VALIDACIÓN DE EXPERIMENTOS

LA ERA POST P<0.05

CiTIUS

Centro Singular de Investigación en Tecnoloxías da Información

Validación de experimentos

Métodos

Dataset	PDFC	NNEP	IS-CHC+INN
a	0.752	0.773	0.785
b	0.727	0.748	0.724
c	0.736	0.716	0.585
d	0.994	0.861	0.88
e	0.508	0.553	0.575
f	0.535	0.536	0.513
g	0.967	0.871	0.954
h	0.831	0.807	0.819
i	0.745	0.702	0.719
j	0.709	0.572	0.669
k	0.722	0.728	0.725
l	0.967	0.947	0.953

Casos de uso

¿De entre estos (columnas)

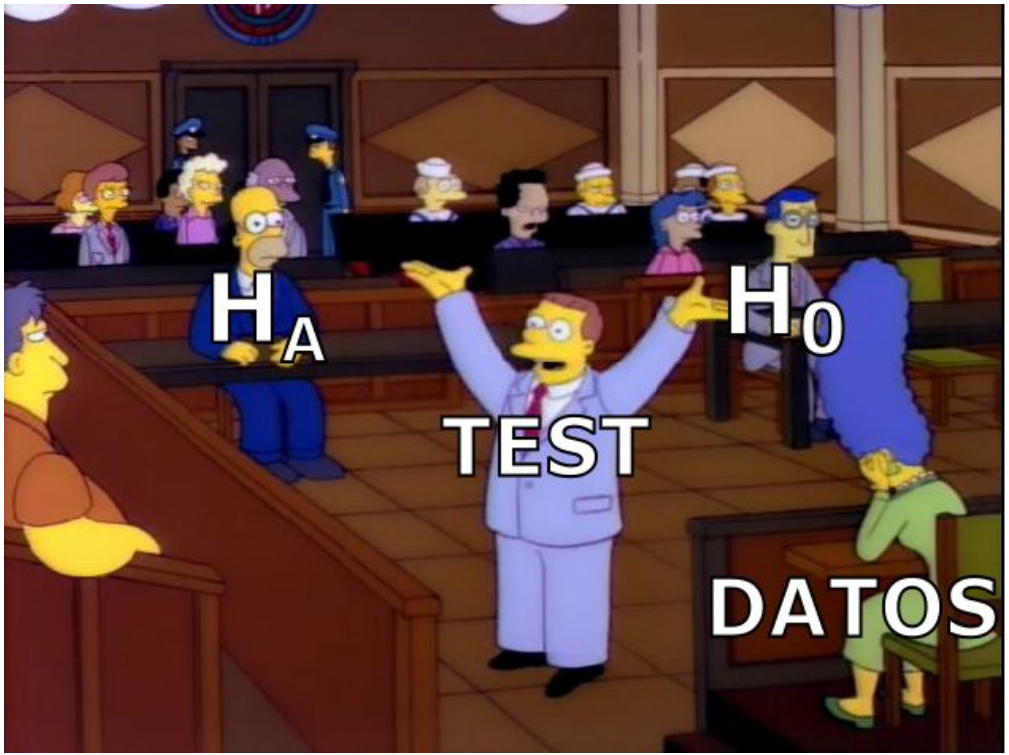
- Métodos
- Algoritmos
- Optimizaciones
- Configuraciones
- …

cuál es el mejor para estos casos de uso (filas)?

Como se hizo de toda la vida

- Transformar el objetivo de trabajo en una hipótesis
  - “Mi aproximación es mejor que la de mi vecino/enemigo/gurú de turno”
  - Difícil de contrastar
  - Hipótesis alternativa (H_A)
- Plantear el caso contrario
  - “Mi aproximación es igual que otra”
  - Fácil de caracterizar y contrastar
  - Hipótesis nula (H_0)
- Se mira la veracidad de H_0 a través de un test estadístico

Contraste de hipótesis



Pasos a seguir:

1. Obtener los datos
2. Calcular un estadístico a partir de los datos
   a. Depende del test
   b. Tiene una distribución asociada (paramétrico)
3. Se calcula el p-valor a partir del estadístico
4. ???
5. Profit!

Cosillas de nomenclatura

α: Probabilidad de decir H_A cuando H_0 es cierta

La has cagado (pero mucho) a posteriori

ß: Probabilidad de decir H_0 cuando H_A es cierta

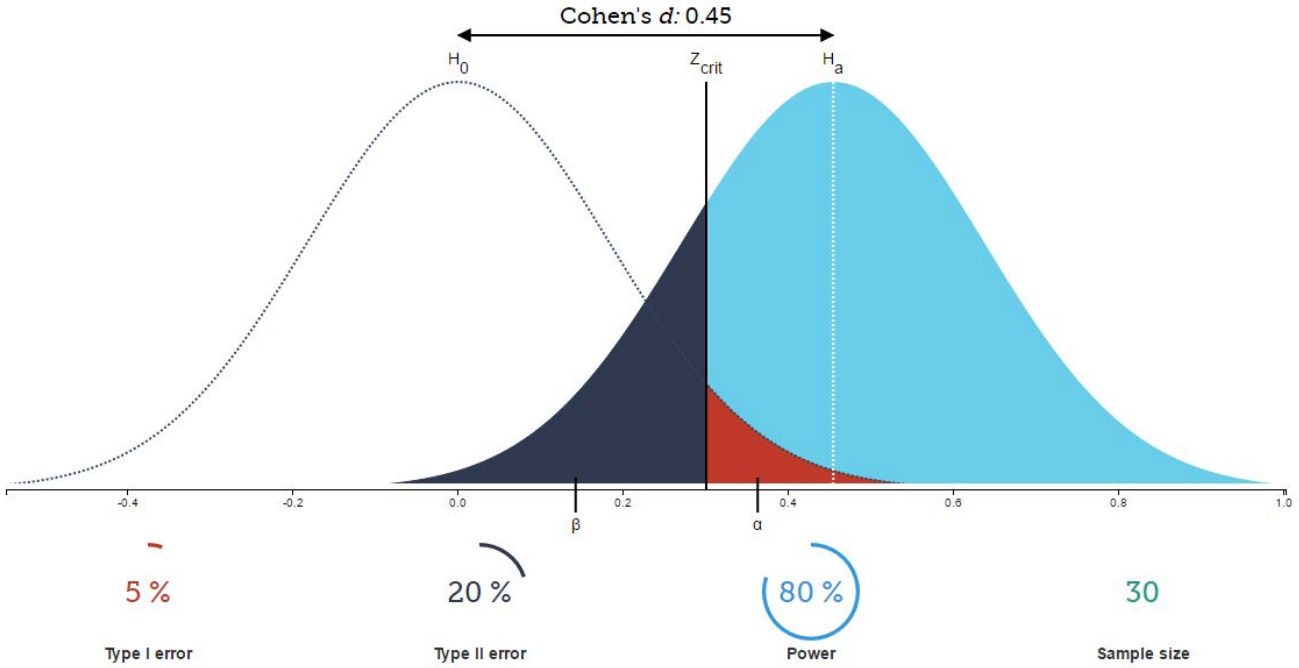
La has cagado a priori. Trabaja más, esclavo!

Potencia: Probabilidad de decir H_A cuando H_A es cierta

Enhorabuena, eres autor de un nuevo paper

	H_0 es cierta	H_A es cierta
Elegimos H_0	?	Error tipo 2 (ß)
Elegimos H_A	Error tipo 1 (α)	Potencia (1 - ß)

El famoso e inigualable p-valor



P-valor: Probabilidad de que, **siendo la H_0 cierta**, se haya obtenido un estadístico al menos tan extremo

“… an instance of a kind of essential mindlessness in the conduct of research"

– Bakan (1966)

Nickerson (2000) "Null hypothesis significance testing: a review of an old and continuing controversy."

Tipos de tests

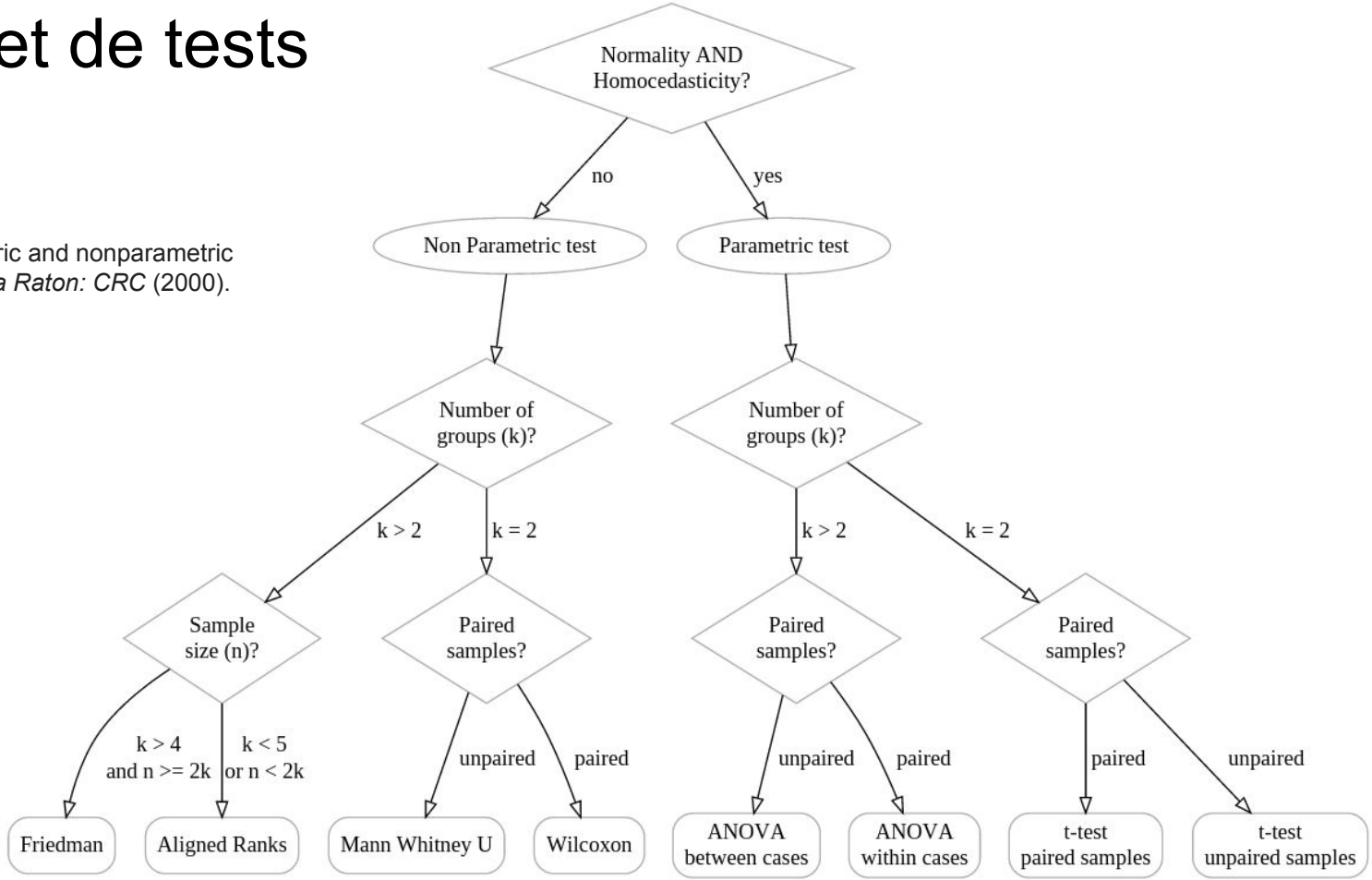
- Una sola muestra
  - Test sobre que distribución sigue, la media, varianza…
- Dos muestras
  - Tienen la misma media, siguen la misma distribución, tienen la misma tendencia…
  - Pueden ser apareados (se usan los mismos casos para ambas muestras) o no
- Multiples muestras
  - Las muestras provienen de la misma distribución
  - Es necesario realizar test post-hoc para comparar por pares a posteriori
    - Evita aumentar error tipo 1 al hacer comparaciones múltiples (FWER)

Paramétricos vs No paramétricos

- Independencia de los datos
  - Folds en cross-validation no son independientes
- Normalidad
  - Las muestras siguen una distribución normal
  - Típico en procesos biológicos, encuestas…
  - No tan común en ciencias de la computación
- Homocedasticidad
  - Las muestras tienen varianzas similares

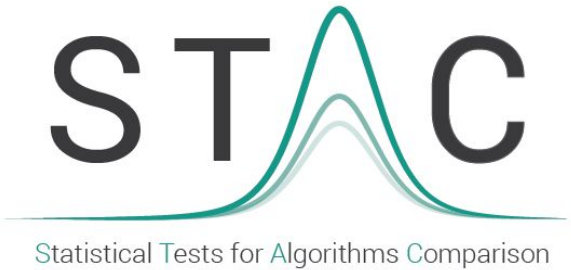
Cheatsheet de tests

Sheskin, David J. "Parametric and nonparametric statistical procedures." *Boca Raton: CRC* (2000).



Herramientas

- Python
  - scipy stats http://docs.scipy.org/doc/scipy-0.17.1/reference/stats.html
- R
  - Varias librerías desperdigadas (como suele pasar)
  - Mejor opción: scmamp https://cran.r-project.org/web/packages/scmamp/index.html



“Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution”

– Schmidt and Hunter (1997)

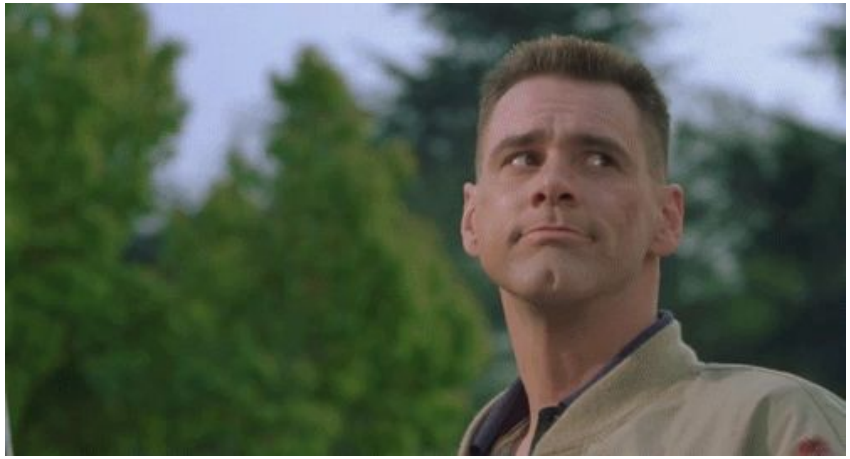
Nickerson (2000) "Null hypothesis significance testing: a review of an old and continuing controversy."

Cuentos de hadas

¿Cuántas muestras necesito para que el test sea significativo?

¿Cuál es el límite del p-valor para que el test sea significativo?

Obtengo un p-valor bajísimo, por lo tanto mi investigación es relevante.



Tamaño muestral

http://rpsychologist.com/d3/NHST/

# p < 0.05

DISTRIBUTIONS 45

only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which P = ·05, or 1 in 20, is 1·96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice *if the data are insufficiently numerous to bring them* out, but no lowering of the standard of significance would meet this difficulty.

R. A. Fisher, Statistical Methods for Research Workers, 1925

Que mide realmente el p-valor

Example 1: Sample size = 10

```
Two-sample T Results

     N   Mean  StDev  SE Mean
C4  10  5.011  0.748     0.24
C5  10  5.020  0.803     0.25

Difference = mu (C4) - mu (C5)
Estimate for difference:  -0.009

95% CI for difference:  (-0.741, 0.723)
T-Test of difference = 0 (vs not =): T-Value = -0.03  P-Value = 0.979  DF = 17
```

With 10 observations, the difference (-0.009) is not statistically significant

Example 2: Sample size = 1,000,000

```
Two-sample T Results

          N  Mean  StDev  SE Mean
C1  1000000  5.01   1.00   0.0010
C2  1000000  5.02   1.00   0.0010

Difference = mu (C1) - mu (C2)
Estimate for difference:  -0.00912

95% CI for difference:  (-0.01189, -0.00635)

T-Test of difference = 0 (vs not =): T-Value = -6.45  P-Value = 0.000  DF =1999994
```

With a million observations, the same difference (-0.009) is statistically significant!

P-hacking

“… despite the awesome pre-eminence this method has attained in our journals and textbooks of applied statistics, it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research”

– Rozeboom (1960)

Nickerson (2000) "Null hypothesis significance testing: a review of an old and continuing controversy."

Lo que dijo ASA

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**
4. Proper inference requires full reporting and transparency.
5. **A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.**
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA's statement on p-values: context, process, and purpose." *The American Statistician* (2016).

Lo que dijo ASA

“The widespread use of “statistical significance” (generally interpreted as “p < 0.05”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.”

“Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise”

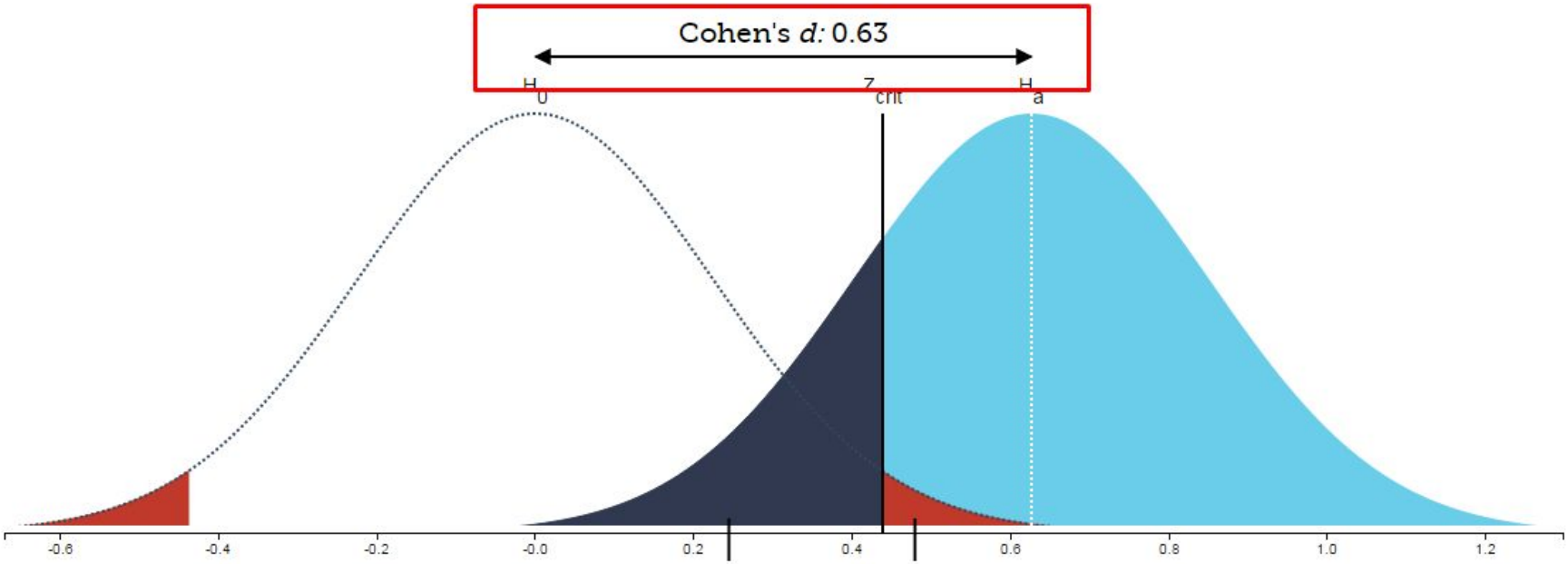
Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA's statement on p-values: context, process, and purpose." *The American Statistician* (2016).

"What's wrong with null hypothesis significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

– Cohen (1994)

Nickerson (2000) "Null hypothesis significance testing: a review of an old and continuing controversy."

Entonces nos comemos los mocos o como va esto



Cheatsheet size effect

- Paramétrico para dos muestras:

$$\text{Cohen's d} = (\mu_2 - \mu_1)/\sqrt{((n_2-1)\sigma_2^2 + (n_1-1)\sigma_1^2)/(n_2+n_1-2)}$$

- Paramétrico para multiples muestras:

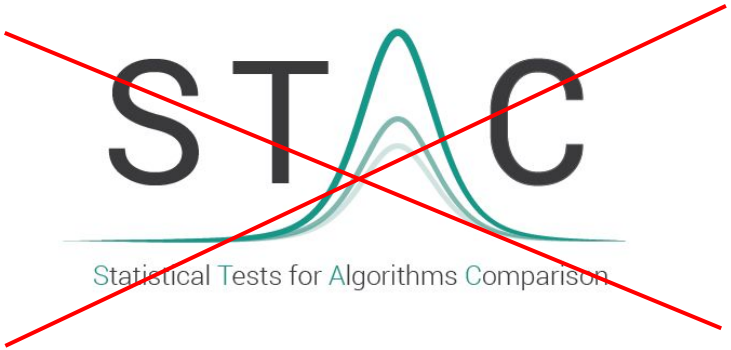
$$\eta^2 = SS_{between}/SS_{total}$$

- No paramétricos:

$$z/\sqrt{n}$$

Herramientas?

- Python
  - ~~scipy stats http://docs.scipy.org/doc/scipy-0.17.1/reference/stats.html~~
- R
  - varias librerías desperdigadas (como suele pasar)
  - ~~Mejor opción: scmamp https://cran.r-project.org/web/packages/scmamp/index.html~~

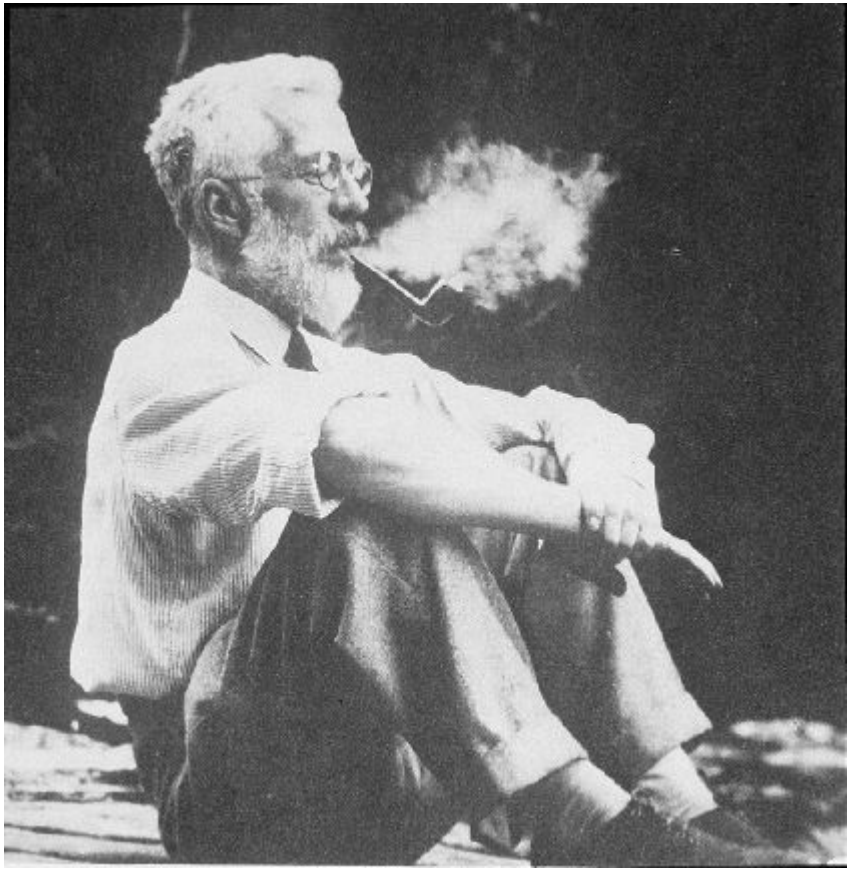


“The textbooks are wrong. The teaching is wrong. The seminar you just attended is wrong. The most prestigious journal in your scientific field is wrong."

– Ziliak and McCloskey (2008)

Nickerson (2000) "Null hypothesis significance testing: a review of an old and continuing controversy."

"The calculation [p-value] is **absurdly academic**, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; **he rather gives his mind to each particular case in the light of his evidence and his ideas**"



VALIDACIÓN DE EXPERIMENTOS

LA ERA POST P<0.05

Ismael Rodríguez

ismael.rodriguez@usc.es (a expirar)
ismael.rodriguez@hpe.com