

WARNING: P-hacking

Juan E. Kamienkowski (juank@dc.uba.ar)

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

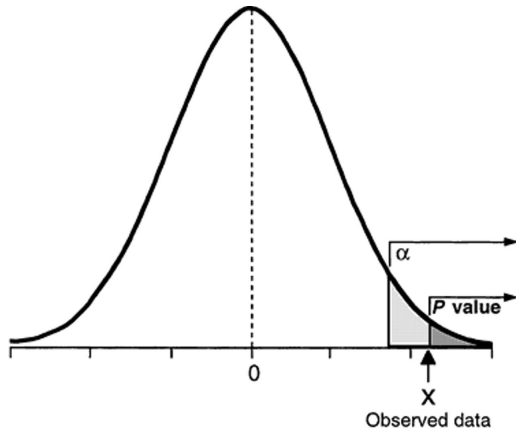
P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

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
≥ 0.1	THIS INTERESTING SUBGROUP ANALYSIS

<http://phdcomics.com/>

¿A qué nos referimos como p-hacking?

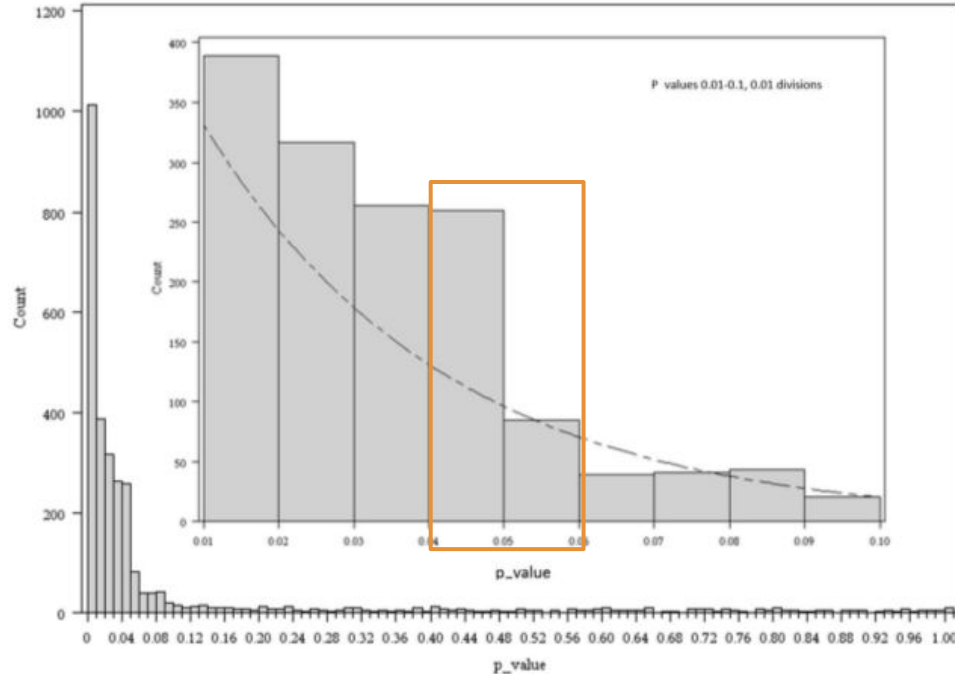


Fig.2 All p values between 0 and 1 are plotted in the *bottom graph*. The *inset* shows p values between 0.01 and 0.1 in 0.01 divisions

<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?

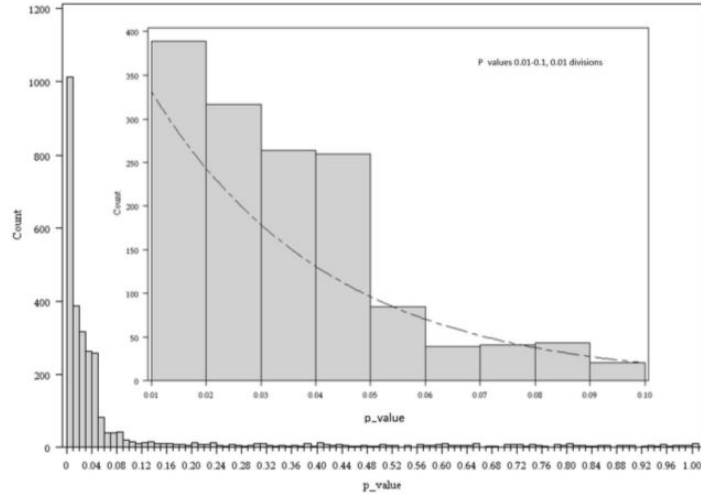
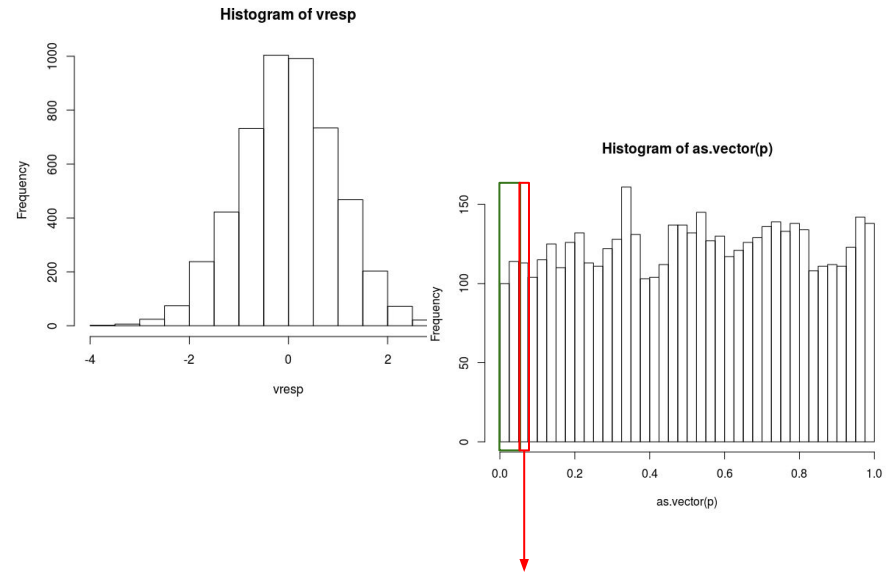


Fig. 2 All p values between 0 and 1 are plotted in the bottom graph. The inset shows p values between 0.01 and 0.1 in 0.01 divisions



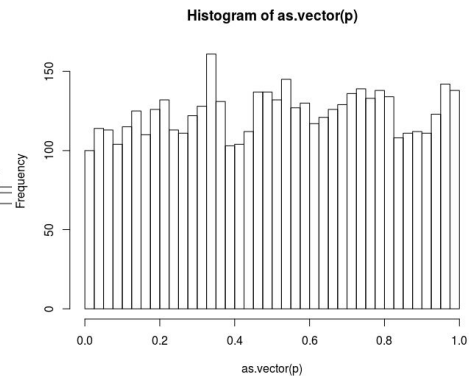
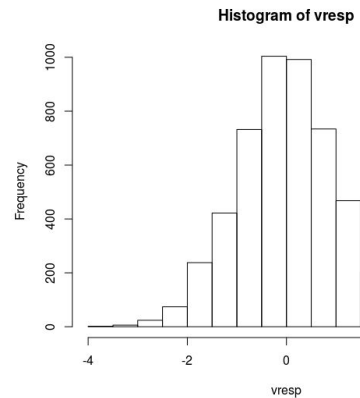
¿Qué pasa si obtengo un resultados con p entre 0.05 y 0.075 y agrego entre algunos ejemplos más? (para decidir si es significativo o no)...

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

```
## Ejemplo 1
rm(list = ls())
nreg <- 5000
vresp <- rnorm(nreg, 0, 1)
hist(vresp)

nsamp = 100
ssamp = 50
vc <- sample(vresp, ssamp, replace = FALSE)
vresp.df <- data.frame(vc)
for(i in 2:nsamp){
  vc <- sample(vresp, ssamp, replace = FALSE)
  vresp.df <- cbind(vresp.df, vc)
}
names(vresp.df) <- paste0("muestra", 1:100)
str(vresp.df)

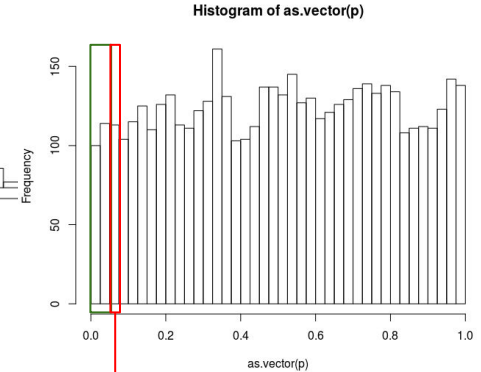
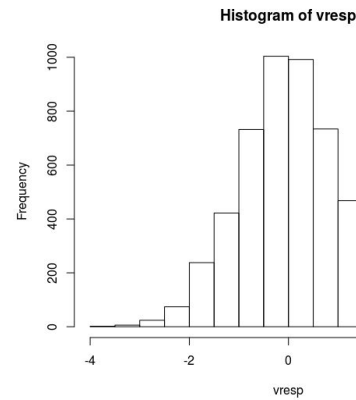
p <- matrix(NA, nsamp, nsamp)
for(i in 1:(nsamp-1)){
  for(j in (i+1):nsamp){
    p[i,j] <- t.test(vresp.df[,i], vresp.df[,j])$p.value
  }
}
hist(as.vector(p), seq(0,1,by=0.025))
```



```
image(1:nsamp,1:nsamp,p,c(0,1),col=gray.colors(20))
```

```
ind <- which(p>0.05 & p<0.075,arr.ind = T)
i <- sample.int(dim(ind)[1],1)
i1 = ind[i,1]
i2 = ind[i,2]
p[i1,i2]
```

```
p2 <- matrix(NA,nsamp,1)
for(i in 1:nsamp){
  p2[i]<-t.test(c(vresp.df[,i1],sample(vresp, 5, replace = FALSE)),
               c(vresp.df[,i2],sample(vresp, 5, replace = FALSE)))$p.value
}
100*sum(p2<0.05)/length(p2)
```



¿Qué pasa si obtengo un resultados con p entre 0.05 y 0.075 y agrego entre 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?

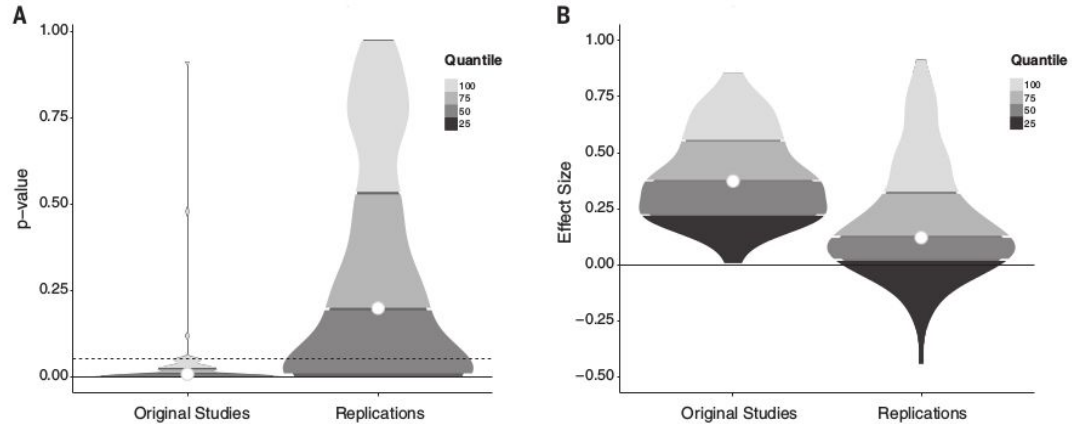


Fig. 1. Density plots of original and replication *P* values and effect sizes. (A) *P* values. (B) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

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>

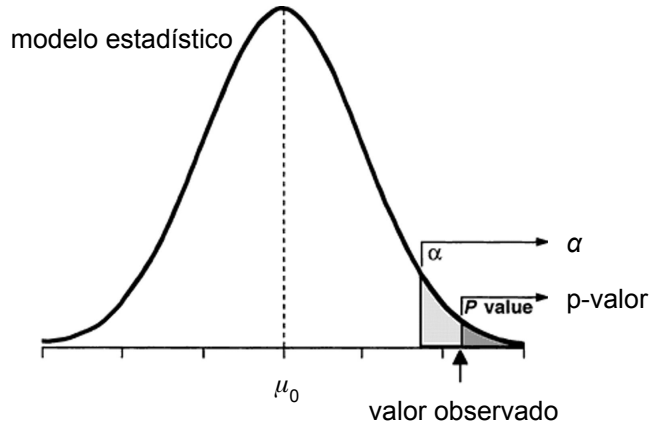
EDITORIAL

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?

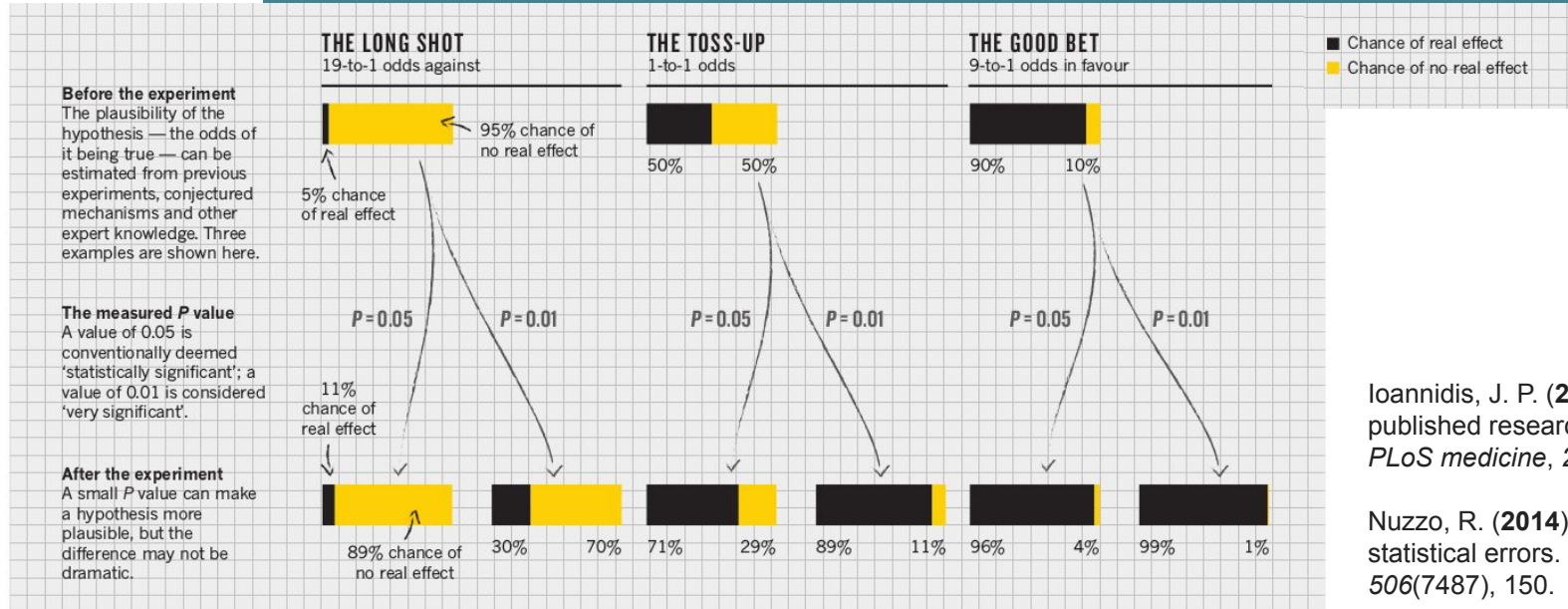
Los p-valores pueden indicar cuán incompatibles son los datos con un modelo estadístico dado.



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?

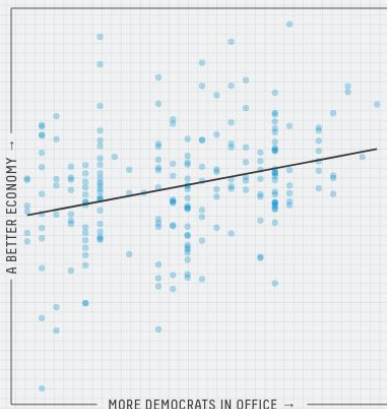
- ☐ Employment
- ☐ Inflation
- ☒ GDP
- ☒ Stock prices

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.

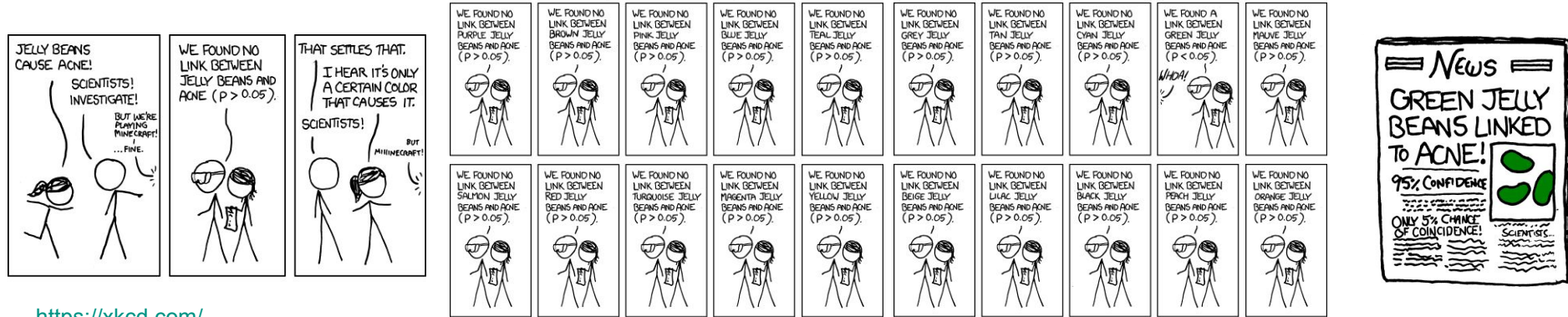
Las conclusiones científicas o políticas NO pueden basarse únicamente en si el p-valor pasa o no un umbral dado.

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

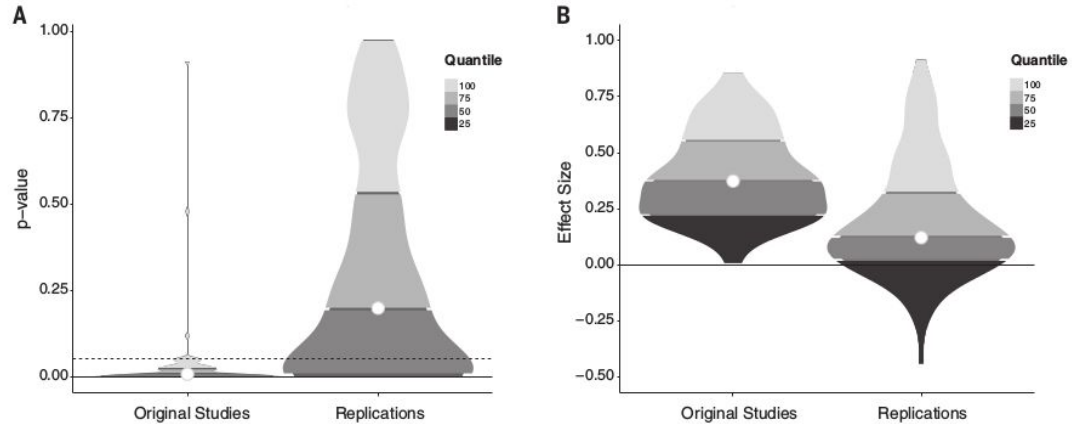


Fig. 1. Density plots of original and replication P values and effect sizes. (A) P values. (B) Effect sizes (correlation coefficients). Lowest quantiles for P values are not visible because they are clustered near zero.

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

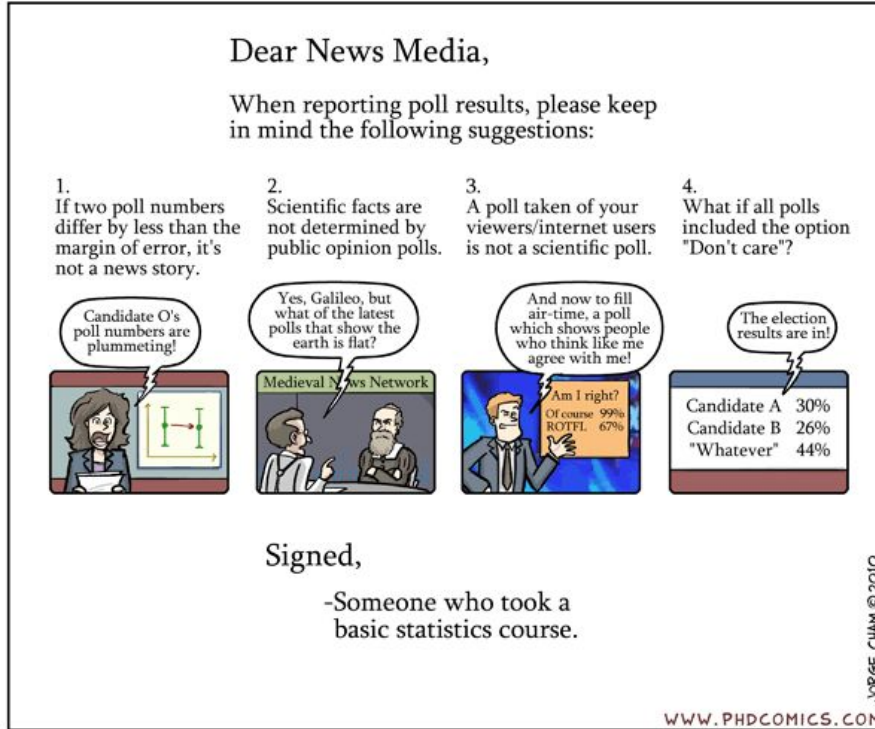
Ten Simple Rules for Effective Statistical Practice

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.



¿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*?

