P-hacking

Contame algo que me guste escuchar

Marcelo A. Soria
Facultad de Agronomía y
Maestría de Explotación de datos FCEN-FI
UBA

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

is characteristic of the field and vary a lot depending on whethe field targets highly likely relatic or searches for only one or a fe true relationships among thous

Why Most Published Research Findings Are False.
John P. A. Ioannidis, 2005, PLoS Med 2(8): e124.

AMSTATNEWS

The Membership Magazine of the American Statistical Association

nature International weekly

Home | News & Comment | Research | Career

Archive > Volume 519 > Issue 7541 > Rese

HOME ABOUT EDITORIAL CALENDAR PDF ARCHIVES ADVERTISE

Home » Additional Features, News and Announcements

ASA Releases 'Statement on Statistical Significance and *P*-Values'

7 MARCH 2016 736 VIEWS 2 COMMENTS

The ASA "Statement on Statistical Significance and *P*-Values" includes six principles underlying the proper use and interpretation of the *p*-value and is intended to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement is published in *The American Statistician*, along with more than a dozen discussion papers to provide further

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

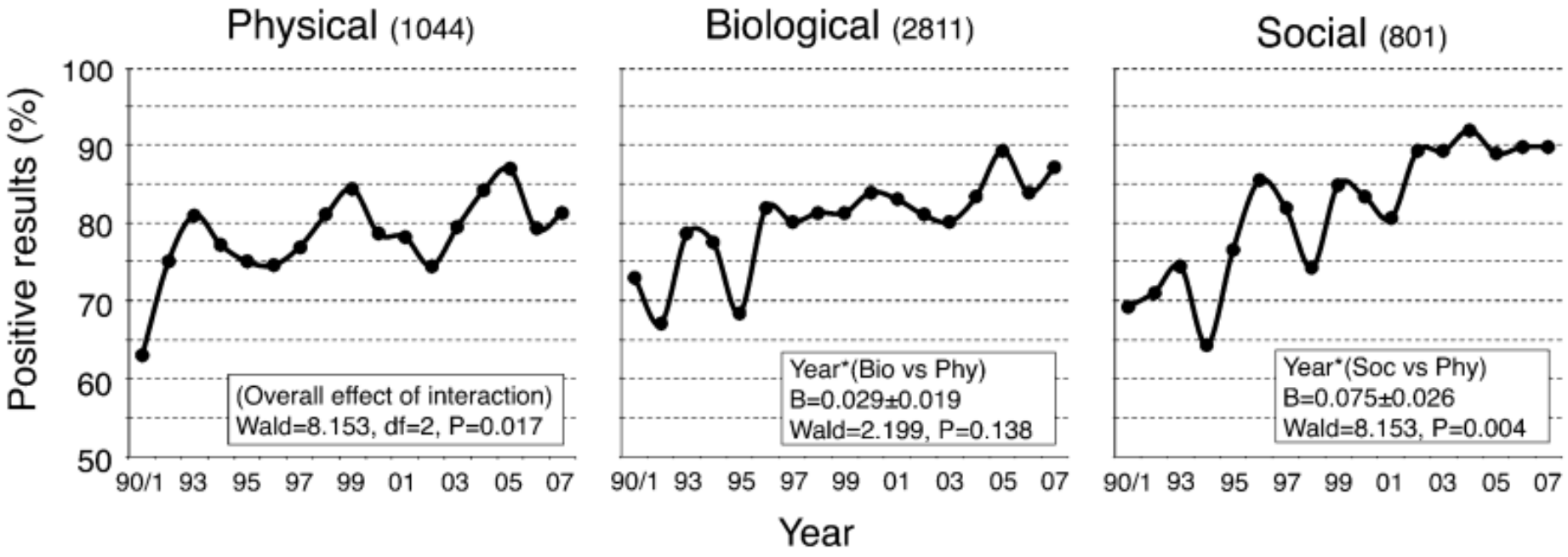
Test for reliability of results 'too easy to pass', say editors.

Uno de los componentes de este problema es la manipulación de los tests estadísticos.

Específicamente cómo se obtiene e informa la significancia estadística.

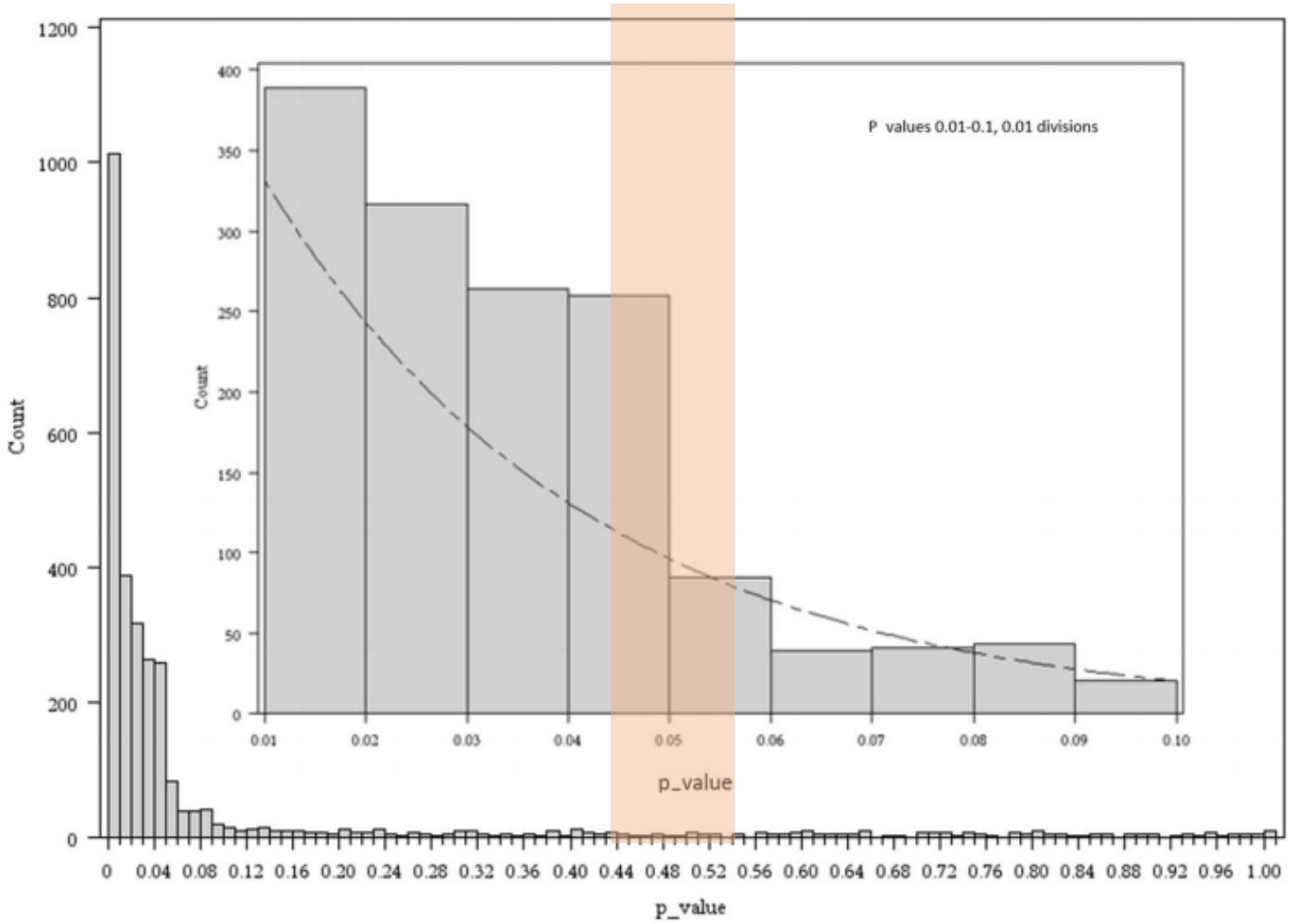
Esto se llama **P-hacking**

Otro resultado preocupante es que la cantidad de resultados positivos está aumentando en el tiempo



En la práctica esto significa que al no reportarse resultados nulos, otro grupo puede intentar reproducir el experimento, gastando tiempo y recursos

La distribución de los valores P en la literatura médica muestra una diferencia muy marcada entre aquellos inmediatamente inferiores y superiores a 0.05

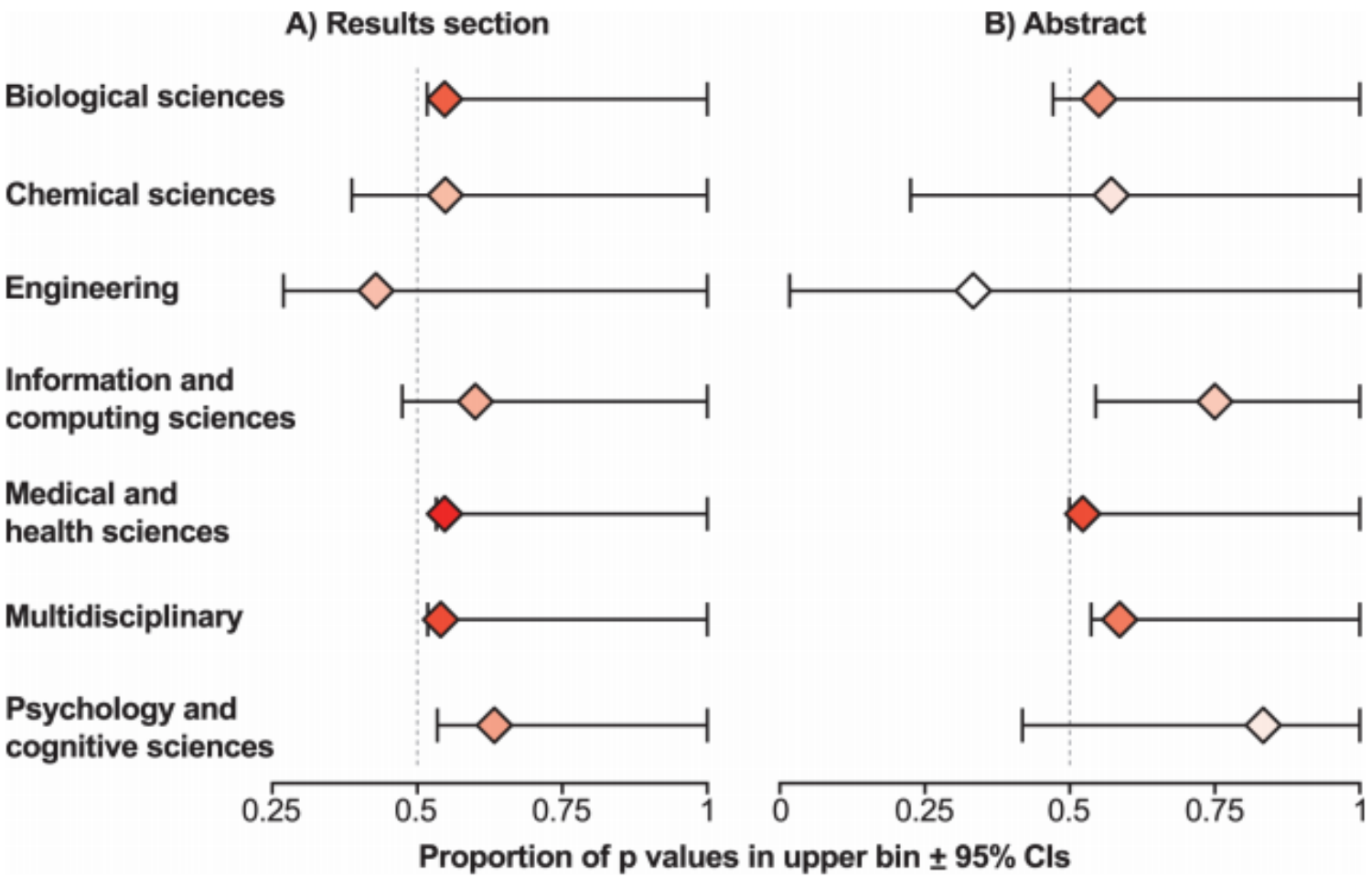


Ginsel et al. BMC Res Notes (2015) 8:721 DOI 10.1186/s13104-015-1691-x

Estas tendencias son más marcadas en algunas disciplinas,

…y paises !!

Frecuencias de valores P en diferentes áreas que se ubican en el rango 0.045 < P < 0.05



The Extent and Consequences of P-Hacking in Science
PLOS Biology | DOI:10.1371/journal.pbio.1002106

Fanelli y Ioannidis (2013) analizaron 1174 resultados de investigaciones en salud publicados en 82 meta-análisis y encontraron:

Los estudios sobre aspectos del comportamiento tienden a informar efectos más extremos que los biológicos.

Los trabajos con al menos un autor basado en Estados Unidos tienden a incluir resultados más extremos

http://www.pnas.org/cgi/doi/10.1073/pnas.1302997110

¿Fraude?

¿Más presión?

¿Más oportunidades?

¿Descuido / desconocimiento?

FiveThirtyEight Science

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

http://fivethirtyeight.com/features/science-isnt-broken/

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [ ] Senators
- [x] Representatives

How do you want to measure economic performance?

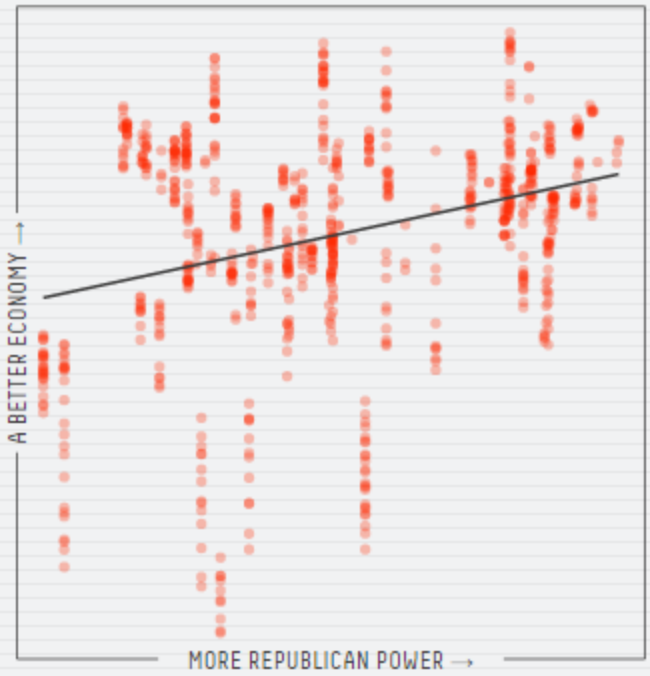
- [x] Employment
- [x] Inflation
- [ ] GDP
- [ ] Stock prices

Other options

- [x] **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 Republicans are in power? 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 **Republicans** 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.

Algunos trucos con los que nos puede engañar un p-hacker

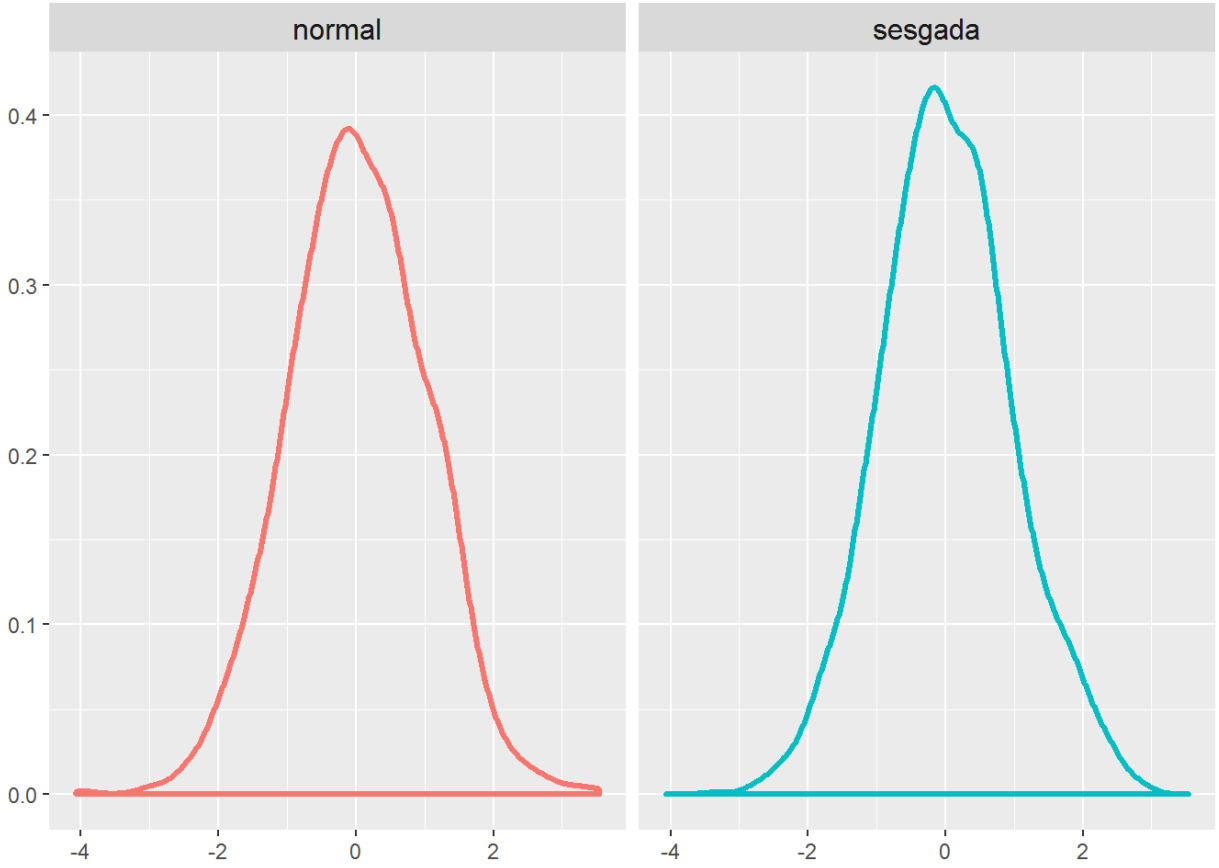
- Analizar muchas métricas, informar solo aquellas que “dan bien”
- Dejar de recolectar datos cerca del valor crítico (asumamos $P_c = 0.05$)
- Excluir casos (y llamarlos outliers) para alcanzar $P < 0.05$

- Transformar los datos para contar una “linda” historia, y significativa
- Registrar muchas variables, informar solo aquellas con $P < 0.05$
- Usar covariables para alterar los valores de P

Algunos ejemplos …

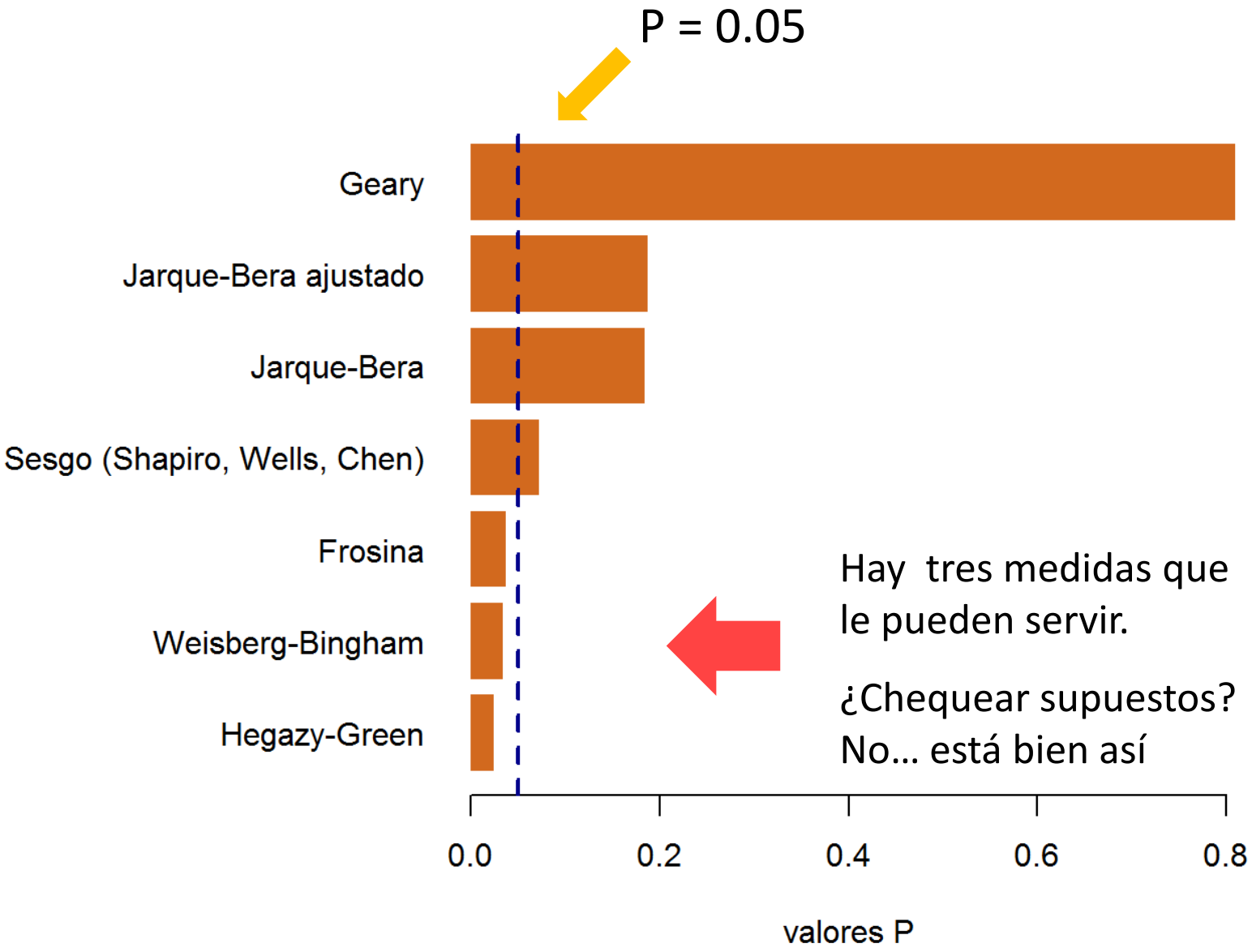
La distribución de la derecha tiene un sesgo leve a la derecha.

Pero la muestra es chica y la asimetría no se ve bien.



Nuestro P-hacker va a demostrar ese sesgo con algún test de normalidad y no piensa buscar más datos

Prueba varias medidas y se queda con la más conveniente:



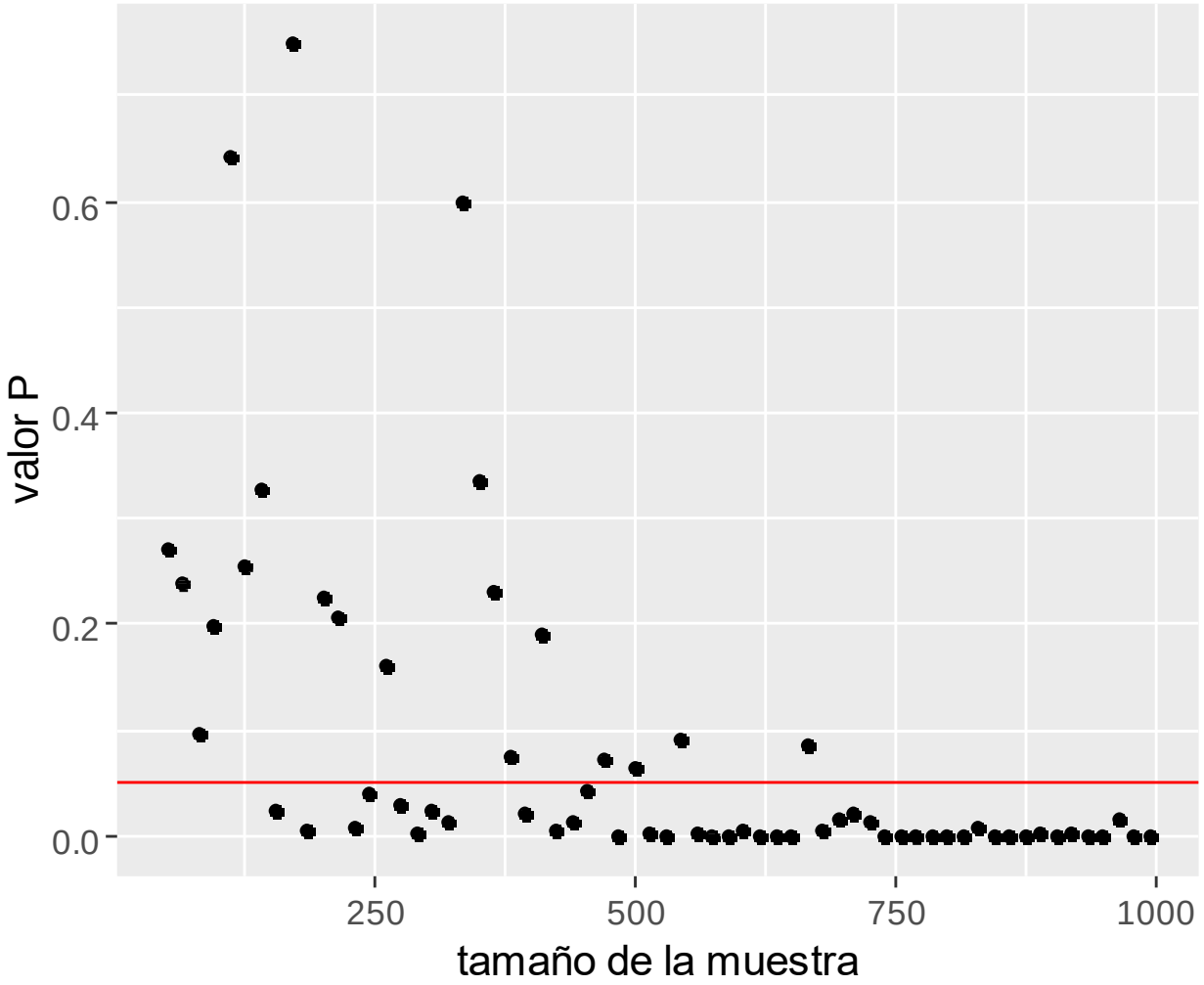
Nuestro P-hacker ahora enfrenta un desafio inverso.

Tiene dos poblaciones que son diferentes y es más cómodo considerar que hay solo una.

Lo que va a hacer es calcular tests de t y jugar con la potencia.

Prueba muestras de tamaño chico y las aumenta hasta que los valores P empiecen a ser significativos.

Valores P obtenidos con muestras de tamaño creciente



Zona de “confort” del P-hacker

El P-hacker tiene este conjunto de datos:

- Una variable de respuesta que son datos extraidos al **azar** de una distribución uniforme
- Las variables explicatorias son 100, todas de tipo lógico (Verdadero / Falso). Y también son extracciones al **azar** de distribuciones binomiales

vresp	feat1	feat2	feat3	feat4	feat5	feat6	feat7	feat8	feat9	feat10
46.05	0	1	1	1	1	0	0	1	0	0
74.82	1	0	0	0	0	0	0	1	0	0
86.19	1	0	0	0	0	0	1	1	1	0
14.90	1	0	0	0	0	1	0	1	1	0
3.76	1	0	0	0	1	1	1	1	0	0
79.89	0	1	0	0	0	0	1	1	1	0

Al P-hacker le encargan encontrar variables asociadas con la respuesta.

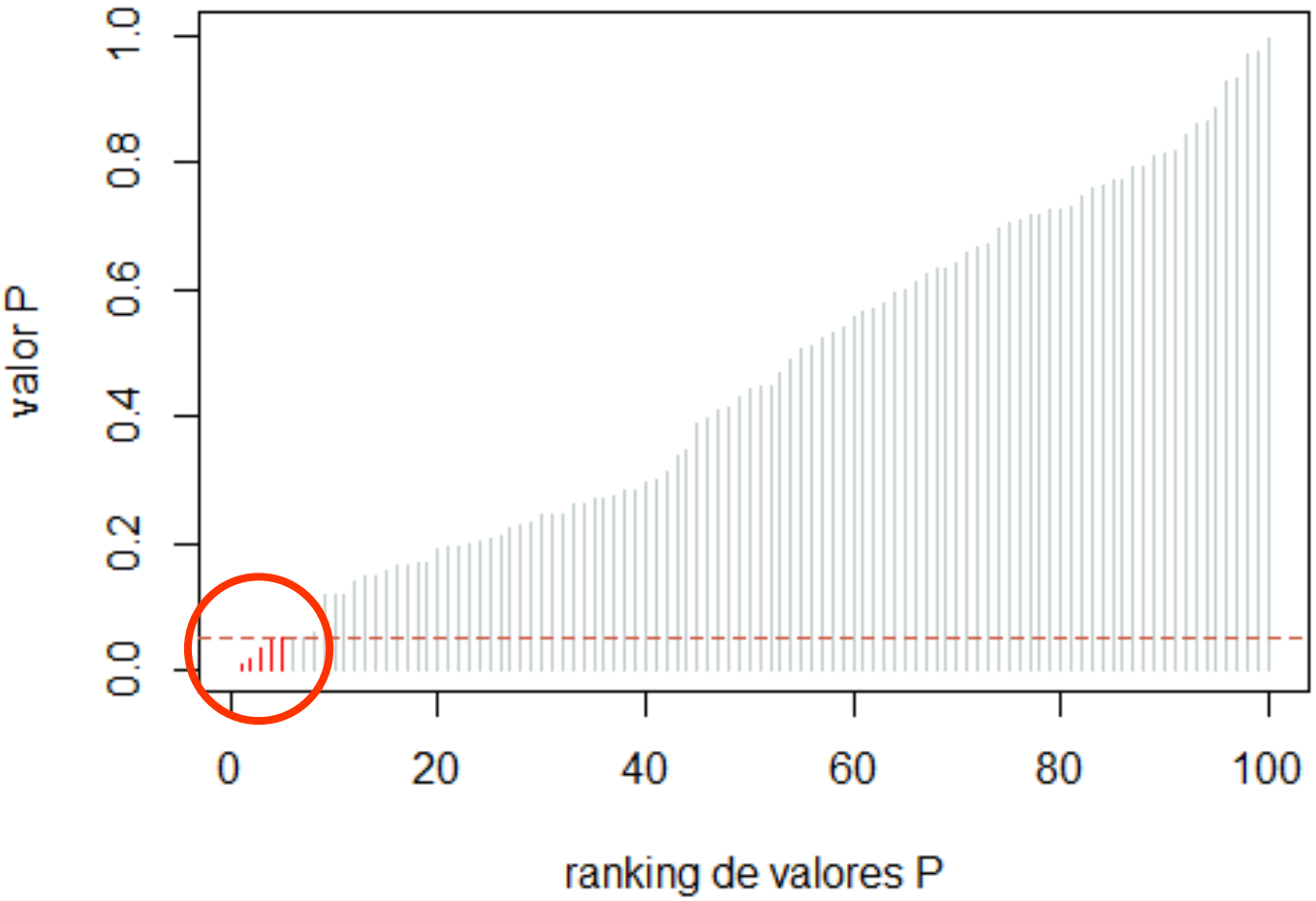
¿Reconocerá nuestro P-hacker que está vez no puede hacer mucho?

De ninguna manera …!

Ofrece aplicar “técnicas especiales de mining” para analizar la asociación con cada una de las variables binarias.

En realidad cualquier test que use con un P_c = 0.05, y sin hacer corrección por comparaciones multiples, va a generar un 5% de falsos positivos

O sea, cinco variables binarias de las 100 van a estar asociadas a la variable de respuesta



No siempre las manipulaciones de datos son intencionales.

A veces ocurren por desconocimiento del dominio

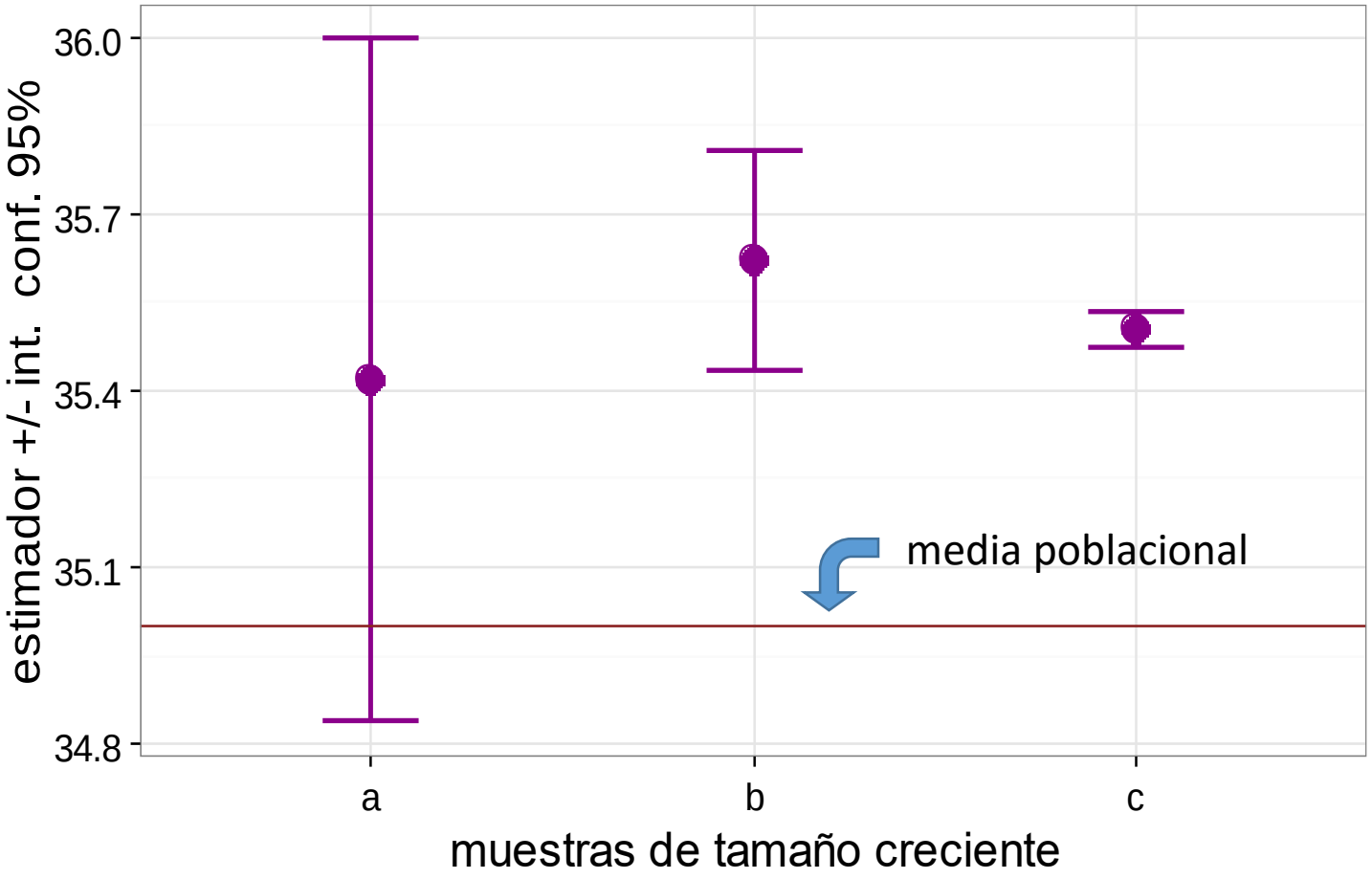
Y otras por falta de conocimiento técnico del científico de datos

Supongamos una variable que tiene una media de 35 en la población.

Para medirla se usa un equipo que para valores menores a 25 la mitad de las veces los registra como “por debajo del nivel de detección".

O sea, quedan registros con datos faltante con un patrón.

Si no tenemos esto en cuenta, al usar muestras cada vez mayores para estimar la media poblacional pasa algo curioso:


36.0
35.7
35.4
35.1
34.8
estimador +/- int. conf. 95%
media poblacional
a
b
c
muestras de tamaño creciente

Conclusiones y recomendaciones

Replicar

Registrar los proyectos

Fomentar la honestidad en los análisis

Documentar – Investigación reproducible

Ser competente en las técnicas

Muchas gracias !