simulated_data %>%
  group_by(sample) %>%
  summarise(
    znx = mean(x) / (1 / sqrt(n())),
    px = 2 * pnorm( abs(znx), lower.tail = FALSE),
    Rx = px < 0.05,
    znx_hacked = if_else(
      Rx,
      znx,
      mean(c(x, rnorm(10, 0, sd = 1))) * sqrt(n() + 10)
    ),
    px_hacked = 2 * pnorm( abs(znx_hacked), lower.tail = FALSE),
    Rx_hacked = px_hacked < 0.05
  )
statistics %>%
  ungroup() %>%
  summarise(
    fprx = sum(Rx) / N * 100,
    fprx_hacked = sum(Rx_hacked) / N * 100,
  )
```

## Tercer problema: $p$ -hacking.

Para simular la parada selectiva de recogida de datos:

```
# A tibble: 1 x 2
  fprx  fprx_hacked
  <dbl>      <dbl>
1   5.21       7.96
```

## Tercer problema: *p*-hacking.

Ejemplo en A/B testing:

Qué es el A/B testing? Extraido de <https://www.abtasty.com/es/ab-testing/>

El A/B testing consiste en comparar dos versiones de una misma página web o aplicación para comprobar cuál de las dos versiones es más eficiente. Estas variaciones, llamadas A y B, se muestran de forma aleatoria a los distintos usuarios de la página web. Una parte de ellos verá la versión A y la parte restante verá la versión B.

# Statistical Significance in A/B testing (and How People Misinterpret Probability)

2019-10-01



Tomi Mester

A few years ago we were running a major homepage A/B test with one of my clients. Huge traffic, huge potential, huge expectations — and huge risk, of course. We did our homework: our new design was well-researched and very promising, so we were all very excited. Especially Phil, the CEO of the company.

We launched the A/B test on the 1st of October and just in a few days the new version performed +20% better than the old one. The statistical significance was climbing slowly up, too: 50%, 60%, 70%... But then on the ~21st of October when I checked the data, our experiment was still not conclusive: +19% in conversion, with 81% significance.

But the client wanted results!

## Tercer problema: *p*-hacking.

¿Qué solución para el *p*-hacking?

## Tercer problema: *p*-hacking.

¿Qué solución para el *p*-hacking?

- Concienciación.

¿Qué solución para el *p*-hacking?

- Concienciación.
- Pedir que se publiquen todos los experimentos realizados

¿Qué solución para el *p*-hacking?

- Concienciación.
- Pedir que se publiquen todos los experimentos realizados
- Distinguir muy bien entre experimento exploratorio y experimento para investigar efecto.

¿Qué solución para el *p*-hacking?

- Concienciación.
- Pedir que se publiquen todos los experimentos realizados
- Distinguir muy bien entre experimento exploratorio y experimento para investigar efecto.

Esto se puede abordar pidiendo el pre-registro de los experimentos antes de la publicación.

¿Qué solución para el *p*-hacking?

- Concienciación.
- Pedir que se publiquen todos los experimentos realizados
- Distinguir muy bien entre experimento exploratorio y experimento para investigar efecto.

Esto se puede abordar pidiendo el pre-registro de los experimentos antes de la publicación.

Por ejemplo, en [aspredicted.org](http://aspredicted.org)

¿Qué solución para el *p*-hacking?

- Concienciación.
- Pedir que se publiquen todos los experimentos realizados
- Distinguir muy bien entre experimento exploratorio y experimento para investigar efecto.

Esto se puede abordar pidiendo el pre-registro de los experimentos antes de la publicación.

Por ejemplo, en [aspredicted.org](http://aspredicted.org)

Eso es parte de Open science.



Create a new AsPredicted pre-registration

CREATE

See your existing AsPredicteds (e.g. approve, make public)

Your email address (used in AsPredicted)

SEE OWN

## What's an AsPredicted?

It is a standardized pre-registration that requires only what's necessary to separate exploratory from confirmatory analyses. You will easily generate a pre-registration document that takes less effort to evaluate than it takes to evaluate the published study itself.

[About](#)   [Terms of use](#)

## How does it work?

- One author briefly answers 9 questions.
- All participating authors receive an email asking for approval.
- If everyone approves, it is saved and stays private until an author acts to make it public, or it remains private forever. ([Why?](#))
- Authors may share anonymous .pdf with reviewers.
- If made public, a single-page .pdf is generated. That document can be used as a supplement. ([See sample](#))
- The .pdf contains a unique URL that can be used for tracking.

## What if things don't go "as predicted"

You can just say so in the paper:

- "Contrary to expectations, we found that..."
- "Unexpectedly, we also found that..."
- "In addition to the analyses we pre-registered we also ran..."
- "We encountered an unexpected situation, and followed our Standard Operating Procedure" (.pdf)



Create a new AsPredicted pre-registration

CREATE

See your existing AsPredicteds (e.g. approve,  
make public)

Your email address (used in AsPredicted)

SEE OWN

### What's an AsPredicted?

It is a standardized pre-registration that requires only what's necessary to separate exploratory from confirmatory analyses. You will easily generate a pre-registration document that takes less effort to evaluate than it takes to evaluate the published study itself.

About

Terms of use



Create a new AsPredicted pre-registration

CREATE

See your existing AsPredicteds (e.g. approve,  
make public)

Your email address (used in AsPredicted)

SEE OWN

### How does it work?

- One author briefly answers 9 questions.
- All participating authors receive an email asking for approval.
- If everyone approves, it is saved and stays private until an author acts to make it public, or it remains private forever. ([Why?](#))
- Authors may share anonymous .pdf with reviewers.
- If made public, a single-page .pdf is generated. That document can be used as a supplement. ([See sample](#))
- The .pdf contains a unique URL that allows for one-click verification. That URL can be included in the paper.
- The .pdf is automatically stored in the web-archive. ([See sample](#))
- There are no accounts, userids, or passwords.



Create a new AsPredicted pre-registration

CREATE

See your existing AsPredicteds (e.g. approve, make public)

Your email address (used in AsPredicted)

SEE OWN

### What if things don't go "as predicted"

You can just say so in the paper:

- "Contrary to expectations, we found that..."
- "Unexpectedly, we also found that..."
- "In addition to the analyses we pre-registered we also ran..."
- "We encountered an unexpected situation, and followed our Standard Operating Procedure" (.pdf)

# Tercer problema: *p*-hacking.



## Sample 5 - SUMMER PROGRAMS - GPA performance, Chicago, July 2016 (#578)

### Author(s)

Larry TheRobot (AsPredicted College) - larry@aspredicted.org

Created: 04/07/2016

Made public: 04/07/2016

### 1) What's the main question being asked or hypothesis being tested in this study?

A month-long academic summer program for disadvantaged kids will reduce the drop in academic performance that occurs during the summer.

### 2) Describe the key dependent variable(s) specifying how they will be measured.

Simple average GPA across all courses during the first semester after the intervention.

### 3) How many and which conditions will participants be assigned to?

Two conditions: Offering summer program: yes vs no.

### 4) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Linear regression predicting the dependent variable with a dummy indicator for having been offered the summer program vs not (intention-to-treat analysis). We will also report results when controlling for baseline levels of the dependent variable (simple GPA average semester before training), gender & household income.

### 5) Any secondary analyses?

The effect may be larger for boys rather than girls, and for children living with one rather than two parents/guardians.

### 6) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

We will offer the program until 500 people have agreed to participate in it or until June 30, 2016 (whichever comes first).

### 7) Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

We include a battery of questions for exploratory purposes including happiness, videogame playing and family activity. We will also collect data on a survey with 24 questions and will report the results of those data in a separate project.

### 8) Have any data been collected for this study already?

No, no data have been collected for this study yet

Merece la pena recorrer las recomendaciones de

R. Wasserstein, A. Schirm & N. Lazar (2019) Moving to a World Beyond  
“ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

Merece la pena recorrer las recomendaciones de

R. Wasserstein, A. Schirm & N. Lazar (2019) Moving to a World Beyond  
“ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

Las resumen en:

ATOM

Merece la pena recorrer las recomendaciones de

R. Wasserstein, A. Schirm & N. Lazar (2019) Moving to a World Beyond  
“ $p < 0,05$ ”, *The American Statistician*, 73:sup1, 1-19. [enlace](#)

Las resumen en:

**ATOM**

- **A**ccept uncertainty.
- Be **T**houghtful.
- Be **O**pen.
- Be **M**odest.

# Conclusiones: ATOM

- **A**ccept uncertainty.

- **A**ccept uncertainty.
- Los métodos estadísticos no quitan la incertidumbre presente en los datos.

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- Be **T**houghtful.

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- Be **T**houghtful.

- Es recomendable reflexionar sobre muchos aspectos:
  - Cuáles son las consecuencias prácticas del estimador?. Cuál es su precisión? El modelo está bien especificado?

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- Be **T**houghtful.

- Es recomendable reflexionar sobre muchos aspectos:
  - Cuáles son las consecuencias prácticas del estimador?. Cuál es su precisión? El modelo está bien especificado?
- Ten en cuenta el contexto científico y el conocimiento previo:
  - Qué magnitud del efecto es relevante en la práctica? Eso antes de recoger datos.

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- Be **T**houghtful.

- Es recomendable reflexionar sobre muchos aspectos:
  - Cuáles son las consecuencias prácticas del estimador?. Cuál es su precisión? El modelo está bien especificado?
- Ten en cuenta el contexto científico y el conocimiento previo:
  - Qué magnitud del efecto es relevante en la práctica? Eso antes de recoger datos.
- Considera más que sólo el *p*-valor en la decisión: aspectos prácticos, calidad de los datos, conocimiento previo, etc...

- **A**ccept uncertainty.

- Los métodos estadísticos no quitan la incertidumbre presente en los datos.
- Presenta estimaciones de los efectos con su margen de error.
- Es bueno realizar estudios de réplica, hay que aceptar que pueden proporcionar resultados diferentes.

- Be **T**houghtful.

- Es recomendable reflexionar sobre muchos aspectos:
  - Cuáles son las consecuencias prácticas del estimador?. Cuál es su precisión? El modelo está bien especificado?
- Ten en cuenta el contexto científico y el conocimiento previo:
  - Qué magnitud del efecto es relevante en la práctica? Eso antes de recoger datos.
- Considera más que sólo el *p*-valor en la decisión: aspectos prácticos, calidad de los datos, conocimiento previo, etc...
- Complementa *p*-valor con otras medidas, considerar otras metodologías.

- Be **O**pen.

- Be **O**pen.
- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **M**odest

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **M**odest

- No confundas estadística y realidad. Un modelo siempre es una representación extremadamente simplificada de la realidad. ⇒ hay muchos modelos posibles.

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **M**odest

- No confundas estadística y realidad. Un modelo siempre es una representación extremadamente simplificada de la realidad. ⇒ hay muchos modelos posibles.
- Sé consciente de las limitaciones de los métodos estadísticos.

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **M**odest

- No confundas estadística y realidad. Un modelo siempre es una representación extremadamente simplificada de la realidad. ⇒ hay muchos modelos posibles.
- Sé consciente de las limitaciones de los métodos estadísticos.
- Anima otros a replicar tu trabajo.

- Be **O**pen.

- Abierto a la transparencia, abierto a las prácticas de “Open Science” (preregistro, comparte datos, comparte código, etc...)
- Abierto en la comunicación de los resultados.
- Abierto al hecho de que, a menudo, un estudio no es suficiente.

- Be **M**odest

- No confundas estadística y realidad. Un modelo siempre es una representación extremadamente simplificada de la realidad. ⇒ hay muchos modelos posibles.
- Sé consciente de las limitaciones de los métodos estadísticos.
- Anima otros a replicar tu trabajo.
- Acepta que siempre habrá incertidumbre.

- Mathieu Kessler, Elías Moreno (2015) It is forbidden to use p-values! BEIO, 31:2, 1 June 2015, 202-206  
<http://www.seio.es/BBEIO/BEIOVol31Num2>
- David Trafimow & Michael Marks (2015) Editorial, Basic and Applied Social Psychology, 37:1, 1-2  
<http://dx.doi.org/10.1080/01973533.2015.1012991>
- Nuzzo (2014) "Scientific method: statistical errors", Nature, News feature.  
<https://www.nature.com/news/scientific-method-statistical-errors-1.14700>
- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, <https://doi.org/10.1080/00031305.2016.1154108>
- ASA symposium on statistical Inference: october 11-13, 2017 "Scientific methods for the 21st century: A World Beyond  $p < 0,05$ .  
**Grabación de las sesiones:** [enlace](#)
- Special issue The American Statistician: Vol 73, March 2019, Issue 1. Statistical Inference in the 21st Century: A World Beyond  $p < 0,05$ . [enlace](#)

- Podcast Roger Peng y Hilary Parker. "Not so standard deviations", <http://nssdeviations.com/>. Episode 77, "Back to statistics".
- The reproducibility crisis in science: A statistical counterattack, Roger Peng, First published: 15 June 2015 [enlace](#)
- La depresión del Excel, Paul Krugman El pais 21 abril 2013 [enlace](#)
- Estimating the reproducibility of psychological science Open Science Collaboration\*,† (270 autores), Science 28 Aug 2015, Vol. 349, Issue 6251 [enlace](#)
- Nature Challenges in irreproducible research", 2018.  
<https://www.nature.com/collections/prbfkwmwvz>
- Final report - Science as an open enterprise, The Royal Society, 2012 [enlace](#)
- El portal de la Unión Europea sobre Open Science [enlace](#)
- John P. A. Ioannidis (2019) What Have We (Not) Learnt from Millions of Scientific Papers with P Values?, The American Statistician, 73:sup1, 20-25, [enlace](#)
- Steven N. Goodman (2019) Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics, The American Statistician, 73:sup1, 26-30, [enlace](#)

- Valentin Amrheim, Sander Greenland, Blake McShane (2019) Retire Statistical Significance”, Nature, 567, 305-307, [enlace](#)
- Thomas Sellke, M. J Bayarri & James O Berger (2001) Calibration of *p*-Values for Testing Precise Null Hypotheses, The American Statistician, 55:1, 62-71, [enlace](#)
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., and Cesarini, D. (2018), “Redefine Statistical Significance,” Nature Human Behaviour, 2, 6. [115,117] [enlace](#)
- Charla “Life after *p*-hacking”, John Simons en el seminario de julio 2018 “Reliable Research in Business”, University of Florida, Warrington College of Business [youtube](#)
- Recomiendo explorar para R los paquetes y la filosofía del “Tidyverse”, promovido por RStudio y colaboradores. Particularmente interesante es la librería “dplyr” y “purrr”. <https://www.tidyverse.org/>