

WARNING: P-hacking

GOT ANY PLANS FOR THE DAY?

I'M GOING TO EAT AN APPLE, AN EGG,
ONE BABY ASPIRIN, AND A PIECE OF
DARK CHOCOLATE, DRINK SIX GLASSES
OF WATER, ONE GLASS OF RED WINE,
A CUP OF COFFEE, AND A CUP OF TEA,
THEN DO 30 MINUTES OF EXERCISE.

THEN BACK TO SLEEP
FOR ANOTHER 8 HOURS!



I ONLY DO THINGS THAT NEWS STORIES HAVE
SPECIFICALLY TOLD ME TO DO ONCE PER DAY.

Dear News Media,

When reporting poll results, please keep
in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.
2. Scientific facts are not determined by public opinion polls.
3. A poll taken of your viewers/internet users is not a scientific poll.
4. What if all polls included the option "Don't care"?



Signed,

-Someone who took a
basic statistics course.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124.

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
<http://dx.doi.org/10.1080/00031305.2016.1154108>

EDITORIAL

The ASA's Statement on p -Values: Context, Process, and Purpose

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p -values: context, process, and purpose. *The American Statistician*, 70(2), 129–133.

STATISTICAL ERRORS

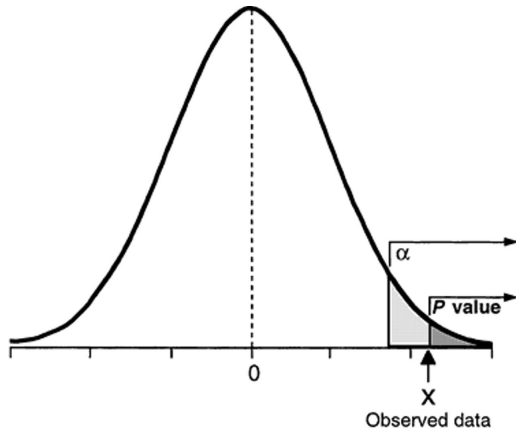
P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, 506(7487), 150.

¿A qué nos referimos como *p-hacking*?

Es la manipulación de los análisis / tests estadísticos con el fin de obtener un resultado significativo ($p < 0.05$). Específicamente se relaciona a cómo se obtiene e informa la significancia estadística.



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

<http://phdcomics.com/>

¿A qué nos referimos como p-hacking?

Q: ¿Por qué se enseña la regla “ $p = 0.05$ ” en tantas universidades?

A: Porque eso es lo que la comunidad científica y los editores de revistas todavía usan.

Q: ¿Por qué tanta gente todavía usa la regla “ $p = 0.05$ ”?

A: Porque eso es lo que se enseña en las universidades.

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
<http://dx.doi.org/10.1080/00031305.2016.1154108>

EDITORIAL

The ASA's Statement on p-Values: Context, Process, and Purpose

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

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

<http://phdcomics.com/>

¿A qué nos referimos como p-hacking?

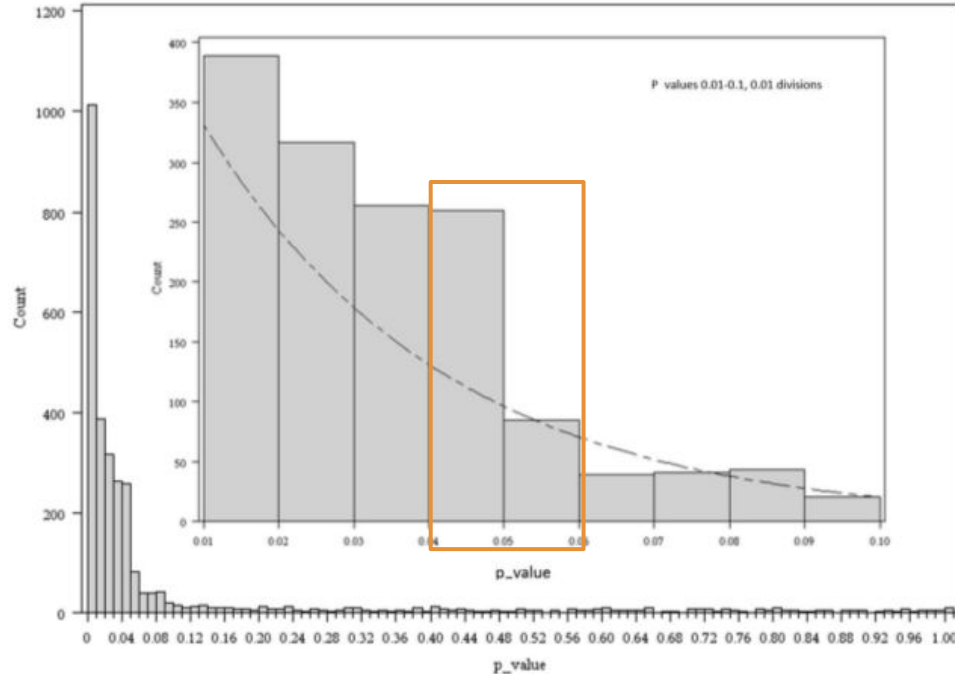


Fig.2 All p values between 0 and 1 are plotted in the *bottom graph*. The *inset* shows p values between 0.01 and 0.1 in 0.01 divisions

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

<http://phdcomics.com/>

Ginsel, B., et al (2015). The distribution of probability values in medical abstracts: an observational study. BMC res. notes, 8(1), 721.

¿Cómo se genera esta curva?

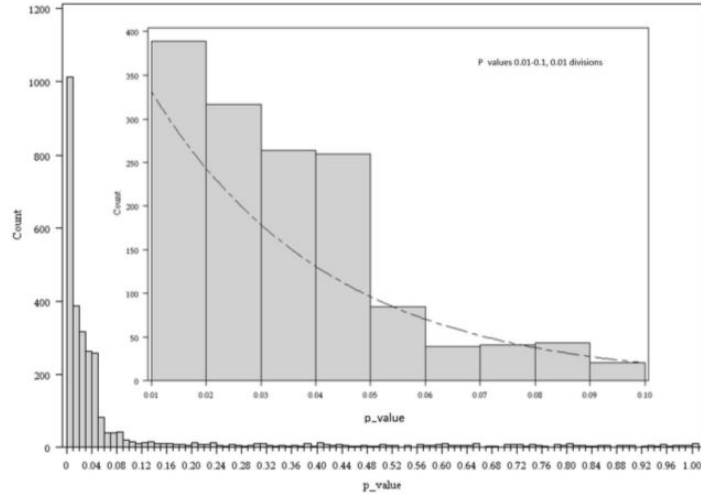
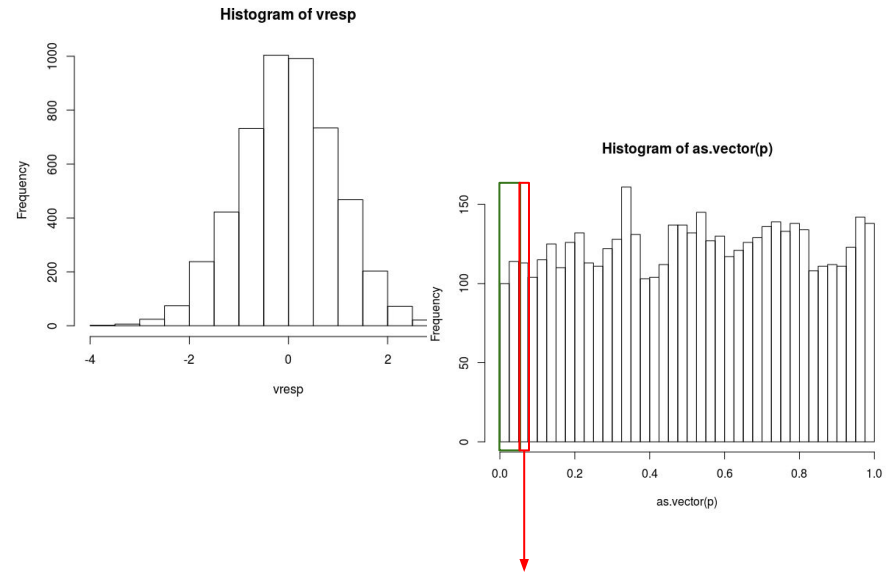


Fig. 2 All p values between 0 and 1 are plotted in the bottom graph. The inset shows p values between 0.01 and 0.1 in 0.01 divisions



¿Qué pasa si obtengo un resultados con p entre 0.05 y 0.075 y agrego algunos ejemplos más? (para decidir si es significativo o no)...

Ahora la probabilidad de caer en los $p < 0.05$ es del orden del ~25% !!!

¿A qué nos referimos como *p*-hacking?

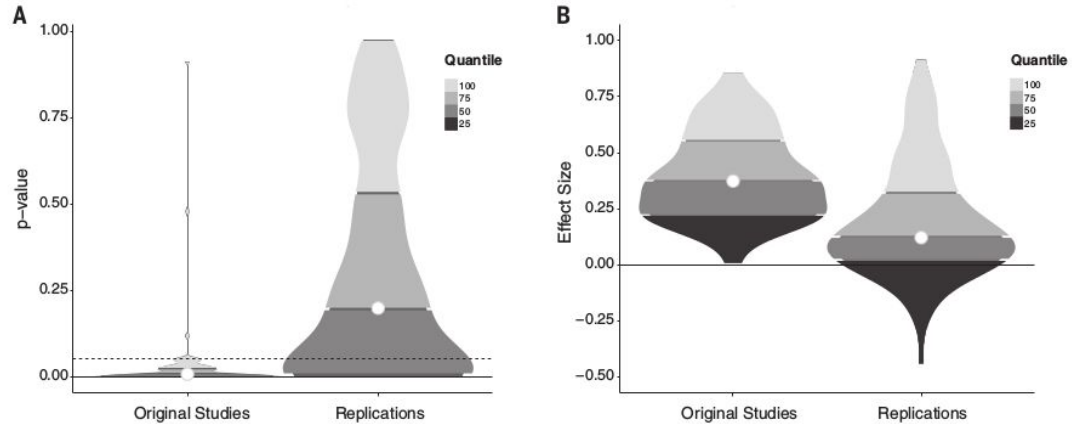


Fig. 1. Density plots of original and replication *P* values and effect sizes. (A) *P* values. (B) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Causas

¿Fraude?

¿Más presión?

¿Más oportunidades?

¿Descuido o desconocimiento?

THE AMERICAN STATISTICIAN

2016, VOL. 70, NO. 2, 129–133

<http://dx.doi.org/10.1080/00031305.2016.1154108>

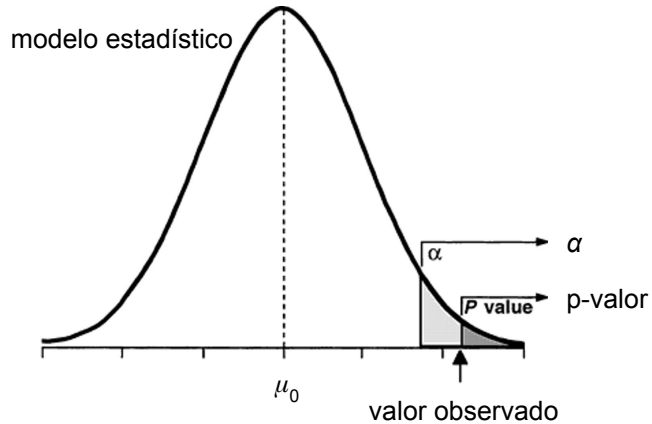
EDITORIAL

The ASA's Statement on p -Values: Context, Process, and Purpose

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p -values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

Principios o ¿Qué es (y qué no) es el p-valor?

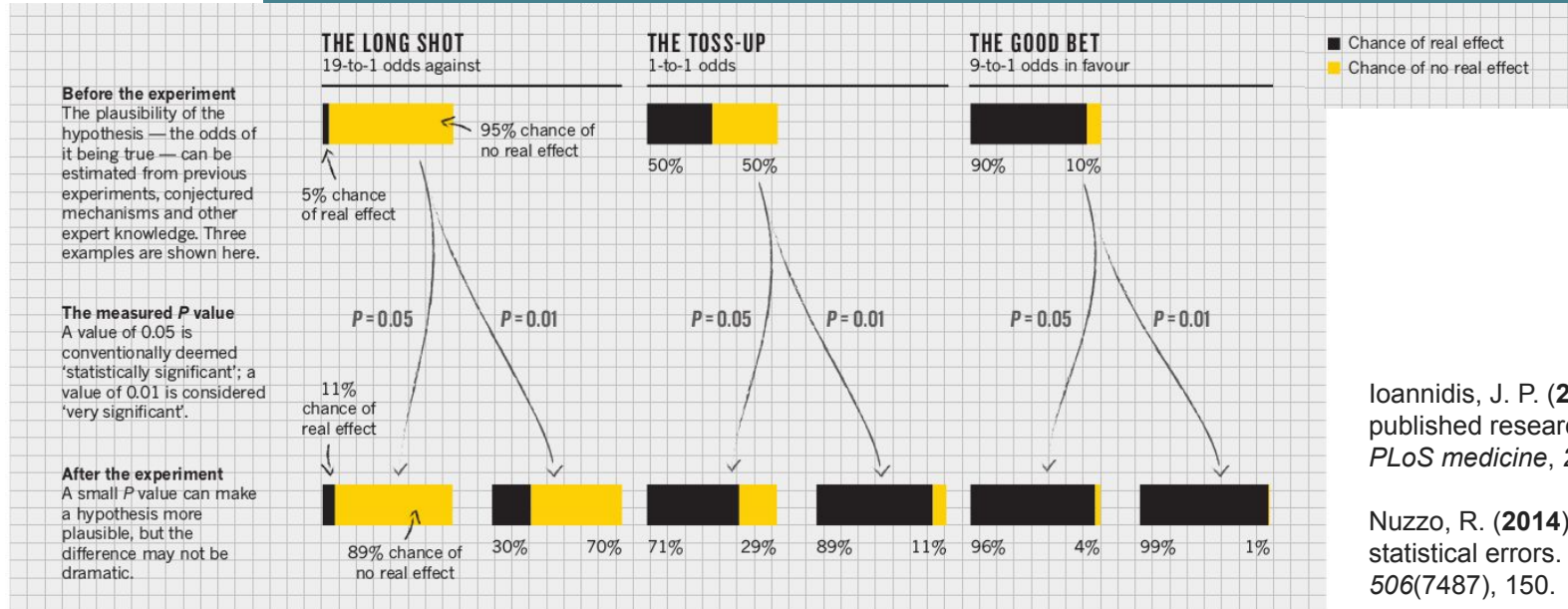
Los p-valores pueden indicar cuán incompatibles son los datos con un modelo estadístico dado.



Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

Principios

Los p-valores NO miden la probabilidad de que una hipótesis sea verdadera, o de la probabilidad de que los datos provengan únicamente del azar.



Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, 506(7487), 150.

Principios

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- ☐ Presidents
- ☒ Governors
- ☐ Senators
- ☐ Representatives

How do you want to measure economic performance?

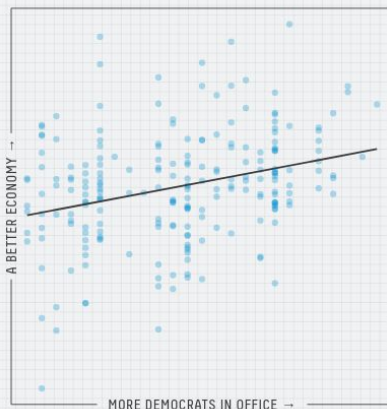
- ☐ Employment
- ☐ Inflation
- ☒ GDP
- ☒ Stock prices

Other options

- ☐ Factor in power
Weight more powerful positions more heavily
- ☒ Exclude recessions
Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



Result: Publishable

You achieved a p-value of less than 0.01 and showed that **Democrats** have a **positive** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

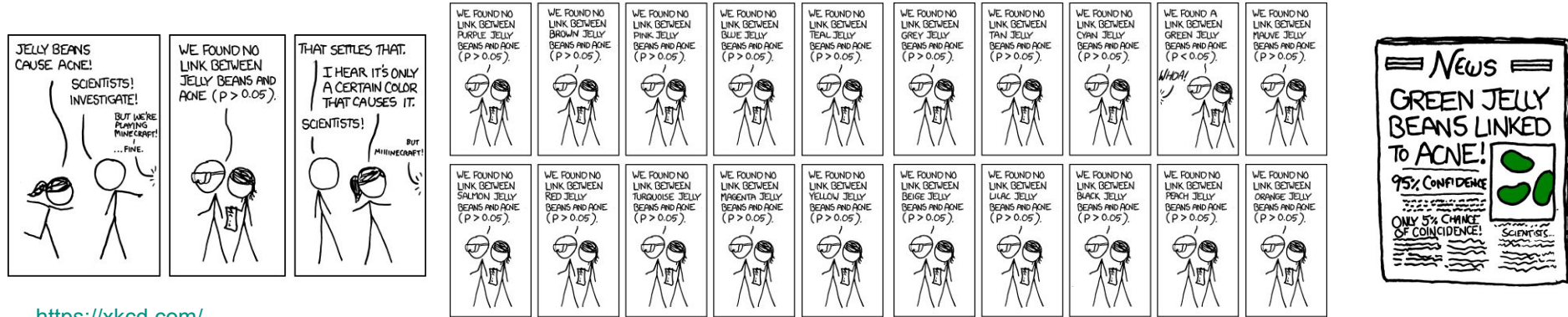
Las conclusiones científicas o políticas NO pueden basarse únicamente en si el p-valor pasa o no un umbral dado.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

<https://fivethirtyeight.com/features/science-isnt-broken/#part1>

Principios

Una inferencia apropiada requiere un reporte completo y transparente.



<https://xkcd.com/>

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

Principios

Los p-valores pueden indicar cuán incompatibles son los datos con un modelo estadístico dado.

Los p-valores NO miden la probabilidad de que una hipótesis sea verdadera, o de la probabilidad de que los datos provengan únicamente del azar.

Una inferencia apropiada requiere un reporte completo y transparente.

Un p-valor, o significancia estadística, NO mide el tamaño de un efecto o la importancia del resultado.

Por sí mismo, un p-valor NO provee una buena medida de la evidencia respecto a un modelo o hipótesis.

Wasserstein RL & Lazar NA (2016) "The ASA's Statement on p-Values: Context, Process, and Purpose", The American Statistician, 0:2, 129-133

Conclusiones. *“Ningún índice puede sustituir el razonamiento científico.”*

Buen diseño del estudio y adquisición de datos (“*garbage in, garbage out*”).

Hacer representaciones numéricas y gráficas buenas y variadas.

Comprender el fenómeno estudiado e interpretar los resultados en contexto.

Realizar reportes completos.

Comprender los métodos de análisis utilizados.

Wasserstein RL & Lazar NA (2016) “The ASA's Statement on p-Values: Context, Process, and Purpose”, The American Statistician, 0:2, 129-133

Recomendaciones

Replicar

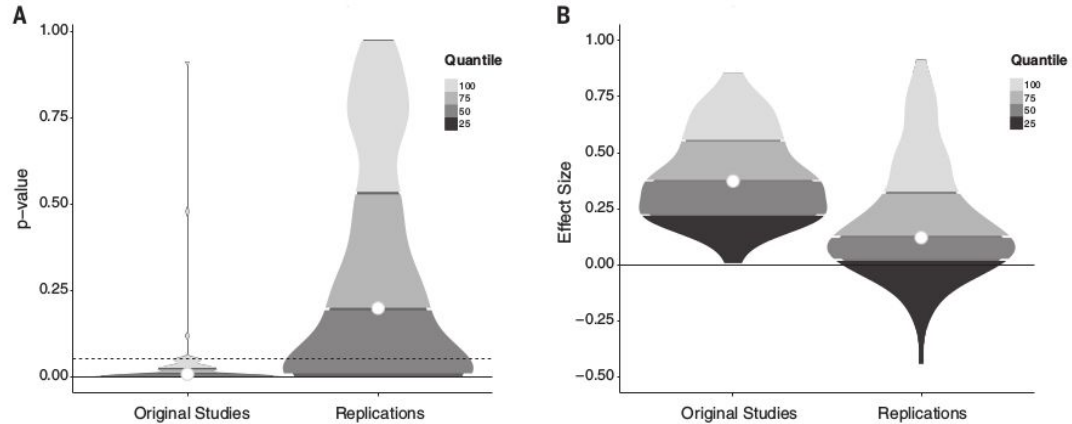


Fig. 1. Density plots of original and replication P values and effect sizes. (A) P values. (B) Effect sizes (correlation coefficients). Lowest quantiles for P values are not visible because they are clustered near zero.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Recomendaciones

Replicar

Registrar (o pre-registrar) proyectos

<http://www.timvanderzee.com/registered-reports/>

Recomendaciones

Replicar

Registrar (o pre-registrar) proyectos

Fomentar la honestidad en los análisis

Documentar, publicar código y datos - Investigación reproducible

Ser competente en las técnicas utilizadas

EDITORIAL

Ten Simple Rules for Effective Statistical Practice

Robert E. Kass¹, Brian S. Caffo², Marie Davidian³, Xiao-Li Meng⁴, Bin Yu⁵, Nancy Reid^{6*}

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., & Reid, N. (2016). Ten simple rules for effective statistical practice. PLoS computational biology, 12(6), e1004961.

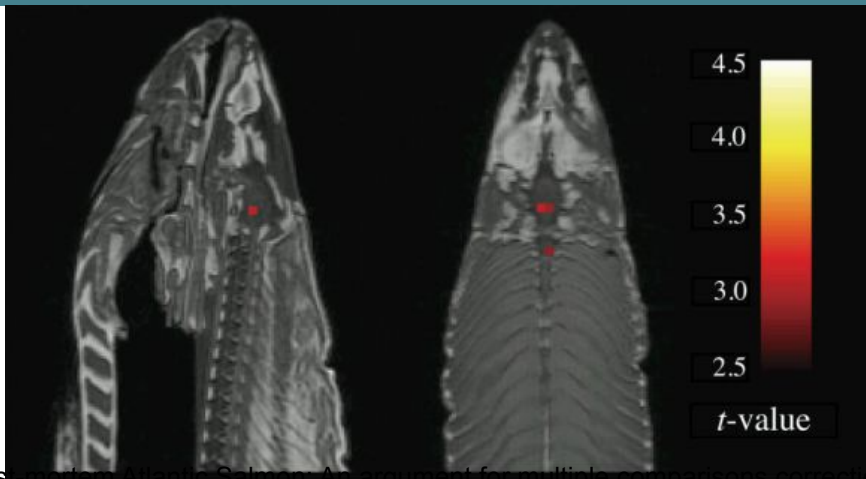
Recomendaciones

¿Cuán expertos son los expertos?

Recomendaciones

¿Cuán expertos son los expertos?

IgNobel: NEUROSCIENCE PRIZE: Craig Bennett, Abigail Baird, Michael Miller, and George Wolford [USA], for demonstrating that brain researchers, by using complicated instruments and simple statistics, can see meaningful brain activity anywhere — even in a dead



salmon. "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction," Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford, poster, 15th Annual Meeting of the Organization for Human Brain Mapping, San Francisco, CA, June 2009.

"Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Multiple Comparisons Correction," Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford, Journal of Serendipitous and Unexpected Results, vol. 1, no. 1, 2010, pp. 1-5.

¿Cuál fue el problema? Comparaciones múltiples



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

Anders Eklund^{a,b,c,1}, Thomas E. Nichols^{d,e}, and Hans Knutsson^{a,c}

^aDivision of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; ^bDivision of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden; ^cCenter for Medical Image Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; ^dDepartment of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; and ^eWMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

Random Field Theory en fMRI requiere que la autocorrelación espacial del ruido sea Gaussiana, y en general no es el caso... por lo que se aumentan los falsos positivos hasta un 70% (en casos extremos).

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28), 7900-7905.

Comparaciones múltiples

Table 1

Diagnoses for which residents with given astrological sign had a higher probability of hospitalization compared to residents born under the remaining astrological signs combined: results from derivation cohort

Astrological sign	ICD-9 code	Diagnosis	P-value	Relative risk
Aries	733	Other disorders of bone and cartilage	0.0402	1.27
	008	Intestinal infections due to other organisms	0.0058	1.41
Taurus	820	Fracture of neck of femur	0.0368	1.11
	562	Diverticula of intestine	0.0006	1.27
Gemini	998	Other complications of procedures, NEC	0.0330	1.15
	303	Alcohol dependence syndrome	0.0154	1.30
Cancer	560	Intestinal obstruction without mention of hernia	0.0475	1.12
	285	Other and unspecified anemias	0.0388	1.27
Leo	578	Gastrointestinal hemorrhage	0.0041	1.23
	V58	Encounter for other and unspecified procedure and aftercare	0.0397	1.17
Virgo	823	Fracture of tibia and fibula	0.0355	1.26
	643	Excessive vomiting in pregnancy	0.0344	1.40
Libra	808	Fracture of pelvis	0.0108	1.37
	430	Subarachnoid hemorrhage	0.0377	1.44
Scorpio	566	Abscess of anal and rectal region	0.0123	1.57
	204	Lymphoid leukemia	0.0395	1.80
Sagittarius	784	Symptoms involving head and neck	0.0376	1.30
	812	Fracture of humerus	0.0458	1.28
Capricorn	799	Other ill-defined and unknown causes or morbidity and mortality	0.0105	1.29
	634	Abortion	0.0242	1.28
Aquarius	413	Angina pectoris	0.0071	1.23
	481	Other bacterial pneumonia	0.0375	1.33
Pisces	428	Heart failure	0.0013	1.13
	411	Other acute and subacute forms of ischemic heart disease	0.0182	1.10

Abbreviation: NEC = not elsewhere classified.

Austin, P. C., Mamdani, M. M., Juurlink, D. N., & Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of clinical epidemiology*, 59(9), 964-969.

Comparaciones múltiples

Table 1

Diagnoses for which residents with given astrological sign had a higher probability of hospitalization compared to residents born under the remaining astrological signs combined: results from derivation cohort

Astrological sign	ICD-9 code	Diagnosis	P-value	Relative risk
Aries	733	Other disorders of bone and cartilage	0.0402	1.27
	008	Intestinal infections due to other organisms	0.0058	1.41
Taurus	820	Fracture of neck of femur	0.0368	1.11
	562	Diverticula of intestine	0.0006	1.27
Gemini	998	Other complications of procedures, NEC	0.0330	1.15
	303	Alcohol dependence syndrome	0.0154	1.30
Cancer	560	Intestinal obstruction without mention of hernia	0.0475	1.12
	285	Other and unspecified anemias	0.0388	1.27
Leo	578	Gastrointestinal hemorrhage	0.0041	1.23
	V58	Encounter for other and unspecified procedure and aftercare	0.0397	1.17
Virgo	823	Fracture of tibia and fibula	0.0355	1.26
	643	Excessive vomiting in pregnancy	0.0344	1.40
Libra	808	Fracture of pelvis	0.0108	1.37
	430	Subarachnoid hemorrhage	0.0377	1.44
Scorpio	566	Abscess of anal and rectal region	0.0123	1.57
	204	Lymphoid leukemia	0.0395	1.80
Sagittarius	784	Symptoms involving head and neck	0.0376	1.30
	812	Fracture of humerus	0.0458	1.28
Capricorn	799	Other ill-defined and unknown causes or morbidity and mortality	0.0105	1.29
	634	Abortion	0.0242	1.28
Aquarius	413	Angina pectoris	0.0071	1.23
	481	Other bacterial pneumonia	0.0375	1.33
Pisces	428	Heart failure	0.0013	1.13
	411	Other acute and subacute forms of ischemic heart disease	0.0182	1.10

Abbreviation: NEC = not elsewhere classified.



Austin, P. C., Mamdani, M. M., Juurlink, D. N., & Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of clinical epidemiology*, 59(9), 964-969.

¿Cuál fue el problema? Comparaciones múltiples

Type I Error



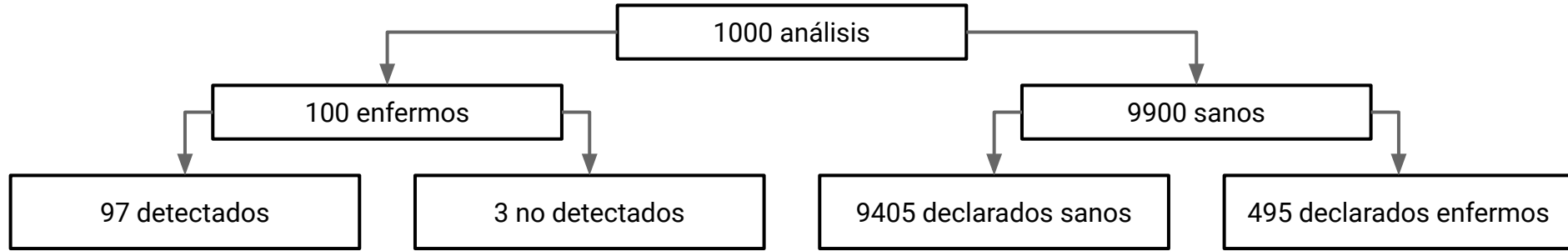
Type II Error



Comparaciones múltiples



Tasa de Falsos Positivos



Prevalencia: 0.01

Una de cada cien personas tiene la condición (100/1000)

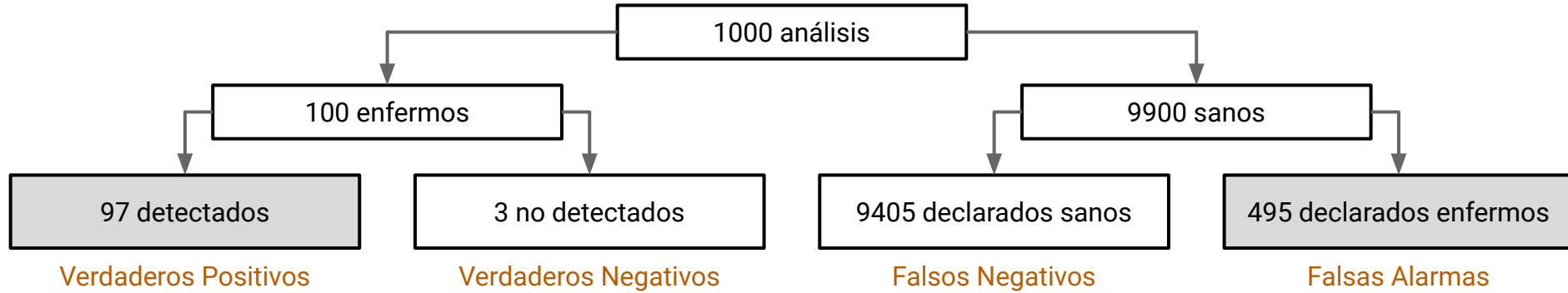
Sensibilidad: 0.97

97% personas con la condición se diagnostican correctamente (97/100)

Especificidad: 0.95

95% de las personas que no tiene la enfermedad es diagnosticada como sana (9405/9900)

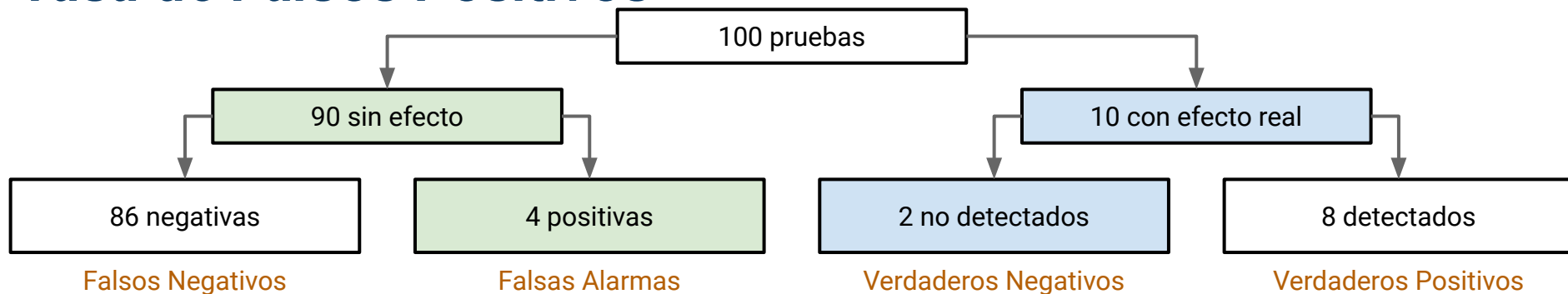
Tasa de Falsos Positivos



**Tasa de falsos descubrimientos
(False discovery rate)**

$$= \frac{\text{Falsas Alarmas}}{\text{Casos Positivos}} = \frac{495}{495 + 97} = 83.6 \%$$

Tasa de Falsos Positivos

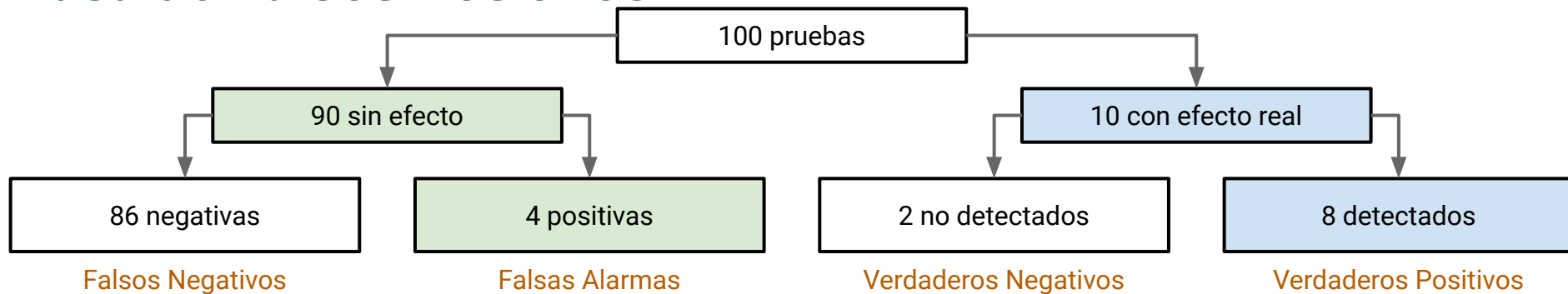


Error de tipo I (α): 0.05
Probabilidad de declarar
un test no significativo
cuando las diferencias
son reales.
(~4/90)

Error de tipo II: 0.20
Probabilidad de declarar
un test significativo
cuando las diferencias no
son reales
(Potencia: $1 - 0.20 = 0.80$)

	Verdadero (H_0 = Sano)	Falso (H_1 = Enfermo)
No rechazo H_0	$1-\alpha$	Error de tipo II (β)
Rechazo H_0	Error de tipo I (α)	$1-\beta$

Tasa de Falsos Positivos



$$\text{Tasa de falsos descubrimientos (False discovery rate)} = \frac{4}{4 + 8} = 33.3 \%$$

Tasa de Falsos Positivos

N tests independientes, la probabilidad de obtener k falsos positivos es:

$$P_{FP}(k) = \frac{N!}{(N-k)!k!} (1-\alpha)^{N-k} \alpha^k$$

(distribución Binomial)

Si N es grande y α es chico:

$$P_{FP}(k) = \frac{(N\alpha)^k \exp(-N\alpha)}{k!}$$

(distribución Poisson)

Comparaciones múltiples

Queremos ajustar los p-valores para compensar las comparaciones múltiples

```
graph LR; A[Queremos ajustar los p-valores para compensar las comparaciones múltiples] --> B[Métodos para control del Family-Wise Error Rate (FWER) (ej. Bonferroni)]; A --> C[Métodos para control del False Discovery Rate (FDR)];
```

Métodos para control del **Family-Wise Error Rate (FWER)** (ej. Bonferroni)

Métodos para control del **False Discovery Rate (FDR)**

Comparaciones múltiples

¿Cuándo hay que corregir?

Muchos tests para diferentes hipótesis. **No es necesario corregir.**

Muchos tests para la misma hipótesis comparando niveles de un mismo tratamiento, con potencial dependencia entre tests. Por ejemplo en una ANOVA. **Hay que corregir.**

Para una misma hipótesis y variable respuesta pero con diferentes comparaciones. Por ejemplo múltiples t-tests probando variables. **Hay que corregir.**

FWER: Bonferroni

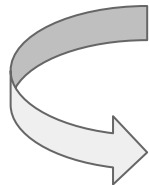
La probabilidad de no cometer Errores Tipo I en N tests es $(1-\alpha)^N$.

Entonces, la probabilidad de cometer al menos uno de estos errores es:

$$\pi = 1 - (1 - \alpha)^N$$

Este error se conoce tasa de error “experiment-wise” o tasa de error “family-wise” (**FWER**).

$$(1 - x)^N \approx 1 - xN$$


$$\alpha = 1 - (1 - \pi)^{1/N}$$

$$\alpha = \pi/N$$

FWER: Holm, Hochberg, Hommel, ...



Método de Holm

Ordenar los N p-valores de menor a mayor

Repetir para $n = 1$ hasta N

Ajustar utilizando la corrección de Bonferroni con $(N-n+1)$

Es decir, para el menor es dividir por N, para el siguiente N-1, N-2, ...

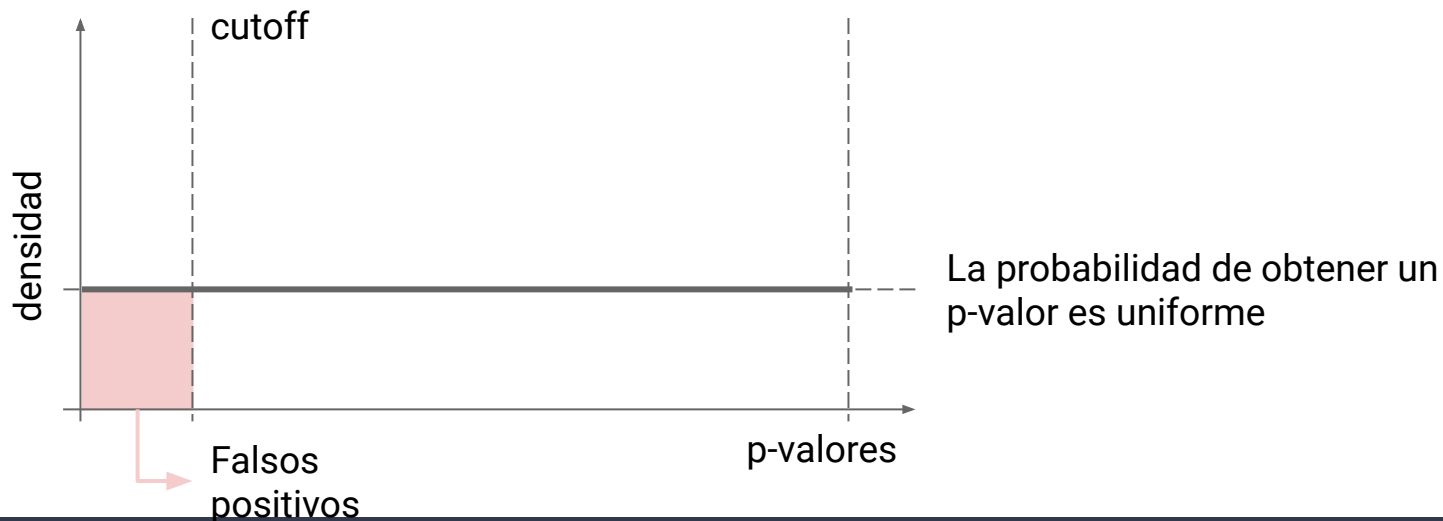
Método de Hochberg

Igual a Holm, pero de mayor a menor. Este método resulta más potente pero sólo es válido para tests independientes.

FDR

Estos métodos tratan de controlar pero no evitar la aparición de Falsos Positivos con la intención de no perder potencia.

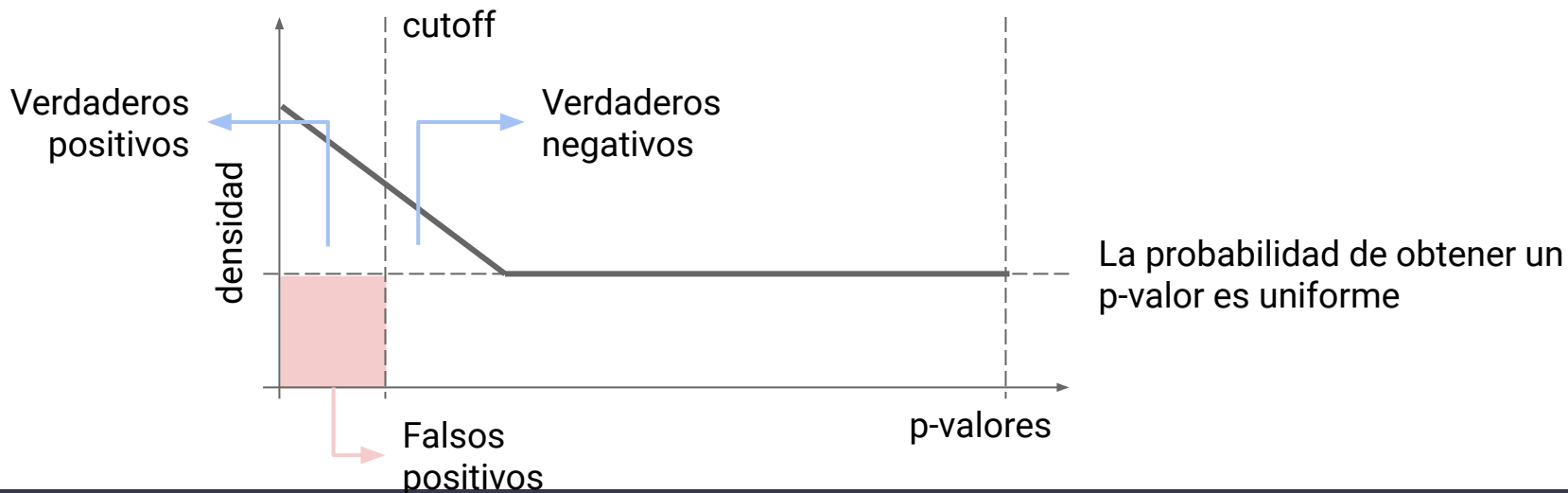
Son los métodos que suelen usarse cuando el N es muy grande y los FWER se vuelven muy restrictivos.



FDR

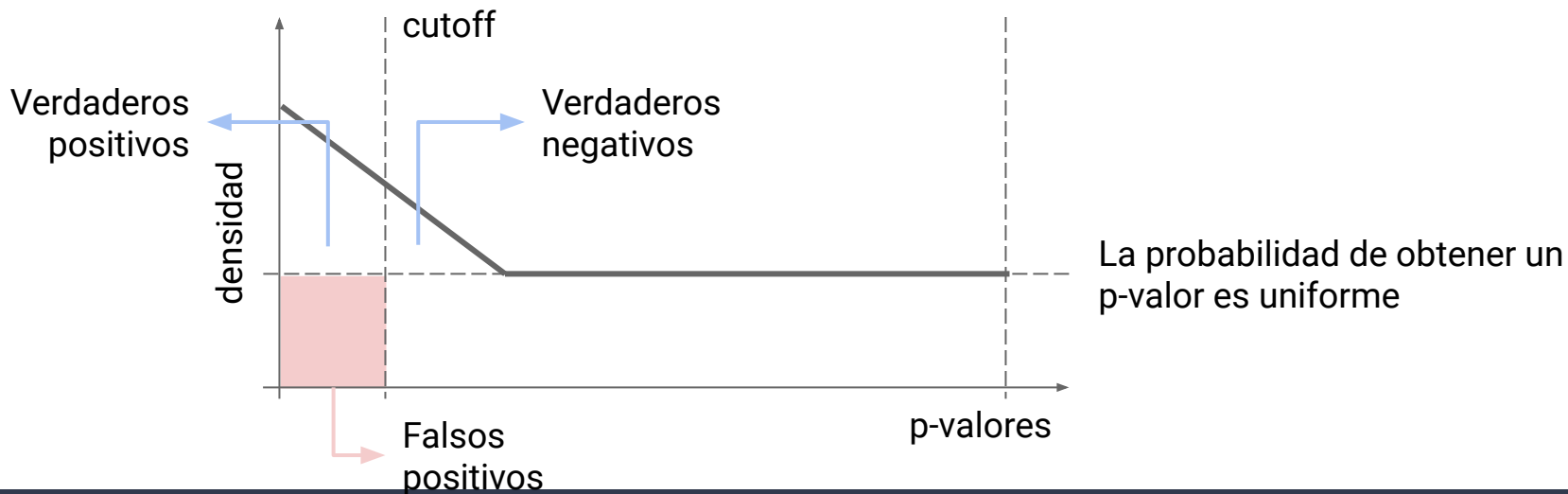
Estos métodos tratan de controlar pero no evitar la aparición de Falsos Positivos con la intención de no perder potencia.

Son los métodos que suelen usarse cuando el N es muy grande y los FWER se vuelven muy restrictivos.



FDR

El método de FDR selecciona un **cutoff** donde queda una buena proporción de p-valores significativos de variables con efecto real y se “escapan” algunos falsos positivos.



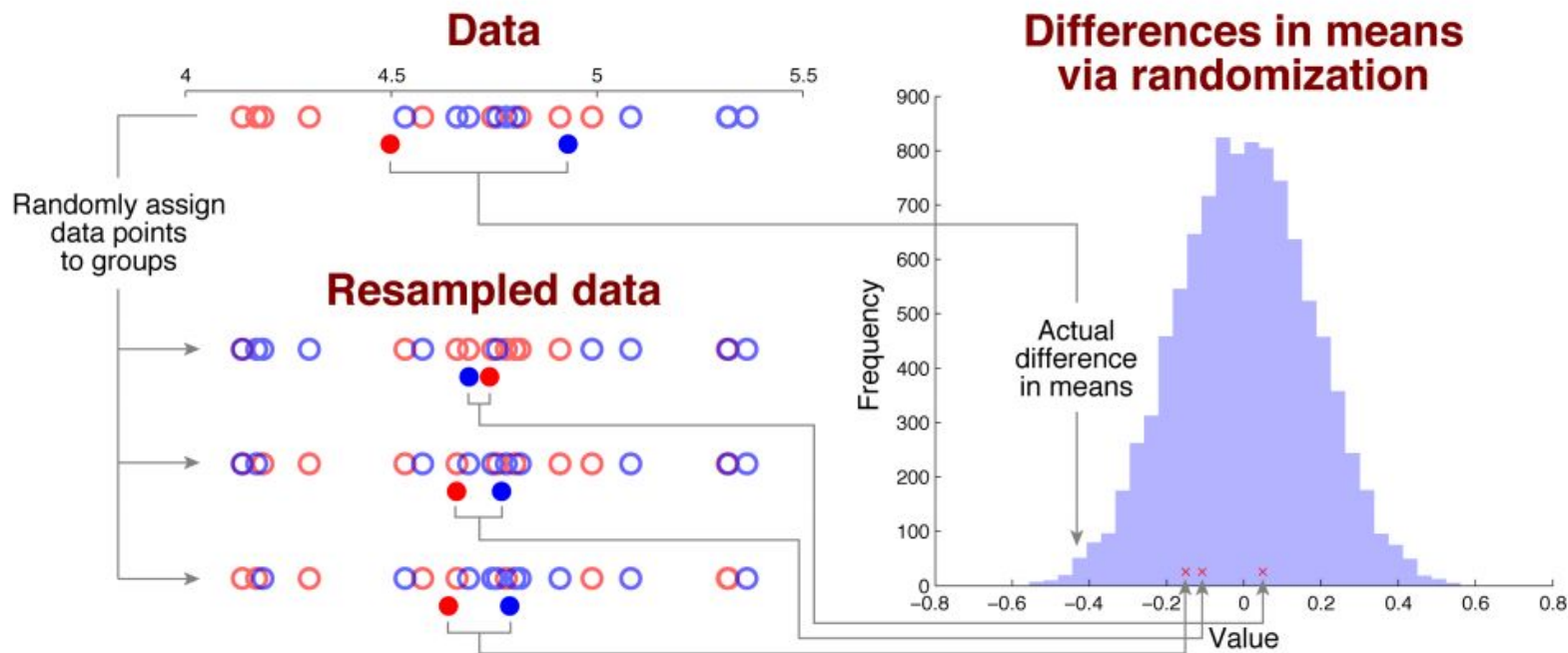
Algunas ideas asociadas al muestreo



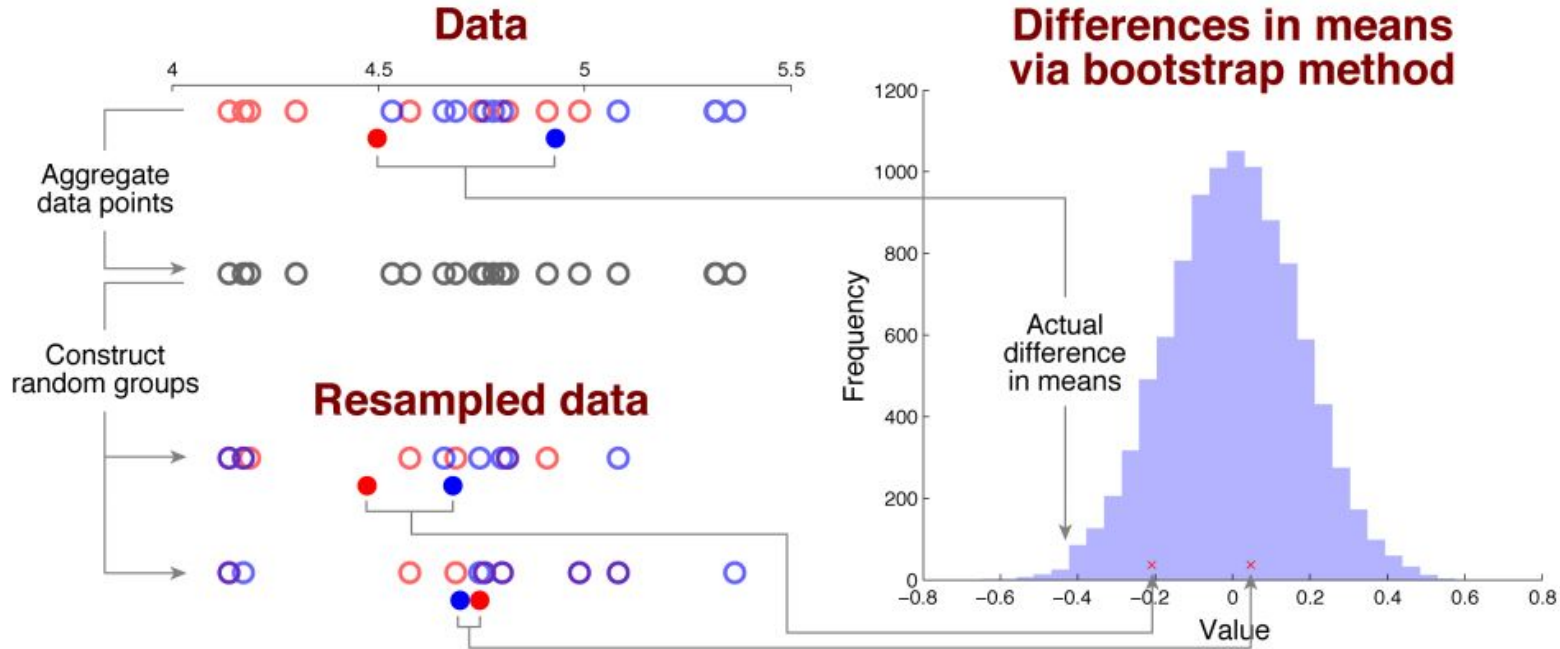
Randomization, Permutations (sin reposición)

Hipótesis nula: Rojos y Azules provienen de la misma distribución. ¿Cuál es la probabilidad de tomar dos subconjuntos que den una diferencia de medias mayor o igual a la original?

Idea: Los mezclo (igual son lo mismo), me genero una distribución de diferencias de medias y me fijo cuales son mayores o iguales al original.



Resampling, Bootstrapping (con reposición)



- No requieren asumir una distribución específica.
- Se pueden combinar con otros tests como regresiones.
- Hay que tener cuidado cuando hay varios factores involucrados, puede no ser trivial como construirse la hipótesis nula.

Randomization, Permutations (sin reposición)

- Evalúa específicamente **Exchangeability** (¿¿¿Intercambiabilidad???), y es más apropiado para **test de hipótesis**.
- Más apropiado para muestras pequeñas.

(ej. Mann-Withney / Wilcoxon)

Resampling, Bootstrapping (con reposición)

- Evalúa más específicamente la **variabilidad** ante un muestreo y resulta más apropiado para estimar **intervalos de confianza**.

Cross-Validation

→ En este caso, se quiere evaluar la capacidad de **generalizar** del modelo.

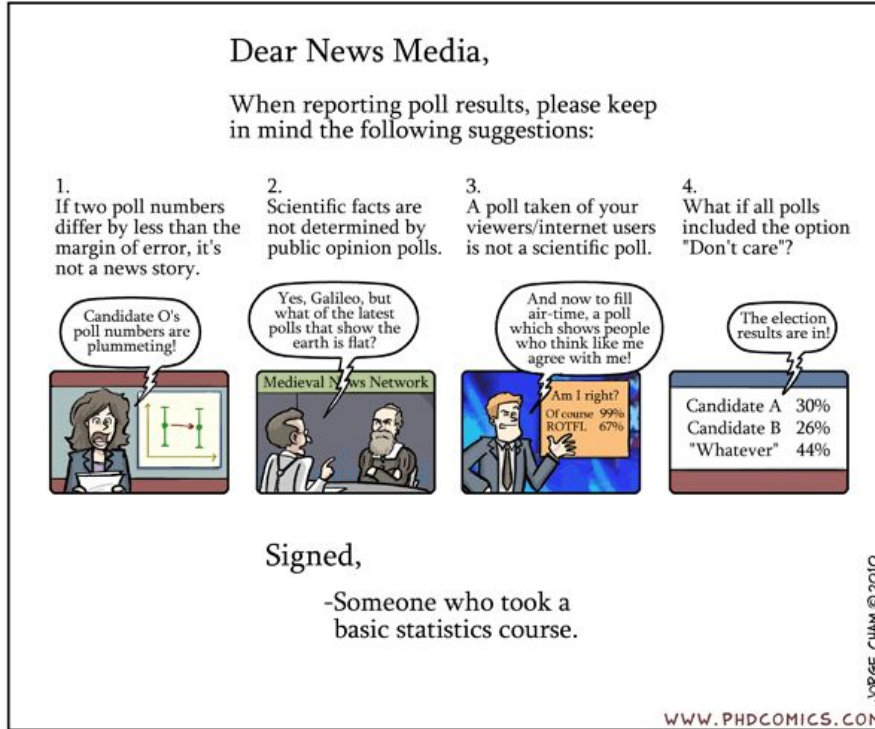
Muestra original				

5-fold cross-validation

- Esta también es una forma de **muestreo**, por lo que hay que tener cuidado de no introducir sesgos al partir la muestra.
- Es generalmente utilizado en *Machine Learning*, no para test de hipótesis.



¿A qué le llamamos p-hacking?



¿A qué le llamamos *p-hacking*?

CLICKBAIT-CORRECTED P-VALUE:

$$P_{CL} = P_{\text{TRADITIONAL}} \cdot \frac{\text{CLICK}(H_1)}{\text{CLICK}(H_0)}$$

NULL HYPOTHESIS

H_0 : ("CHOCOLATE HAS NO EFFECT
ON ATHLETIC PERFORMANCE")

ALTERNATIVE HYPOTHESIS

H_1 : ("CHOCOLATE BOOSTS
ATHLETIC PERFORMANCE")

FRACTION OF TEST SUBJECTS

$\text{CLICK}(H)$: WHO CLICK ON A HEADLINE
ANNOUNCING THAT H IS TRUE

¿A qué le llamamos *p-hacking*?

