

Estadística

Tests de hipótesis y p -Values

Crisis de replicabilidad y Buenas prácticas

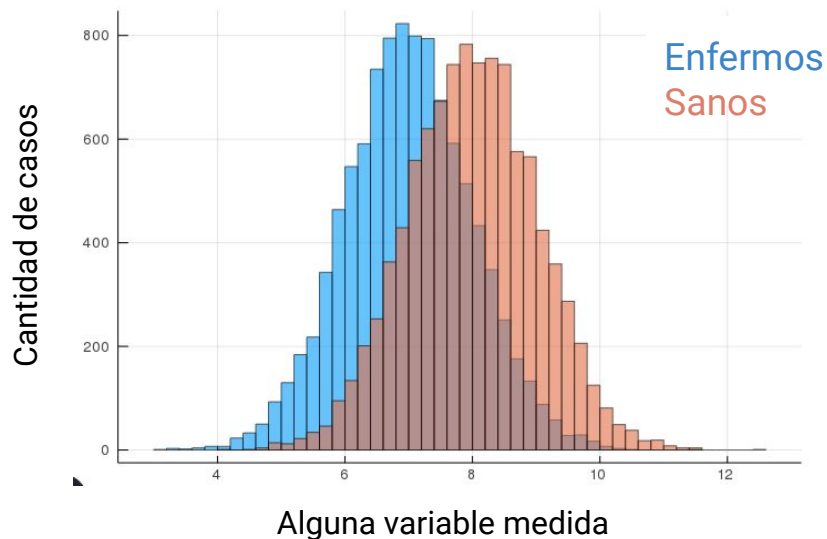
Comparaciones múltiples

Remuestreo: Permutaciones y Bootstrapping

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

¿Cómo comparar muestras de datos?

Pregunta clásica: ¿Mediciones realizadas bajo condiciones diferentes son diferentes o no?



¿La variable medida en personas sanas cómo se compara con la observada en personas enfermas?

Respuesta ordinal:
Es mayor, menor o igual.

No hay afirmaciones sobre el tamaño del efecto sobre la variable

Tests de hipótesis

- Para dar respuestas ordinales, utilizamos un procedimiento que se llama test de hipótesis.
- Para ello, se calcula el p -value que determina si hay una diferencia estadísticamente significativa entre los grupos de mediciones.
- Con “significativo” queremos decir: digno de atención. No es el punto final del análisis.
- Un test de hipótesis intenta discriminar entre señales dignas de atención y ruido aleatorio en los datos empíricos.
- La significancia estadística no es lo mismo que la significancia práctica (que además sugiere que el tamaño del efecto observado es lo suficientemente grande como para tener consecuencias reales).
- Un p -value nos provee una metodología para que no nos engañe el ruido aleatorio ni el sesgo de confirmación del experimentador.

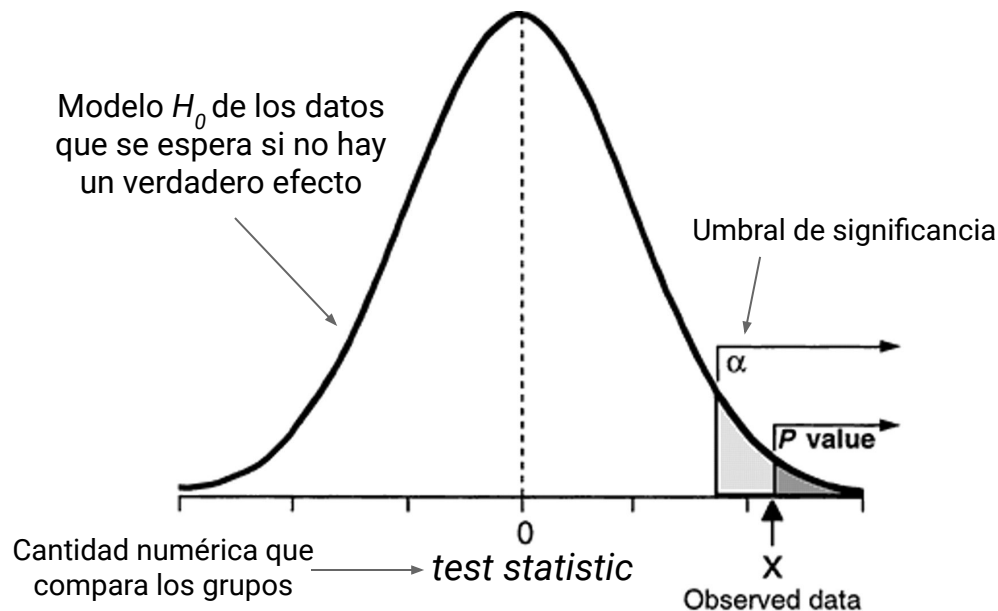
Tests de hipótesis

Pero, ¿qué es un p -value?

Tests de hipótesis

Pero, ¿qué es un p -value?

Test de significancia frente a una Hipótesis Nula (H_0) (Ronald Fisher)



p -value:

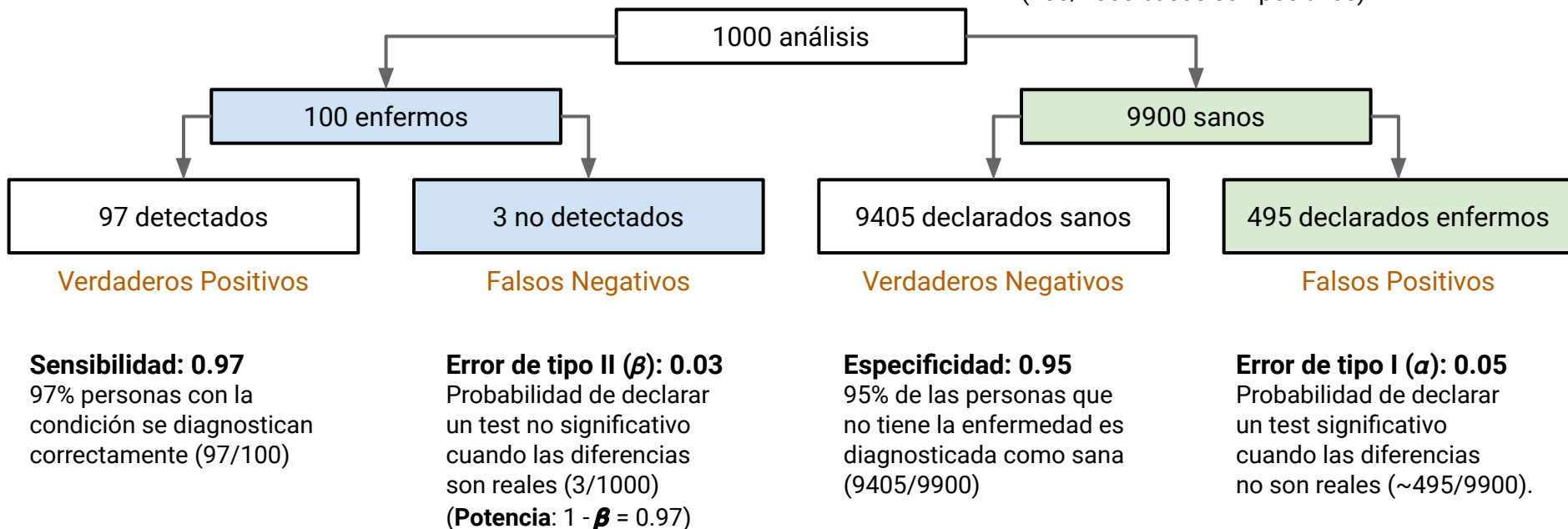
“Es la probabilidad de observar un estadístico que sea tan extremo o más extremo que el observado, si la hipótesis nula es verdadera.”

El p -valor es la probabilidad de que los datos observados sean tan extremos o más extremos que los observados, si la hipótesis nula es verdadera.

Tests de hipótesis

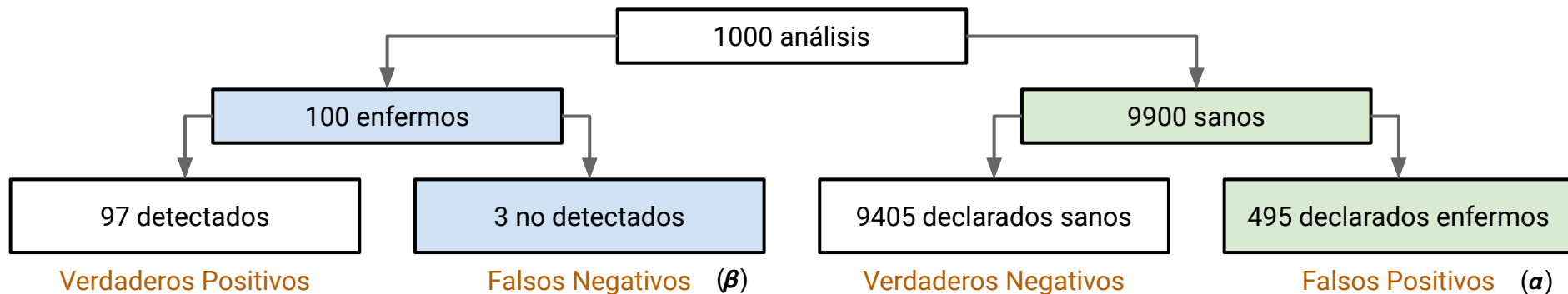
Otra manera de pensar el problema

Prevalencia: 0.01
(100/1000 casos son positivos)



Tests de hipótesis

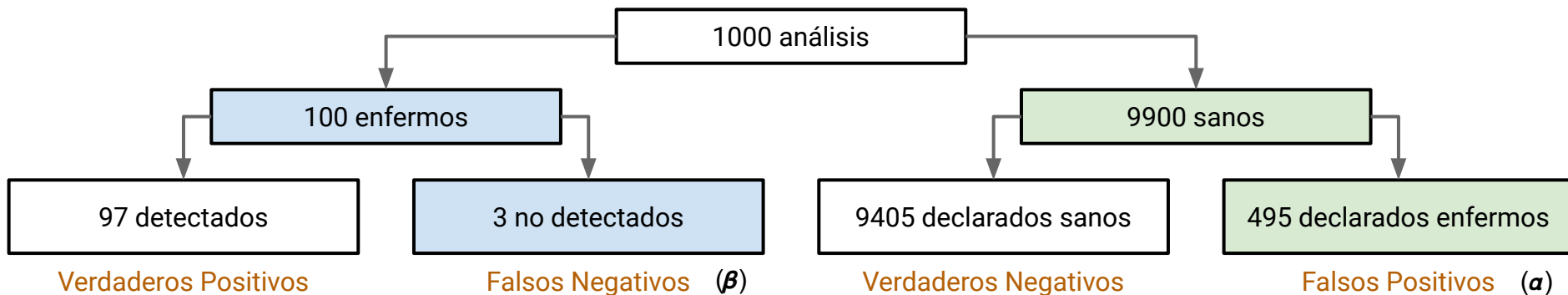
Otra manera de pensar el problema



	H_0 Verdadera	H_1 Verdadera
No rechazo H_0	$1-\alpha$ Verdaderos Negativos	Error de tipo II (β) Falsos Negativos
Rechazo H_0	Error de tipo I (α) Falsos Positivos	$1-\beta$ Verdaderos Positivos

Tests de hipótesis

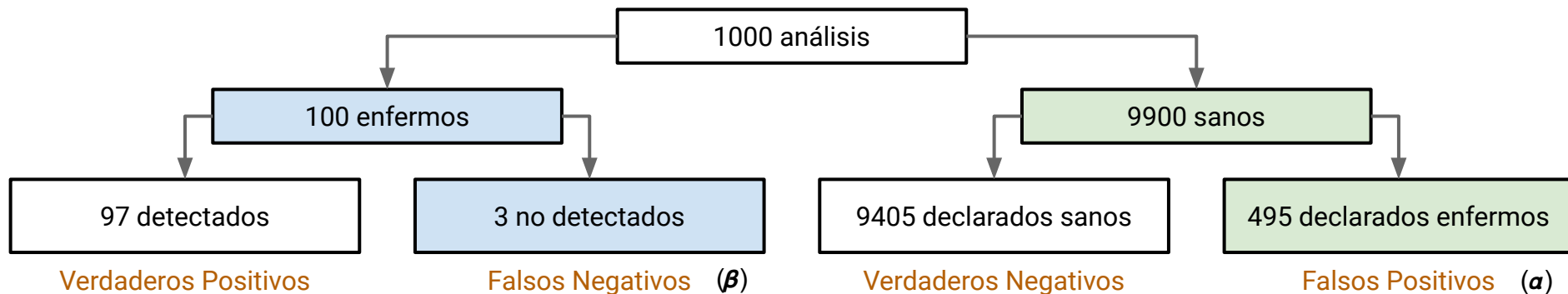
Otra manera de pensar el problema



	H_0 Verdadera	H_1 Verdadera
No rechazo H_0	$1-\alpha$ Verdaderos Negativos	Error de tipo II (β) Falsos Negativos
Rechazo H_0	Error de tipo I (α) Falsos Positivos	$1-\beta$ Verdaderos Positivos

Tests de hipótesis

Otra manera de pensar el problema

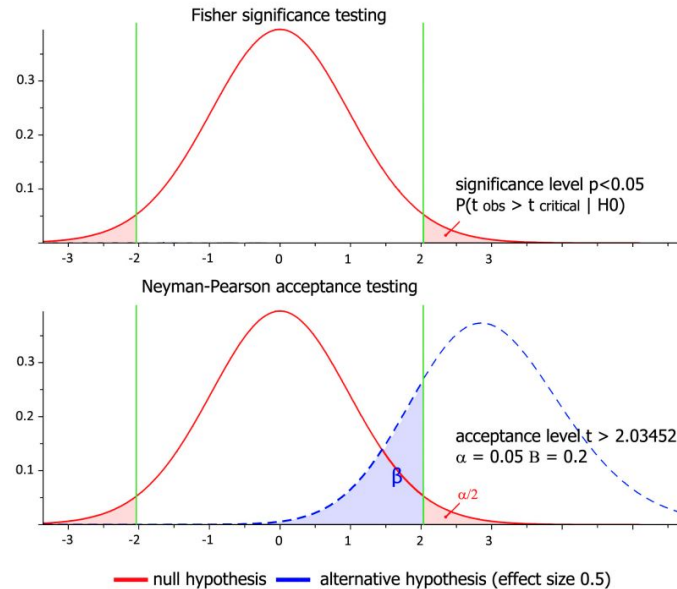


Siempre tenemos un compromiso entre los tipos de error que podemos aceptar

	H_0 Verdadera	H_1 Verdadera
No rechazo H_0	$1-\alpha$ Verdaderos Negativos	Error de tipo II (β) Falsos Negativos
Rechazo H_0	Error de tipo I (α) Falsos Positivos	$1-\beta$ Verdaderos Positivos

Tests de hipótesis

Tests de hipótesis como un proceso de toma de decisiones (Neyman-Pearson)



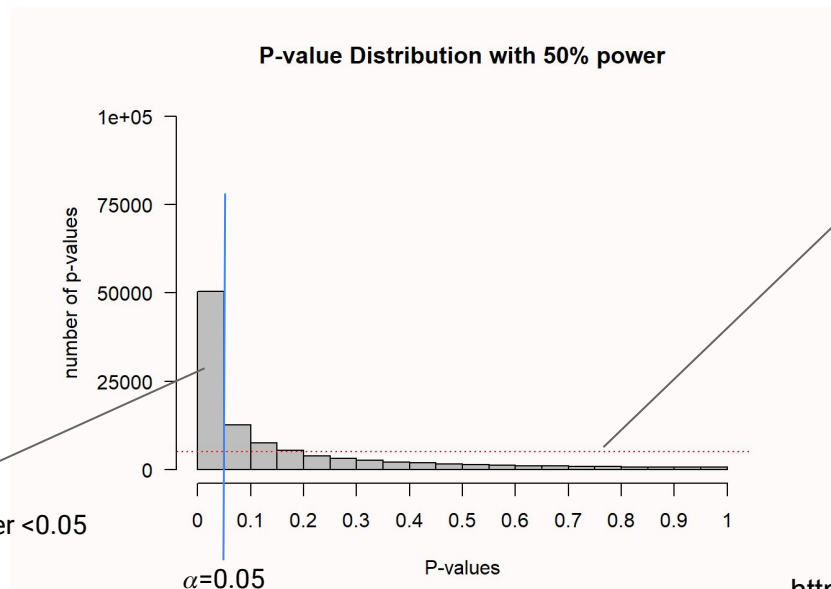
- Determinar dos hipótesis estadísticas: H_0 y H_1 , y fijar el α , β , y el tamaño de la muestra antes del experimento, considerando costos-beneficios. Esto define *regiones de rechazo* para cada hipótesis.
- Si los datos caen en la región de rechazo de H_0 , aceptar H_1 , sino aceptar H_0 . Aceptar significa **actuar** como si fuera verdadera, **no creerlo**.

Tests de hipótesis

¿Cómo se comporta el p-value **cuando hay un efecto real**?

Su comportamiento va a depender del poder estadístico ($1-\beta$) del experimento

100.000 experimentos
con poder de 50%



cantidad de experimentos
esperados bajo H_0 con un $\alpha=0.05$
(serían 5000)

Si hay efecto, cuanto
mayor sea el poder, los
 p -value estarán más
concentrados en
valores menores

50% de los p -values van a ser <0.05

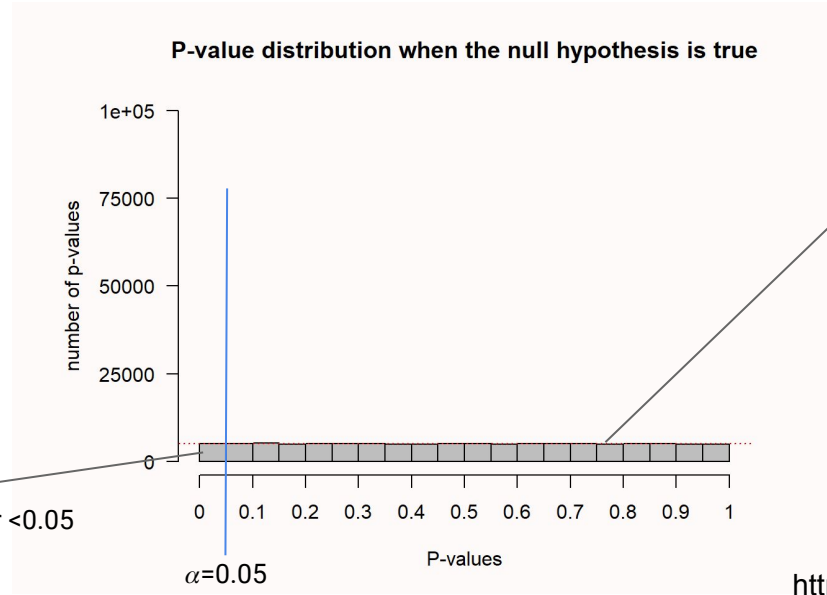
https://lakens.github.io/statistical_inferences/

Tests de hipótesis

¿Cómo se comporta el p-value **cuando NO hay un efecto real?**

¡La distribución va a ser **uniforme**! Cualquier p -value es igualmente posible

100.000 experimentos
con poder de 5%



cantidad de experimentos
esperados bajo H_0 con un $\alpha=0.05$
(serían 5000)

Cuando H_0 es verdadera, el
 α % de p-values deben caer
bajo el nivel de α .
Esto solo puede ocurrir si
los p-values se distribuyen
uniformemente.

5% de los p -values van a ser <0.05

https://lakens.github.io/statistical_inferences/

Crisis de replicabilidad y Buenas prácticas

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

La replicabilidad de estudios puede ser baja en algunos campos

Replicando 100 experimentos publicados

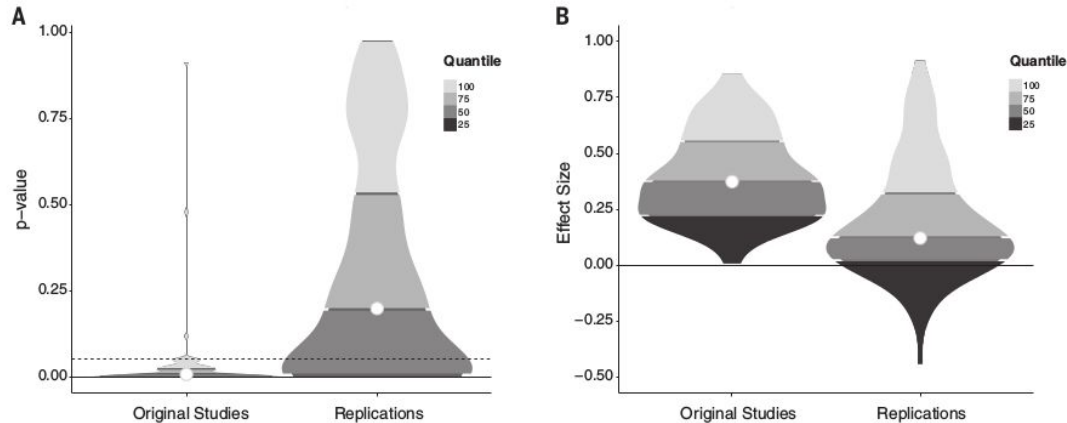


Fig. 1. Density plots of original and replication P values and effect sizes. (A) P values. (B) Effect sizes (correlation coefficients). Lowest quantiles for P values are not visible because they are clustered near zero.

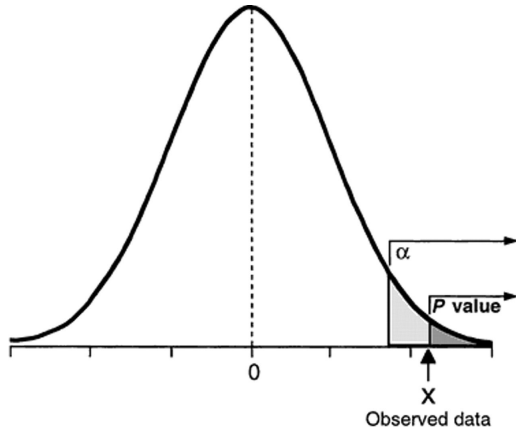
No hay una única medida de “replicabilidad”, pero:

- El tamaño del efecto promedio bajó a la mitad en los estudios de replicación
- Los resultados con $p < 0.05$ pasaron del 97% al 36%

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

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/>

p-hacking

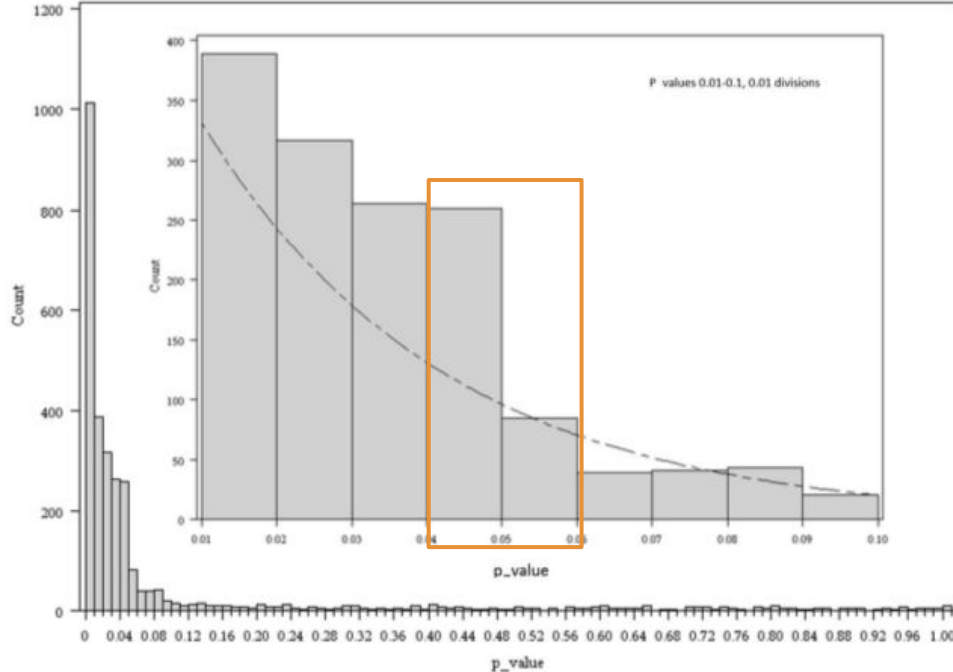


Fig.2 All p values between 0 and 1 are plotted in the *bottom graph*. The *inset* shows p values between 0.01 and 0.1 in 0.01 divisions

<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.

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.

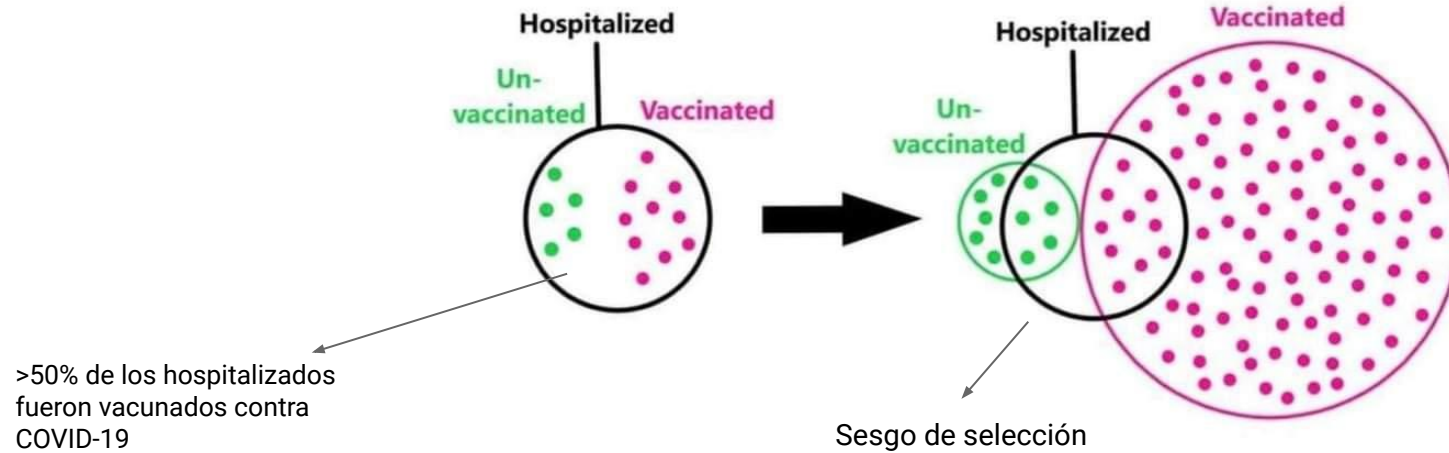
Q: ¿Por qué tanta gente todavía usa la regla “ $p = 0.05$ ”?

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

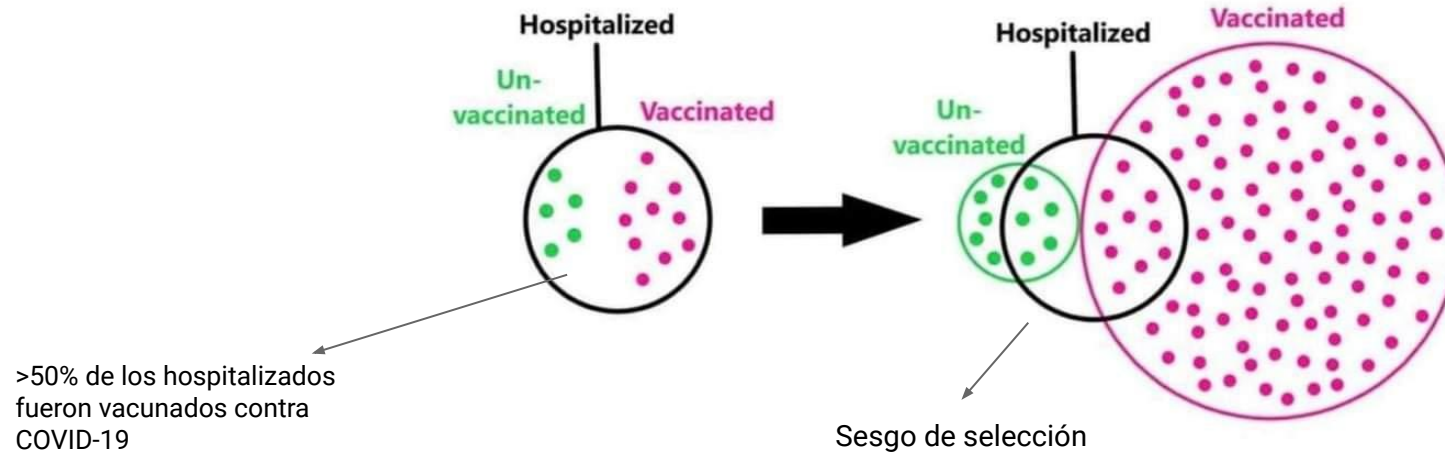
<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS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

<http://phdcomics.com/>

Sesgo de publicación



Sesgo de publicación



- En el ámbito de las publicaciones científicas hay un “sesgo de publicación”: sólo se publican los resultados significativos (por ej., $p < 0.05$)
- ¿Cómo evaluar la **proporción de resultados en la literatura que esperamos sean efectivamente verdaderos** si no tenemos acceso al **conjunto total de estudios**, inclusive los resultados no significativos obtenidos? (o sea, no conocemos los falsos negativos + verdaderos negativos)

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.

¿Qué porcentaje de los resultados publicados son verdaderos?

Vamos a modelar la Positive Predictive Value (PPV) de los artículos científicos publicados en ciencias médicas

$$PPV = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}} \longrightarrow \# \text{ de artículos publicados}$$

Why Most Published Research Findings Are False

John P. A. Ioannidis

¿Si obtenemos un resultado significativo ya está?

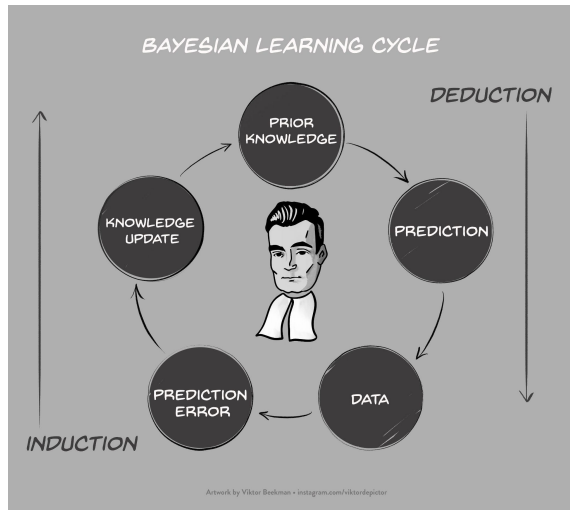
¿La plausibilidad de nuestra hipótesis H_1 no importa?

Significancia de un estudio en particular

	H_1 Verdadera	H_0 Verdadera
Resultado Significativo	$(1-\beta)$ Verdaderos Positivos	α Falsos Positivos
Resultado No Significativo	β Falsos Negativos	$(1-\alpha)$ Verdaderos Negativos

Why Most Published Research Findings Are False

John P. A. Ioannidis

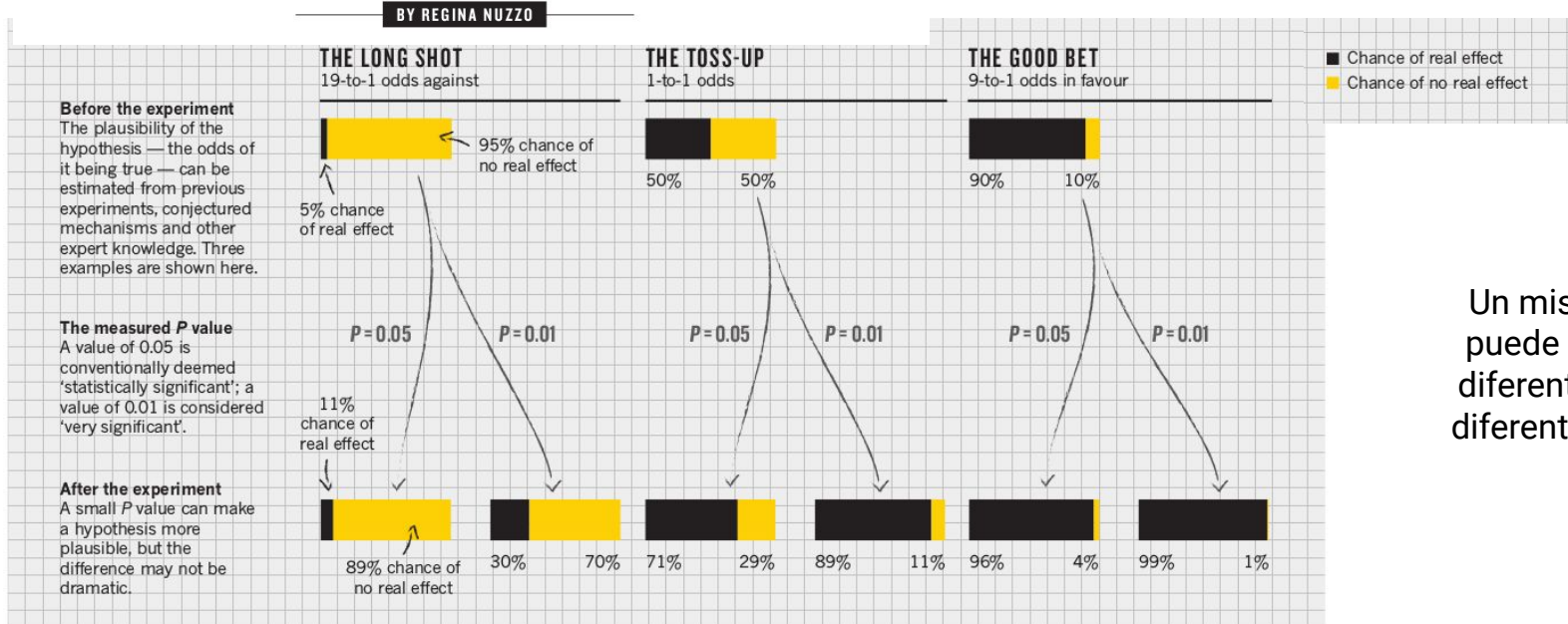


Tenemos que tomar en cuenta nuestra evaluación (¿subjetiva? ¿objetiva?) de que nuestra hipótesis H_1 sea verdadera antes de ver los datos

STATISTICAL ERRORS

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

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



Un mismo *p*-Value puede querer decir diferentes cosas en diferentes contextos

Why Most Published Research Findings Are False

John P. A. Ioannidis

¿Cuántos tipos de resultados espero obtener?

	H_1 Verdadera	H_0 Verdadera
Resultado Significativo	$(1-\beta) P(H_1=\text{True})$ Verdaderos Positivos	$\alpha P(H_0=\text{True})$ Falsos Positivos
Resultado No Significativo	$\beta P(H_1=\text{True})$ Falsos Negativos	$(1-\alpha) P(H_0=\text{True})$ Verdaderos Negativos

$$P(H_1=\text{True}) + P(H_0=\text{True}) = 1$$

Why Most Published Research Findings Are False

John P. A. Ioannidis

Consideremos estudios con nivel de α de 5%, poder estadístico ($1-\beta$) de 80% y con 50% de probabilidades de que H_1 sea verdadera

$$PPV = \frac{VP}{VP+FP} = 94\%$$

¿Cuántos tipos de resultados espero obtener?

	H_1 Verdadera	H_0 Verdadera
Resultado Significativo	40% Verdaderos Positivos	2.5% Falsos Positivos
Resultado No Significativo	10% Falsos Negativos	47.5% Verdaderos Negativos

Why Most Published Research Findings Are False

John P. A. Ioannidis

Definamos sesgo como la combinación de varios factores relacionados al diseño experimental, análisis de dato y presentación de resultados que tienen a producir resultados publicables que no deberían haber sido producidos.

- u es la proporción de esos resultados publicados como resultado de este sesgo

¿Cuántos tipos de resultados espero obtener?

	H_1 Verdadera	H_0 Verdadera
Resultado Significativo	$((1-\beta) + u\beta) P(H_1=T)$ Verdaderos Positivos	$(\alpha + u(1-\alpha)) P(H_0=T)$ Falsos Positivos
Resultado No Significativo	$\beta (1-u) P(H_1=T)$ Falsos Negativos	$(1-\alpha) (1-u) P(H_0=T)$ Verdaderos Negativos

Why Most Published Research Findings Are False

John P. A. Ioannidis

Ioannidis quantifica plausibilidad *a priori* de H_1 en función de:

- Odds (posibilidades) de *True* : *Not True*
- $R = \text{True} / \text{Not True}$

(Por ej. odds 1:2, con $R=1/2$)

$$P(H_1 = \text{True}) = R / (R + 1)$$

$$P(H_0 = \text{True}) = 1 / (R + 1)$$

Razón entre hipótesis
verdaderas y no
verdaderas

Sesgo

$$PPV = \frac{VP}{VP + FP}$$

$1 - \beta$	R	u	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

RCT, randomized controlled trial.

DOI: 10.1371/journal.pmed.0020124.t004

Why Most Published Research Findings Are False

John P. A. Ioannidis

Dependiendo del tipo de estudio podemos tener PPVs muy pobres

Razón entre hipótesis
verdaderas y no
verdaderas

Sesgo

$$PPV = \frac{VP}{VP + FP}$$

$1 - \beta$	R	u	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

RCT, randomized controlled trial.

DOI: 10.1371/journal.pmed.0020124.t004

Causas de baja replicabilidad

¿Fraude?

¿Más presión?

¿Más oportunidades?

¿Descuido o desconocimiento?

The Nine Circles of Scientific Hell

Neuroskeptic¹

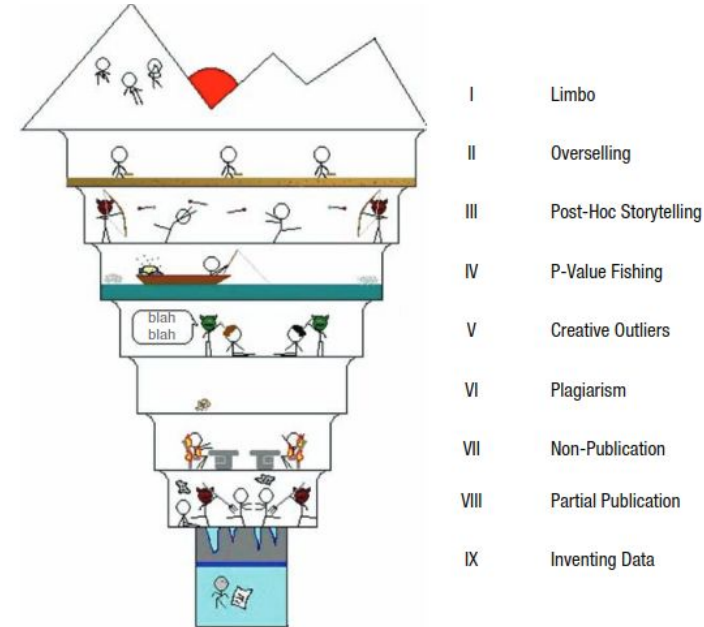


Fig. 1. The nine circles of scientific hell (with apologies to Dante and xkcd)

THE AMERICAN STATISTICIAN

2016, VOL. 70, NO. 2, 129–133

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

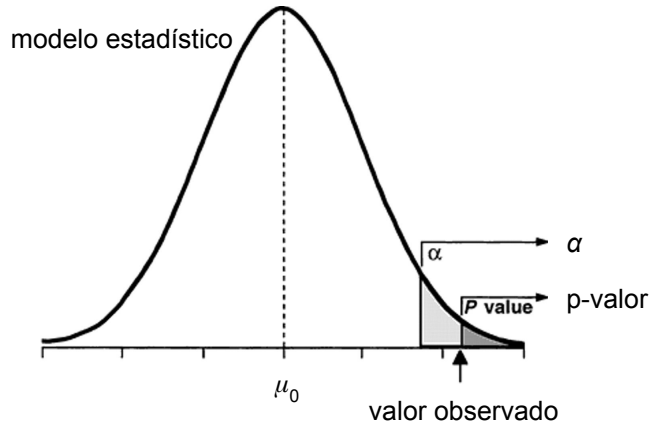
EDITORIAL

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

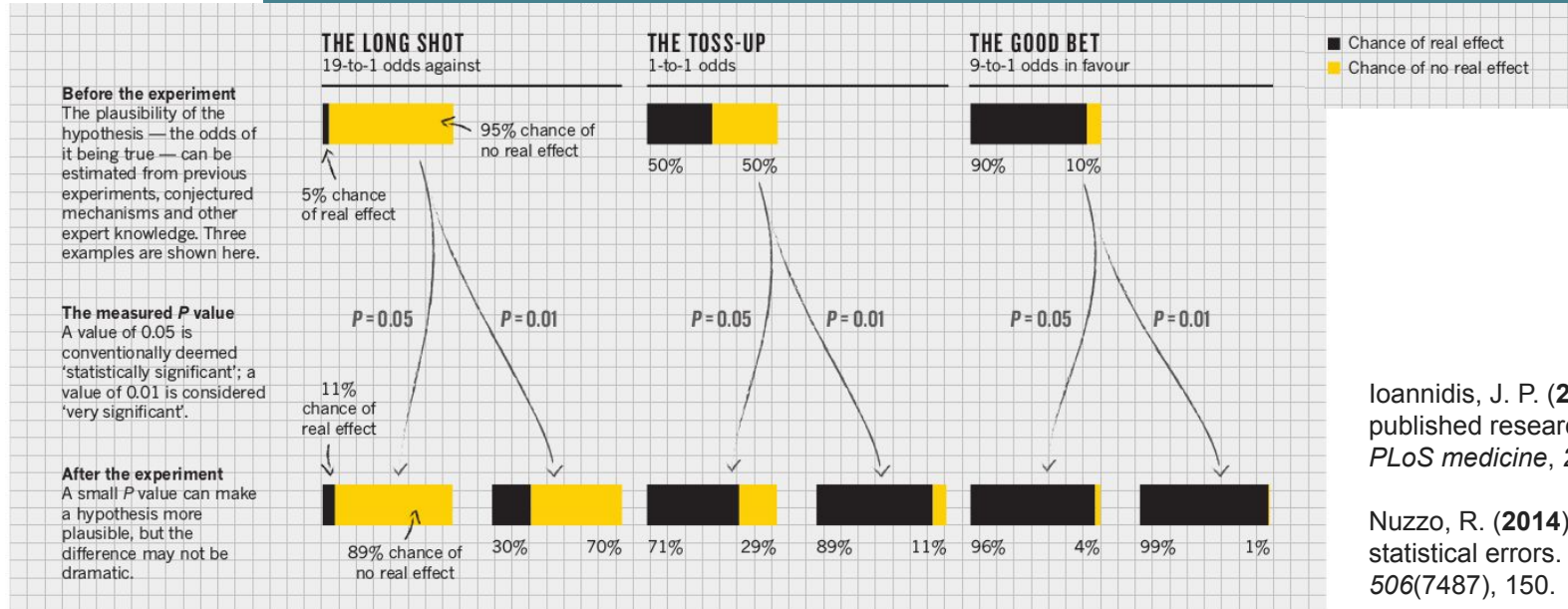
Los p-valores pueden indicar cuán incompatibles son los datos con un modelo estadístico dado.



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.



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?

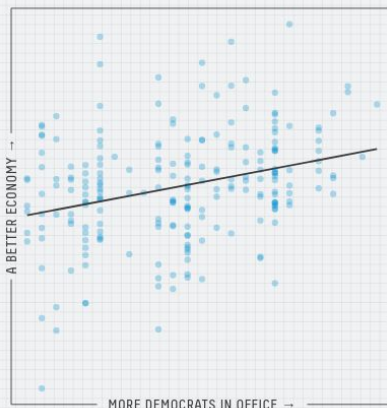
- ☐ Employment
- ☐ Inflation
- ☒ GDP
- ☒ Stock prices

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.

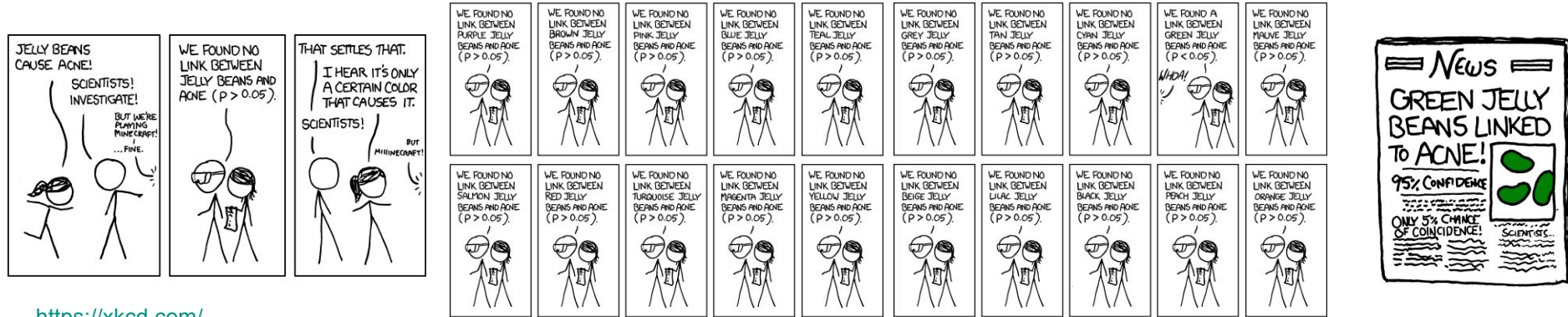
Las conclusiones científicas o políticas NO pueden basarse únicamente en si el p-valor pasa o no un umbral dado.

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.



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.

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.

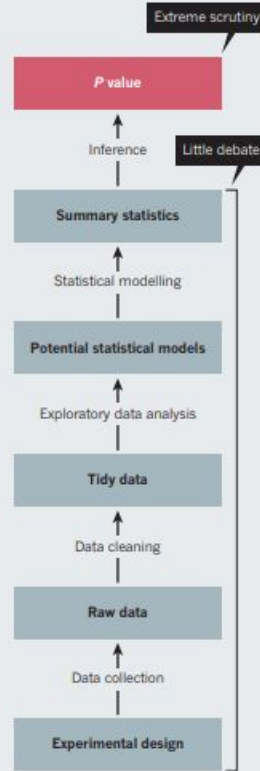
Wasserstein RL & Lazar NA (2016) “The ASA's Statement on p-Values: Context, Process, and Purpose”, The American Statistician, 0:2, 129-133

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.



Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature News*, 520(7549), 612.

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

¿Cuán expertos son los expertos?

Sobre los peligros de aplicar a ciegas paquetes de análisis estadísticos



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).

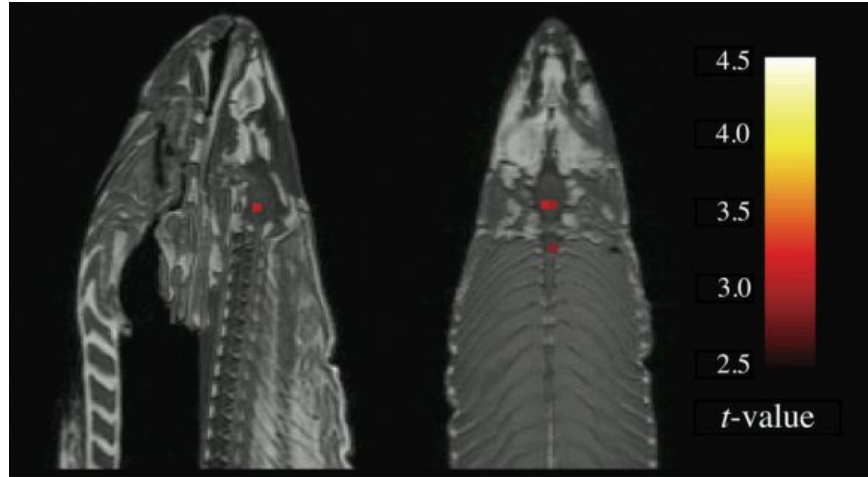
Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28), 7900-7905.

¿Cuán expertos son los expertos?

Respuestas neuronales de un salmón muerto a caras humanas felices o tristes

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.

<https://www.improbable.com/ig-about/winners/#ig2009>



Bennett CM, Baird AA, Miller MB and Wolford GL “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction” poster, 15th Annual Meeting of the Organization for Human Brain Mapping, San Francisco, CA, June 2009.

Bennett CM, Baird AA, Miller MB and Wolford GL “Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Multiple Comparisons Correction” Journal of Serendipitous and Unexpected Results, vol. 1, no. 1, 2010, pp. 1-5.

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.

¿Cuál fue la hipótesis H_1 evaluada?

- ¿Tauro está relacionado con fracturas del cuello del fémur? (o sea, 1 sola evaluación)
- ¿Algún signo está relacionado con alguna enfermedad? (o sea, múltiples evaluaciones)

“Data fishing”

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

Las correcciones tienen un costo: tienden a reducir la potencia estadística. ¿Cuándo hay que corregir?

Comparaciones múltiples

Las correcciones tienen un costo: tienden a reducir la potencia estadística. ¿Cuándo hay que corregir?

- Si basta que haya una comparación significativa para hacer una afirmación, **hay que corregir**
- Si la afirmación se hace solo cuando todas las comparaciones son significativas, **no hay que corregir**

Lo que tenemos que preguntarnos es: **¿cuál es la hipótesis que estamos evaluando?**

Comparaciones múltiples

Las correcciones tienen un costo: tienden a reducir la potencia estadística. ¿Cuándo hay que corregir?

- Si basta que haya una comparación significativa para hacer una afirmación, **hay que corregir**

“Algo voy a encontrar si comparo tal variable de respuesta para todos estos grupos”

- Si la afirmación se hace solo cuando todas las comparaciones son significativas, **no hay que corregir**

“Si comparo todos estos grupos tal variable de respuesta va a ser diferente”

Lo que tenemos que preguntarnos es: **¿cuál es la hipótesis que estamos evaluando?**

Comparaciones múltiples

Las correcciones tienen un costo: tienden a reducir la potencia estadística. ¿Cuándo hay que corregir?

Estamos ante:

- ¿un **estudio confirmatorio** con un conjunto de hipótesis bien definidas en un contexto argumentativo?
- ¿un **estudio exploratorio** en el que muchas variables predictoras son observadas y evaluadas?

(Nota: variable predictora = variable explicativa = variable independiente)

Comparaciones múltiples

¿Cuándo hay que corregir?

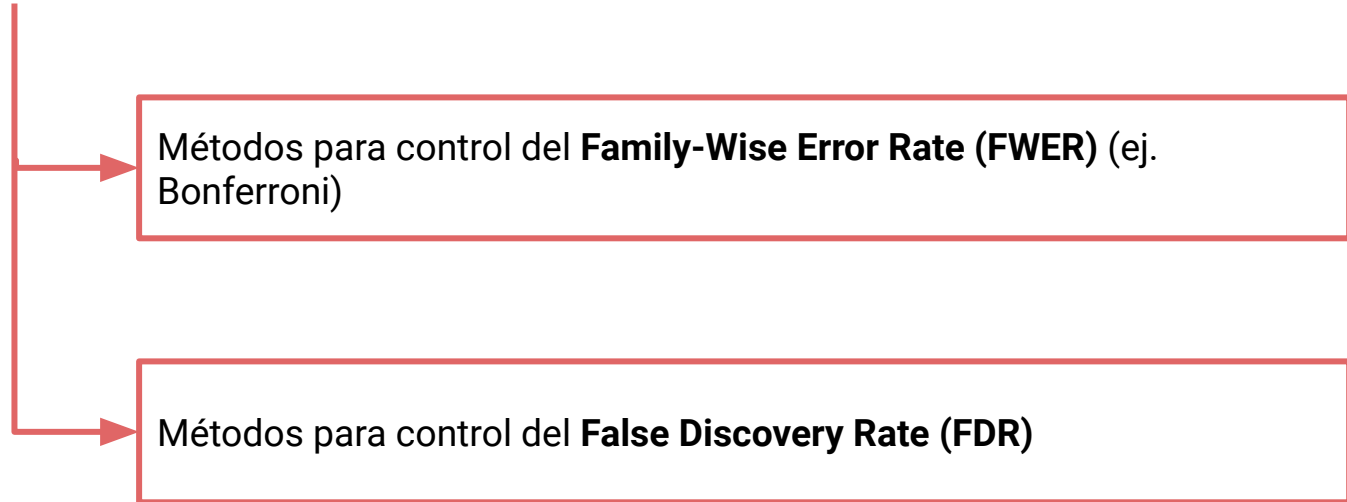
Muchos tests para diferentes hipótesis. **No es necesario corregir.**

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

Correcciones para comparaciones múltiples

Queremos ajustar los p-valores para compensar las comparaciones múltiples



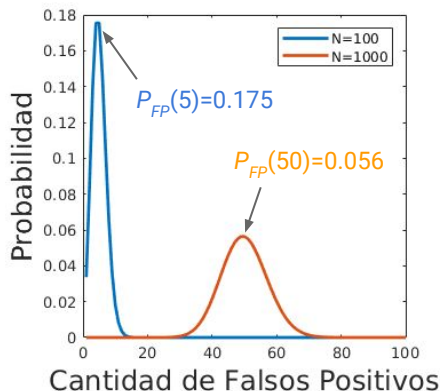
Probabilidad de Falsos Positivos en Comparaciones Múltiples

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)

Para $\alpha = 0.05$:



Si N es grande y α es chico:

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

(distribución Poisson)

Para N moderado la probabilidad de obtener $\alpha\%$ de falsos positivos es mucho mayor al caso asintótico

FWER: Bonferroni

“Quiero acotar la probabilidad de cometer al menos un error de tipo 1 (falso positivo)”

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

\swarrow
 $k=0$

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

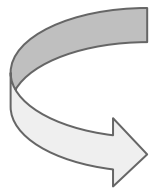
Entonces, la probabilidad de cometer al menos uno de estos errores es: $\pi = 1 - (1-\alpha)^N$

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

Despejo α y pido que π sea chico

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

Si x es chico



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

$$\alpha = \pi/N$$

Si quiero $\pi = 0.05$ para el experimento y $N=100$, entonces α de cada test debe ser: 0.0005

¡Baja mucho la potencia del test!

FWER: Holm, Hochberg, Hommel, ...

Opciones menos conservadoras que la de Bonferroni

Método de Holm

Ordenar los N p-valores de todos los tests de menor a mayor

Repetir para $n = 1$ hasta N

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

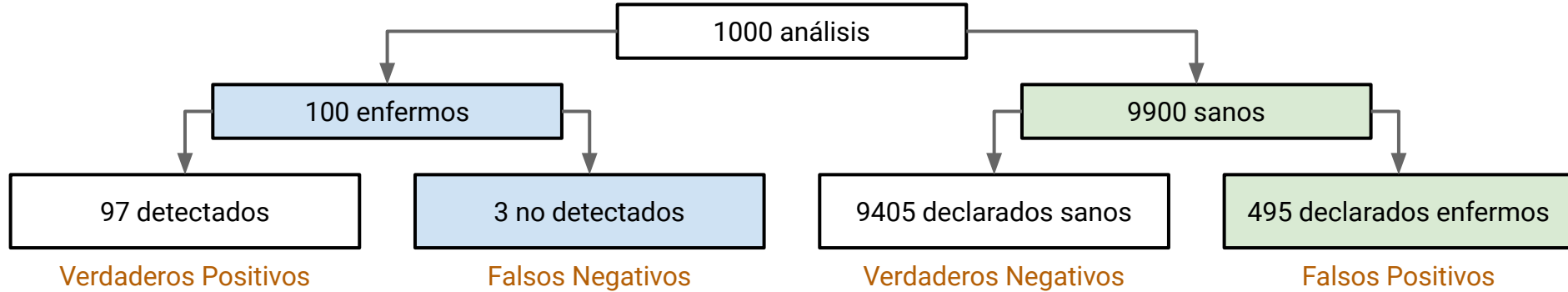
Es decir, para el menor p-value usar $\alpha=\pi/N$, para el siguiente $\alpha=\pi/(N-1)$, ...

Método de Hochberg

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



Tasa de Falsos Descubrimientos (FDR)



$$\begin{aligned} \text{Tasa de falsos descubrimientos (False Discovery Rate)} &= \frac{\text{Falsos Positivos}}{\text{Casos Positivos}} = \frac{\text{Falsos Positivos}}{\text{Falsos Pos.} + \text{Verdaderos Pos.}} = \frac{495}{495 + 97} = 83.6 \% \end{aligned}$$

Si hay poca prevalencia (enfermos/sanos), habrá pocos Verdaderos Positivos y la FDR puede ser alta

FDR

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

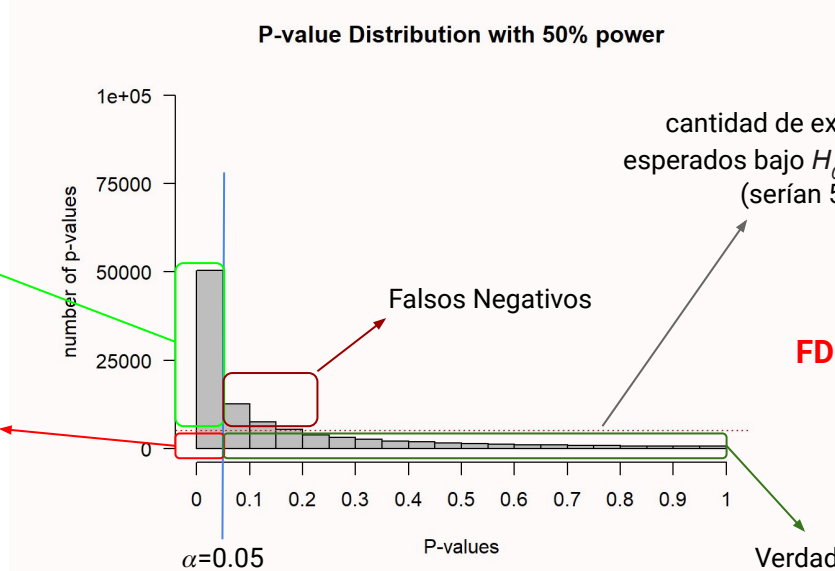
Son los métodos que suelen usarse cuando el N es muy grande y los FWER se vuelven muy restrictivos.

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.

100.000 experimentos
con poder de 50%



$$\text{FDR} = \frac{\text{Falsos Positivos}}{\text{Positivos}} = \frac{5000}{50000} = 0.1$$

Ajusto α para llegar a un FDR deseado

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.”

Hay diferentes implementaciones, por ejemplo Benjamini-Hochberg

Algunas ideas asociadas al muestreo

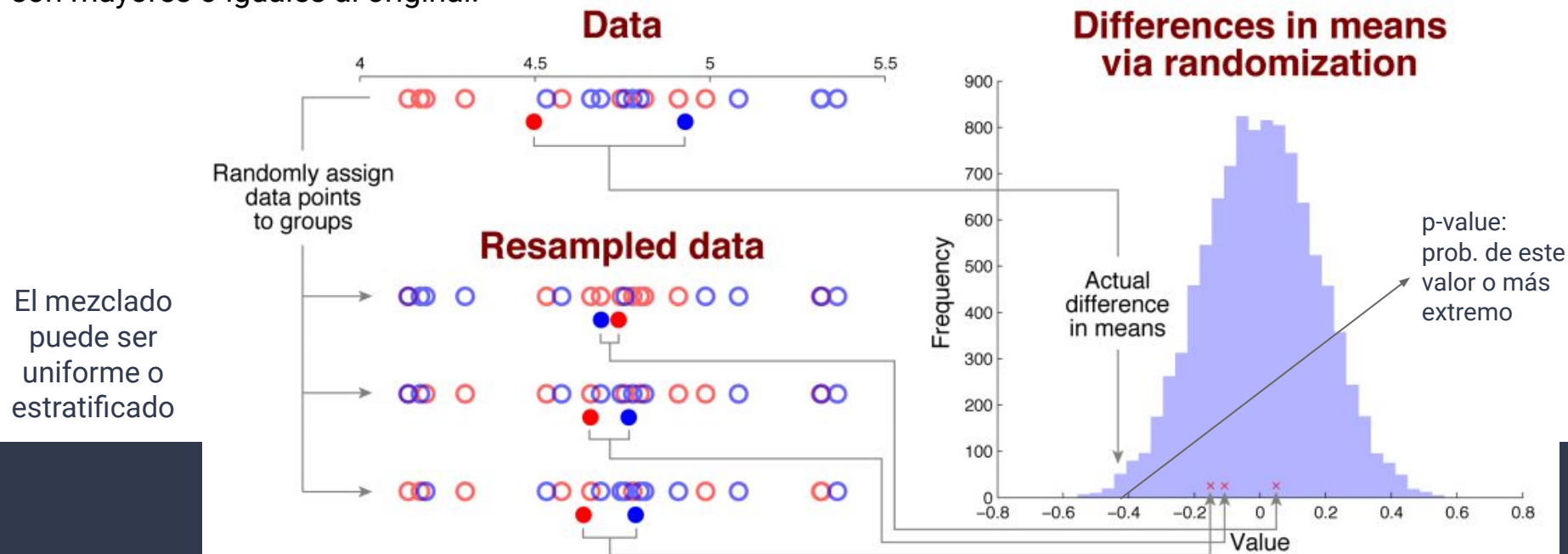


Randomization, Permutations (sin reposición)

Orientado a hacer un test de hipótesis “data-driven”, no paramétrico (sin suposiciones)

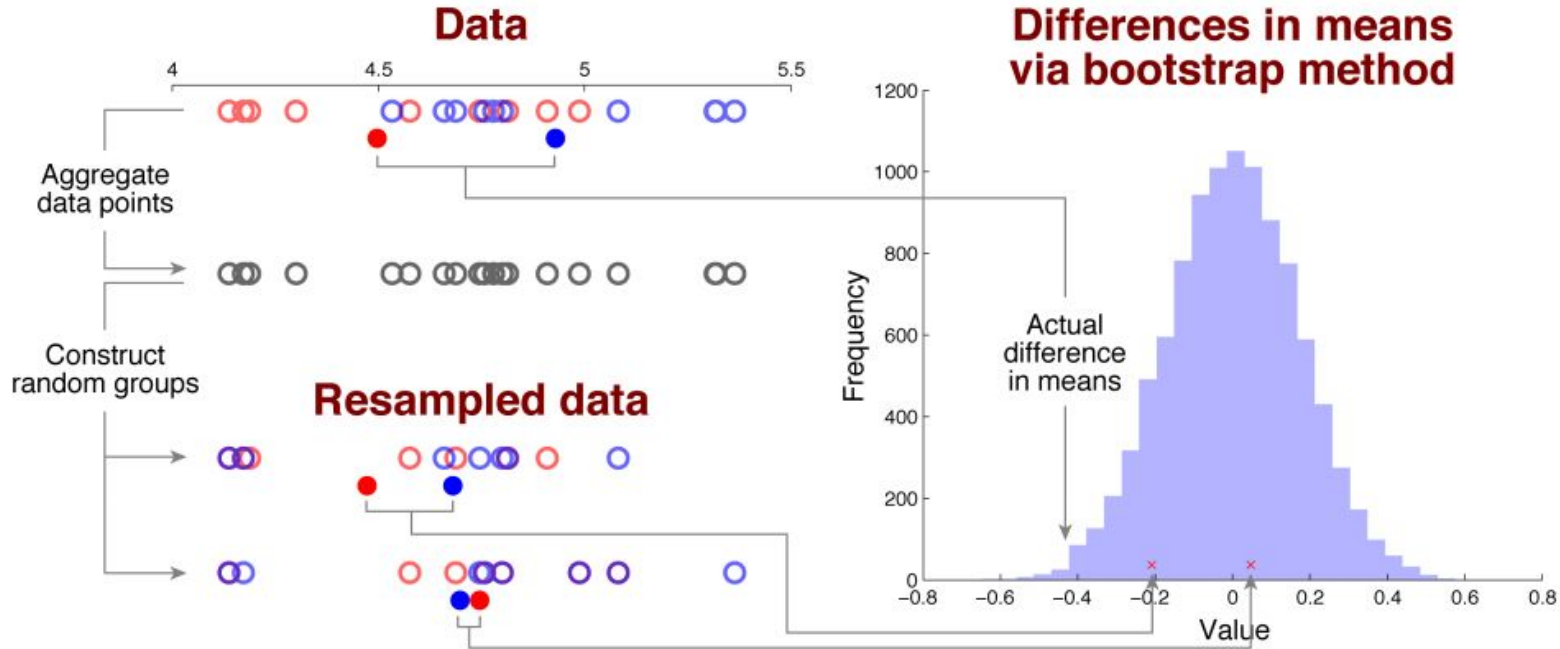
Hipótesis nula: **Rojos** y **Azules** provienen de la misma distribución. ¿Cuál es la probabilidad de tomar dos subconjuntos que den una diferencia de medias mayor o igual a la original?

Idea: Mezclo etiquetas **rojas** y **azules**, me genero una distribución de diferencias de medias y me fijo cuales son mayores o iguales al original.



Resampling, Bootstrapping (con reposición)

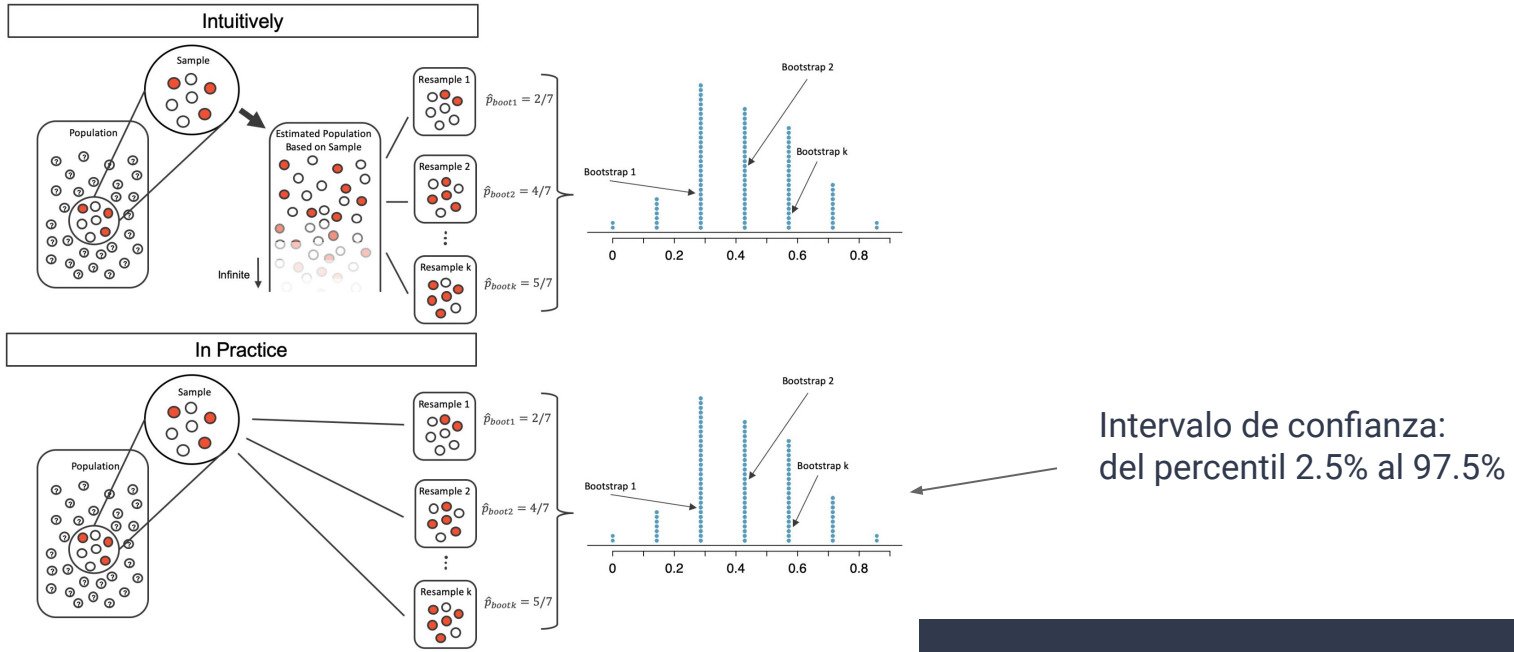
Orientado a cuantificar variabilidad e inferir intervalos de confianza, también de forma no paramétrica



Resampling, Bootstrapping (con reposición)

Orientado a cuantificar variabilidad e inferir intervalos de confianza, también de forma no paramétrica

¿Cómo inferir un rango de confianza si tengo una muestra única?



- 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.

Algunos ejemplos (pero que valen
en general): TP2 2022

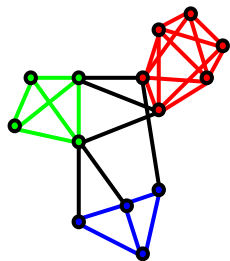
A dark blue, curved, triangular shape that starts from the bottom left corner and extends diagonally upwards towards the right, filling the bottom half of the slide.

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p -value.” (Tagliazucchi, et al., 2013)

Sujeto 1 (S1), Wake (W)

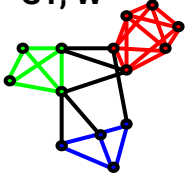


Permutaciones + Rand Index

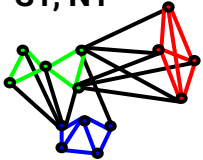
Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p -value.” (Tagliazucchi, et al., 2013)

S1, W



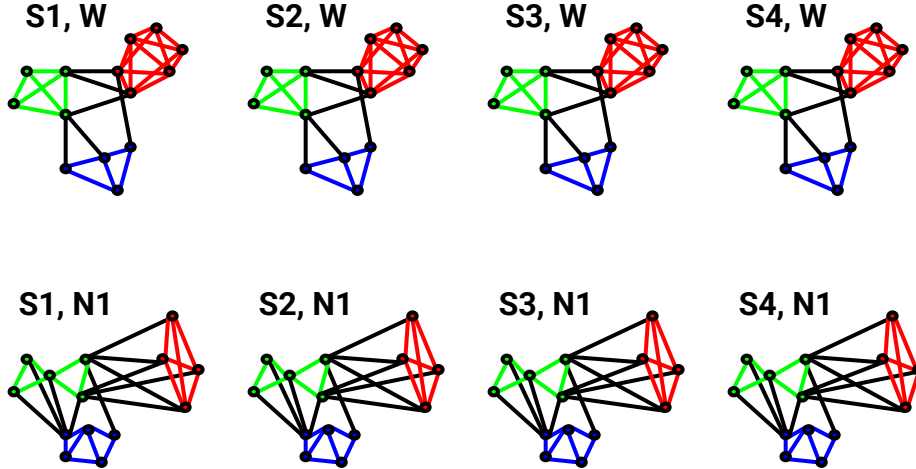
S1, N1



Permutaciones + Rand Index

Statistical testing

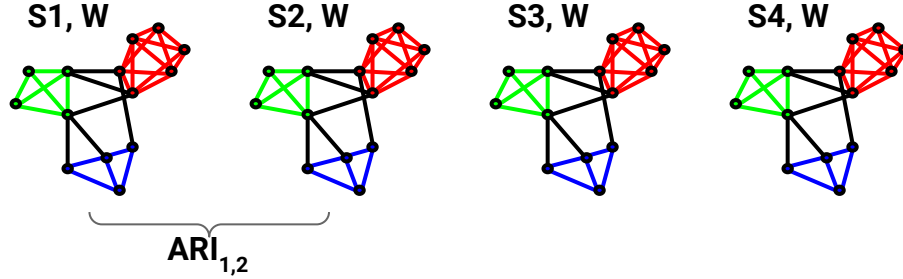
“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)



Permutaciones + Rand Index

Statistical testing

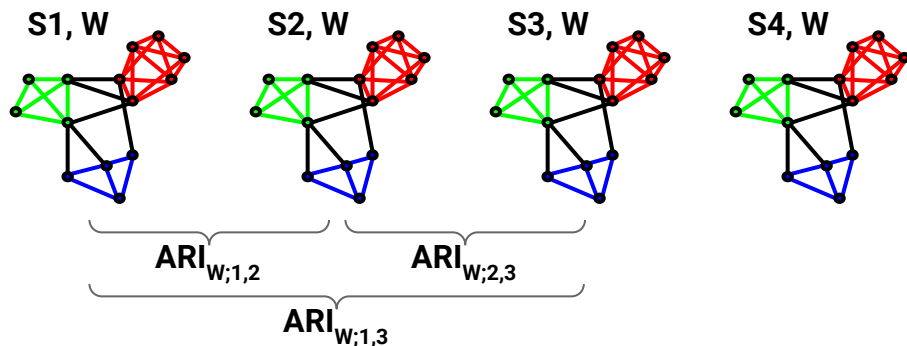
“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)



Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)



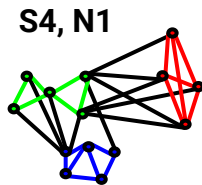
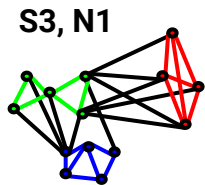
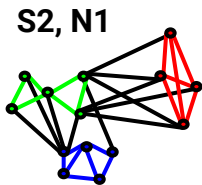
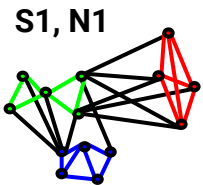
ARI_W = “...within-group similarity (computed with the adjusted-for-chance Rand index) of the real data...” =
Promedio de todos los pares dentro de Wake.

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)

ARI_{N1} = “...within-group similarity (computed with the adjusted-for-chance Rand index) of the real data...” =
Promedio de todos los pares dentro de N1.



Permutaciones + Rand Index

Statistical testing

*“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)*

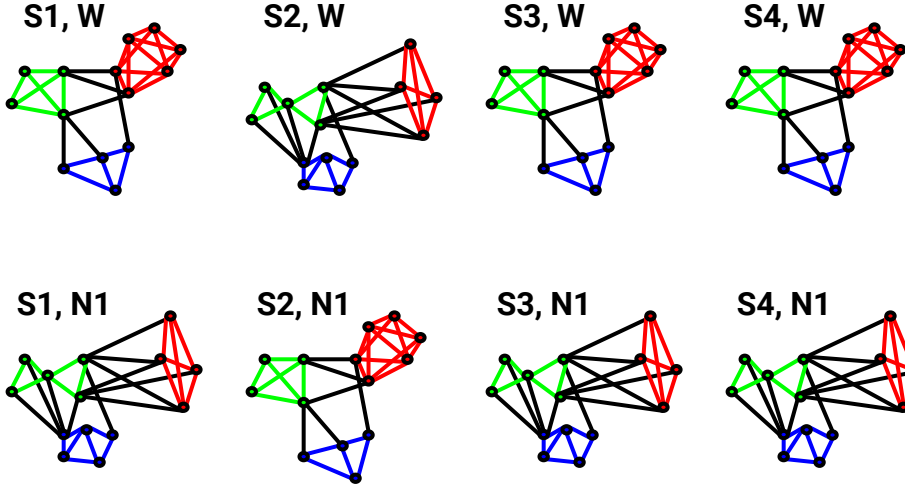
“...within-group similarity (computed with the adjusted-for-chance Rand index) of the real data...” =
Promedio de ARI_w y ARI_{N1} para los datos reales.

•

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)

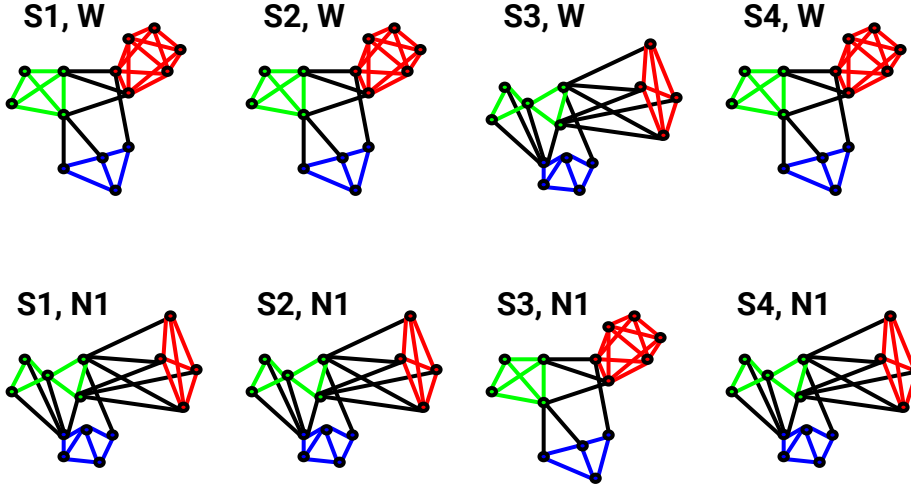


“... permuted data (i.e. randomized data from both groups)...”

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)

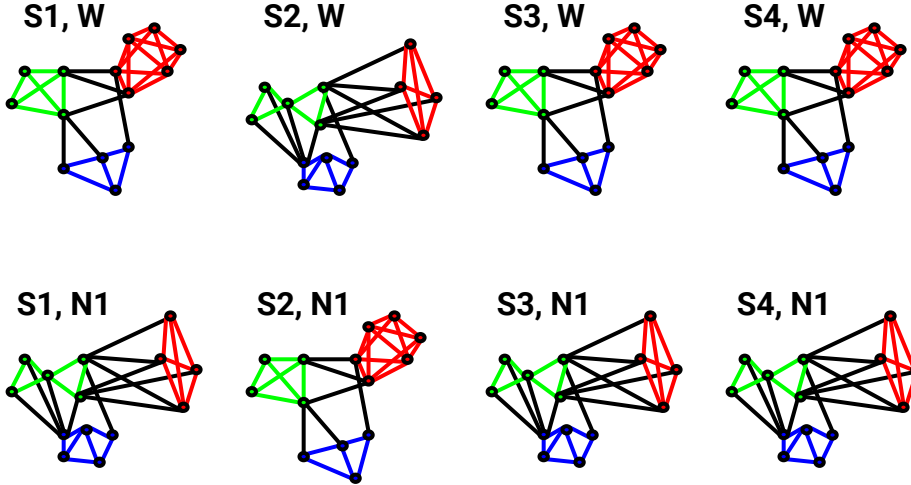


“... permuted data (i.e. randomized data from both groups)...”

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)



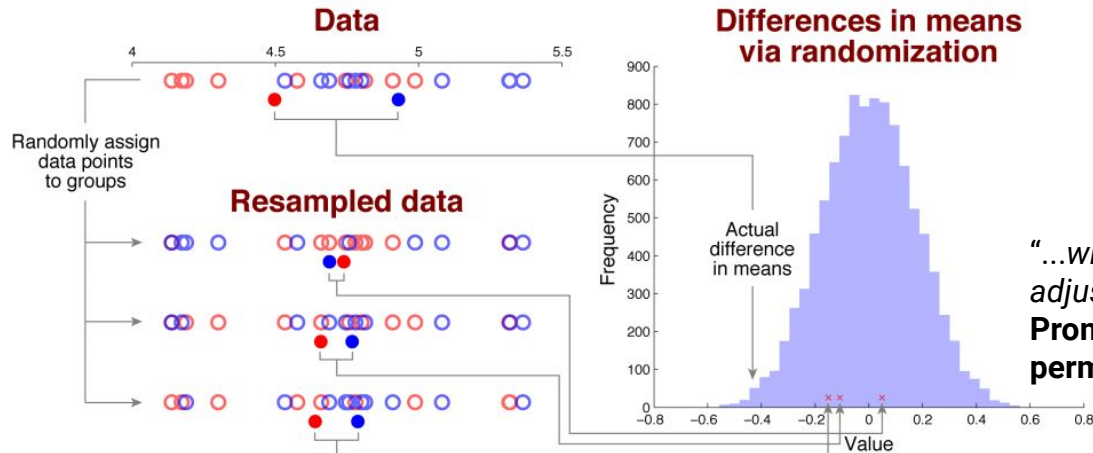
“... permuted data (i.e. randomized data from both groups)...”

“...within-group similarity (computed with the adjusted-for-chance Rand index) of the permuted data...” = Promedio de ARI_w y ARI_{N1} para cada una de las permutaciones.

Permutaciones + Rand Index

Statistical testing

“... Statistical significance of module membership differences was assessed using a permutation testing procedure, as introduced by Alexander-Bloch and colleagues (**Alexander-Bloch et al., 2012**). Briefly, the within-group similarity (computed with the adjusted-for-chance Rand index) of the real data and permuted data (i.e. randomized data from both groups) was computed. The number of instances in which the within-group similarity of the non-permuted data exceeded that of the permuted data was divided by the number of permutations (in this case, 1000) to obtain the p-value.” (Tagliazucchi, et al., 2013)



“...within-group similarity (computed with the adjusted-for-chance Rand index) of the permuted data...” = **Promedio de ARI_w y ARI_{N1} para cada una de las permutaciones.**