Píndoles estadístiques UEB-VHIR

La crisi de la significació estadística: què diuen i què no diuen els p-valors

Santiago Pérez-Hoyos /Alex Sànchez
Unitat d'Estadística i Bioinformàtica

Divendres 28 de Juny de 12:30 a 13:30
Sala d'Actes de Traumatologia i Rehabilitació

Les píndoles estadístiques son sessions divulgatives, organitzades per la Unitat d'Estadística i Bioinformàtica (UEB) del VHIR, on es presenten problemes i solucions estadístiques dirigides als professionals interessats del Campus Vall d'Hebron

Vall d'Hebron
Institut de Recerca
VHIR

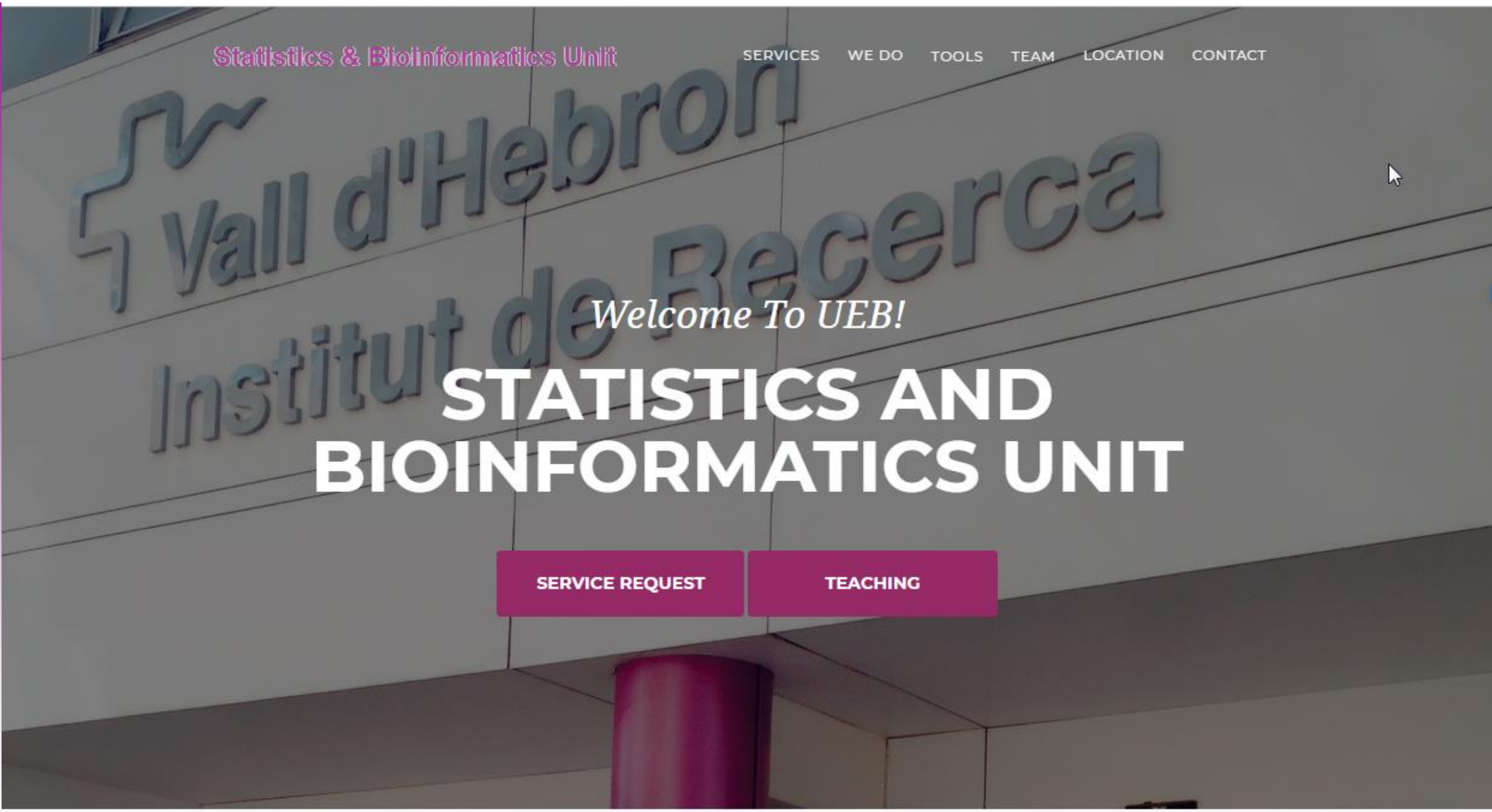
ueb.vhir.org

GRBIO

UNIVERSITAT DE BARCELONA

Outline

- Introduction and motivation
- A quick review:
  - Significance & Hypothesis tests, Confidence Intervals
- P-values drawbacks (1): The real ones
- P-values drawbacks (2): Misconceptions
- Alternatives and recommendations



Statistics & Bioinformatics Unit

SERVICES WE DO TOOLS TEAM LOCATION CONTACT

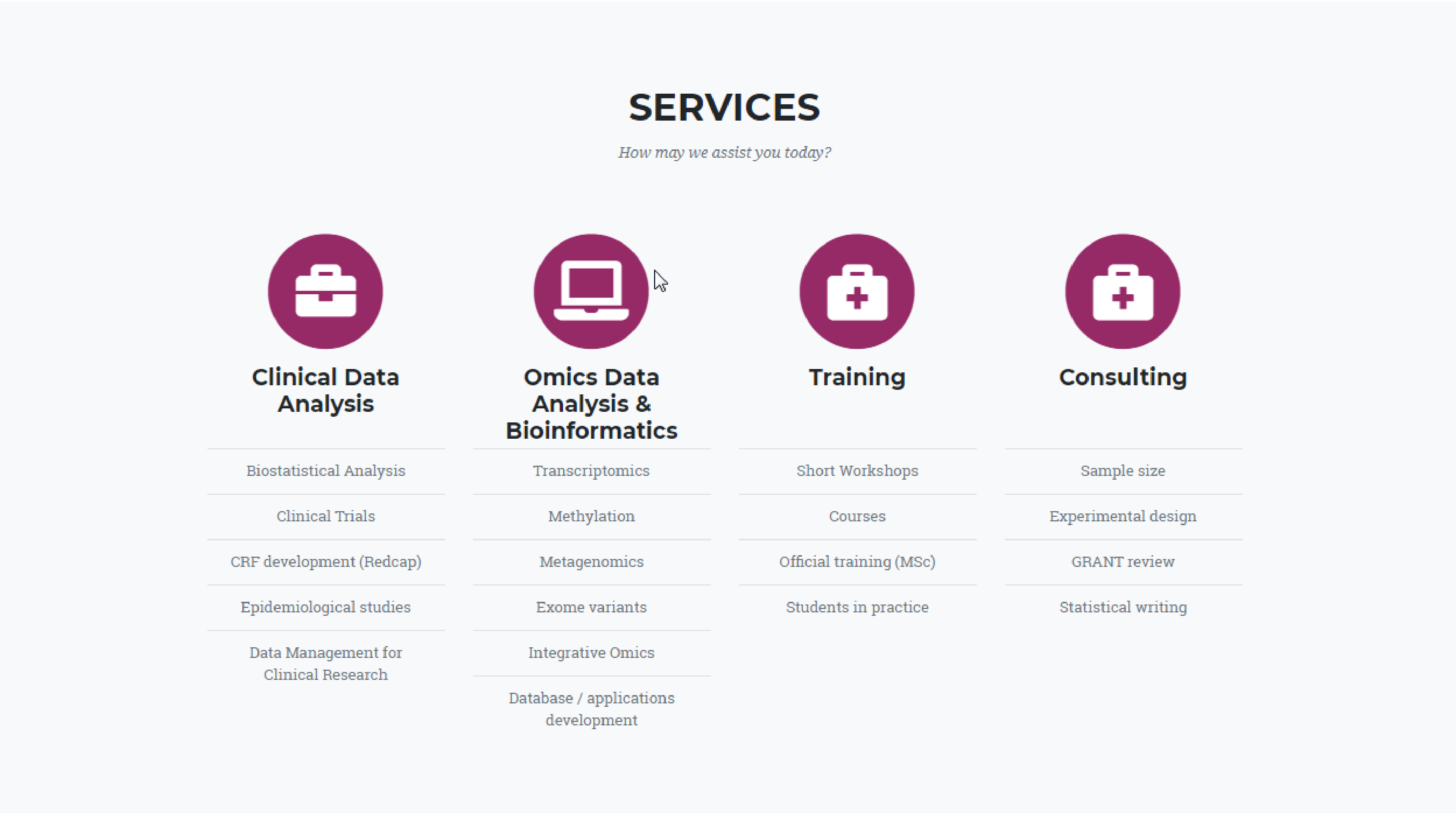
Welcome To UEB!

STATISTICS AND BIOINFORMATICS UNIT

SERVICE REQUEST

TEACHING

ueb.vhir.org



SERVICES

How may we assist you today?

Clinical Data Analysis

- Biostatistical Analysis
- Clinical Trials
- CRF development (Redcap)
- Epidemiological studies
- Data Management for Clinical Research

Omics Data Analysis & Bioinformatics

- Transcriptomics
- Methylation
- Metagenomics
- Exome variants
- Integrative Omics
- Database / applications development

Training

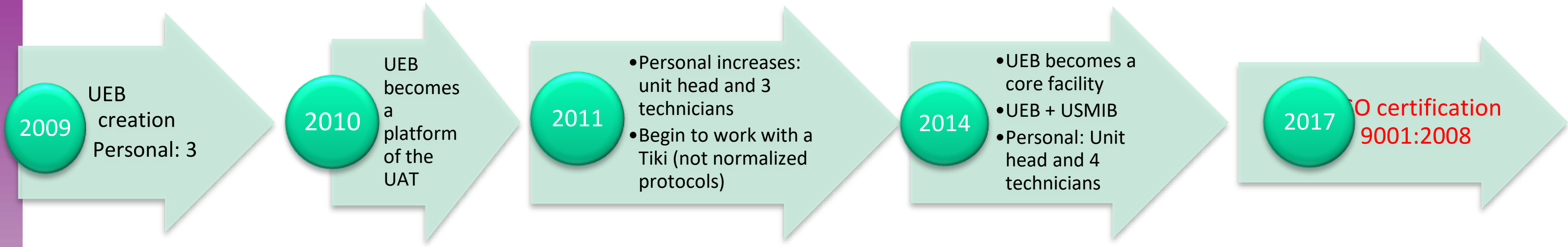
- Short Workshops
- Courses
- Official training (MSc)
- Students in practice

Consulting

- Sample size
- Experimental design
- GRANT review
- Statistical writing

ueb.vhir.org

UEB evolution



2009

UEB creation

Personal: 3

2010

UEB becomes a platform of the UAT

2011

- Personal increases: unit head and 3 technicians
- Begin to work with a Tiki (not normalized protocols)

2014

- UEB becomes a core facility
- UEB + USMIB
- Personal: Unit head and 4 technicians

2017

O certification 9001:2008

Statistical pills

Suma de talent al servei del pacient d'avui i del demà

#VHambTu

www.vallhebron.com

PINDOLES ESTADÍSTIQUES UEB -VHIR

p-valors
Estratègies per a la construcció de models i biomarcadors
Visualització moderna de dades
Endreçant les idees
Mesurant l'associació entre dues variables
mutiple testing

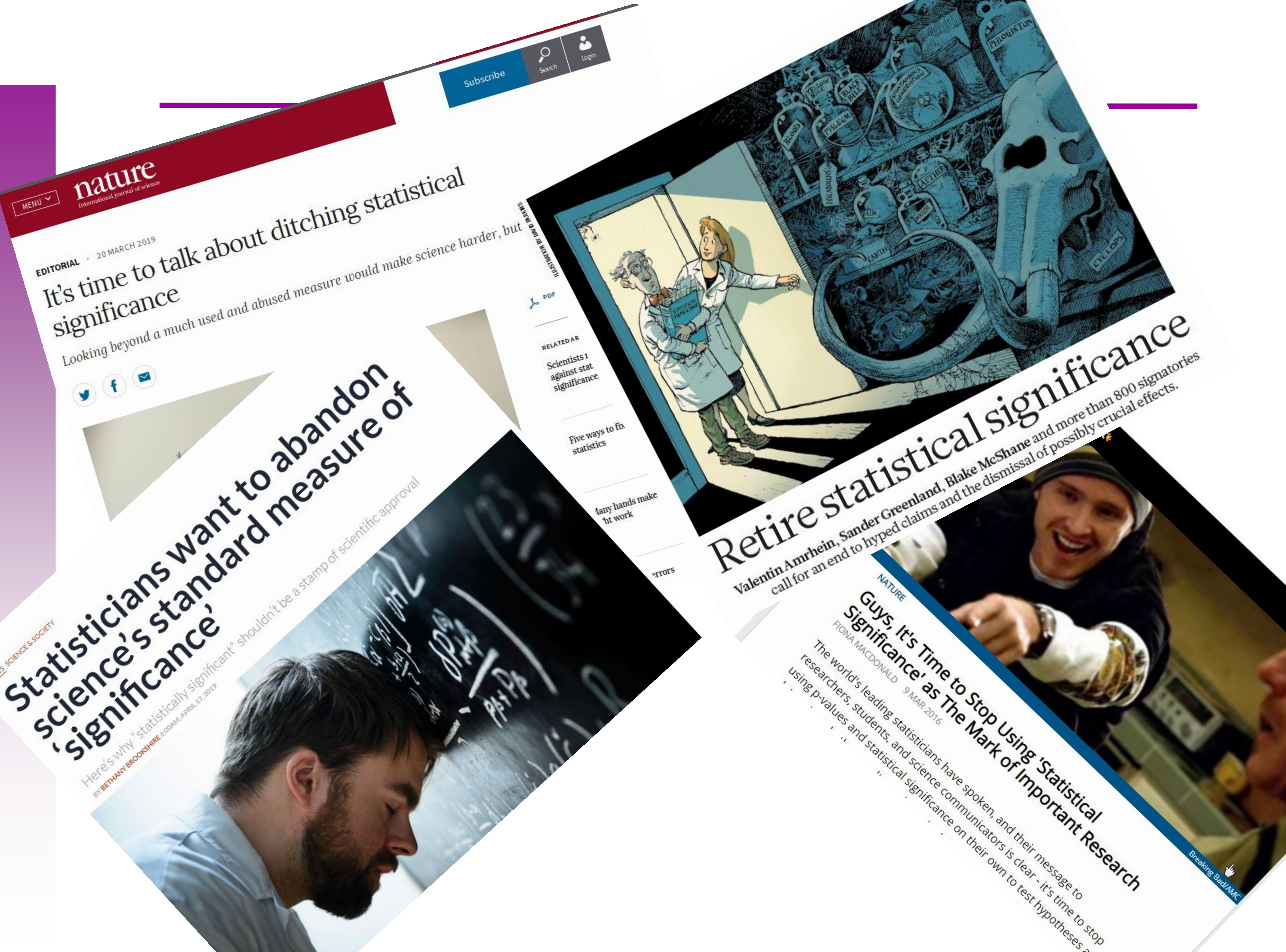
Vall d'Hebron Barcelona Hospital Campus

Generalitat de Catalunya Departament de Salut

Institut Català de la Salut

Vall d'Hebron Institut de Recerca

A significance crisis?



EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but

RELATED AR

Scientists r against stat significance

Five ways to fix statistics

Many hands make ht work

errors

SCIENCE & SOCIETY

Statisticians want to abandon science's standard measure of 'significance'

Here's why "statistically significant" shouldn't be a stamp of scientific approval

BY BETHANY BROOKSHIRE 6:00AM, APRIL 17, 2019

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

NATURE

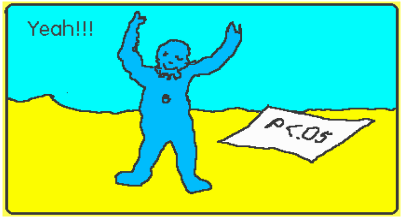
Guys, It's Time to Stop Using 'Statistical Significance' as The Mark of Important Research

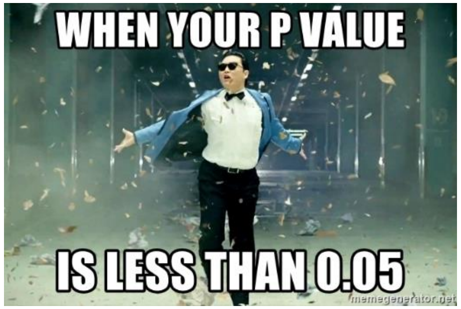
FIONA MACDONALD 9 MAR 2016

The world's leading statisticians have spoken, and their message to researchers, students, and science communicators is clear - it's time to stop using p-values and statistical significance on their own to test hypotheses a

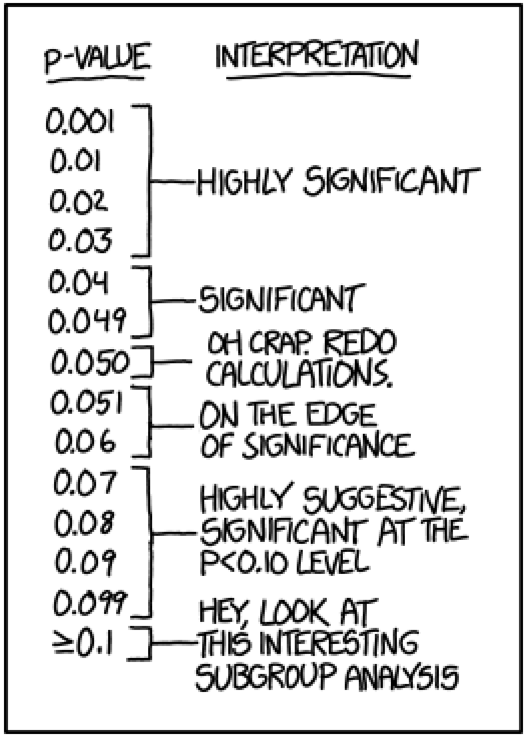
The Question

- Research is dominated by the concept of statistical significance which, at its side ...
- is determined by p-values thresholds.
- In short:

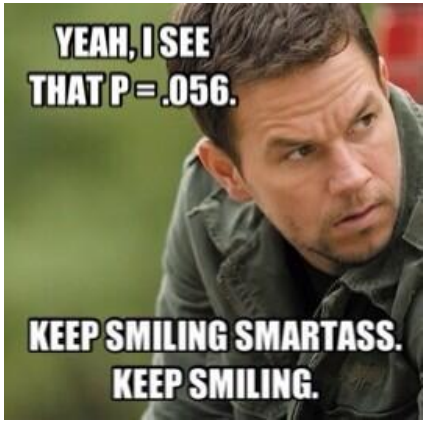




😃

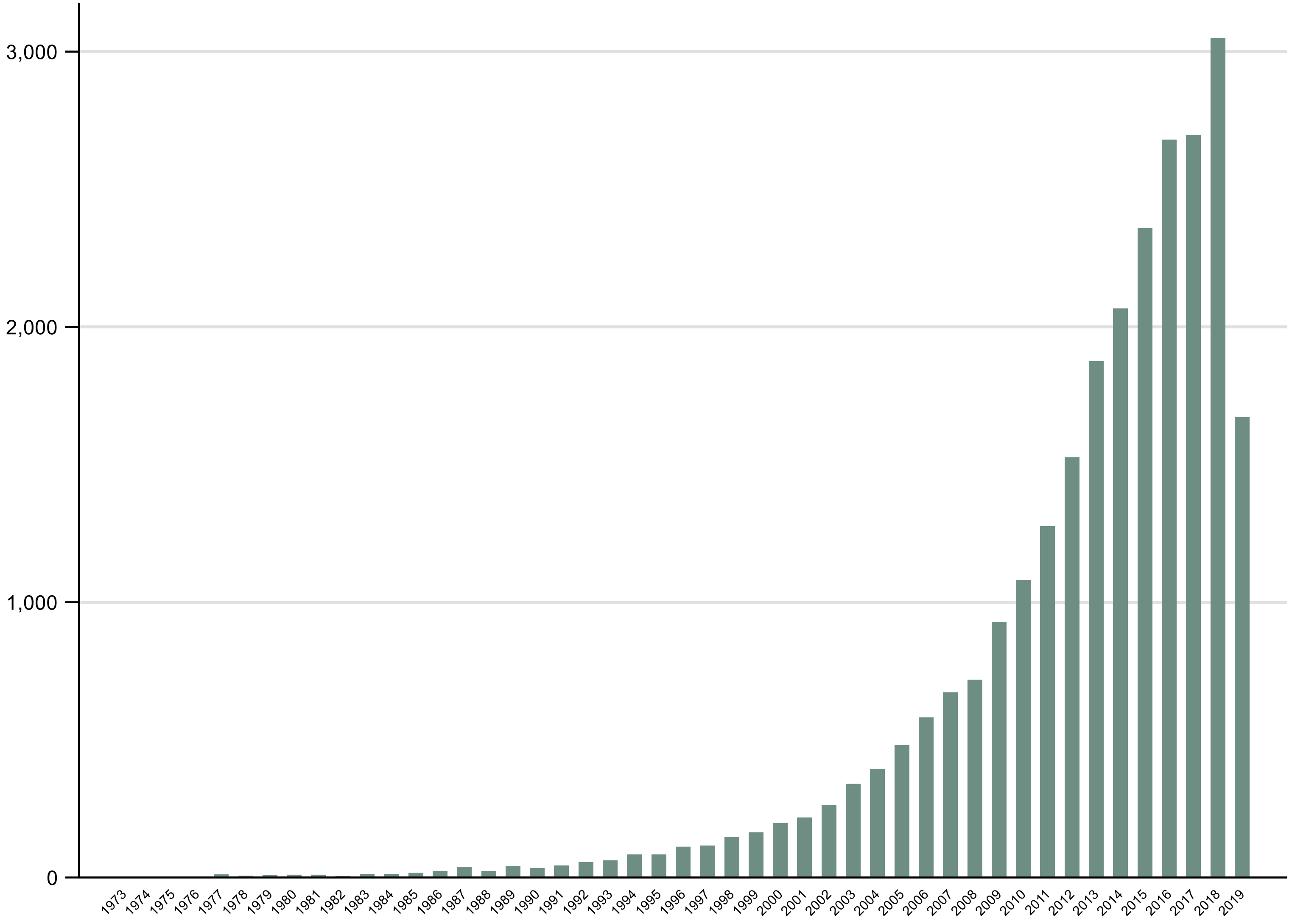


P-VALUE	INTERPRETATION
0.001 0.01 0.02 0.03	HIGHLY SIGNIFICANT
0.04 0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051 0.06	ON THE EDGE OF SIGNIFICANCE
0.07 0.08 0.09 0.099	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS



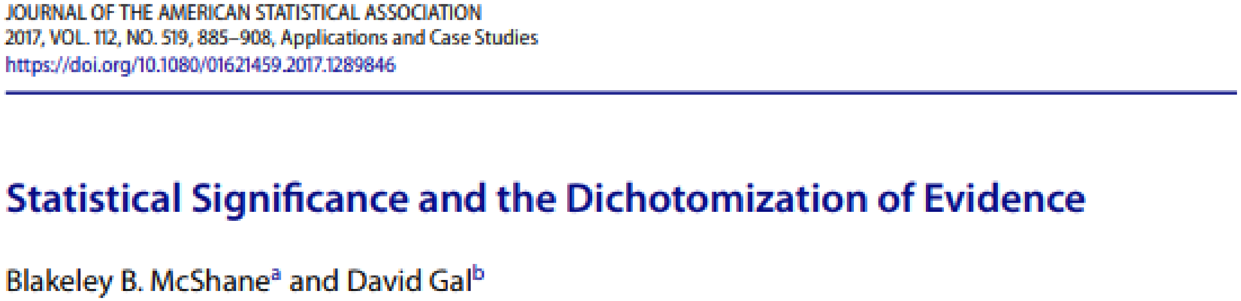
☹

Frequency of abstracts mentioning signific* p value in pubmed



A false dichotomization

- Statisticians (and a few scientists from other fields) have been long claiming against this



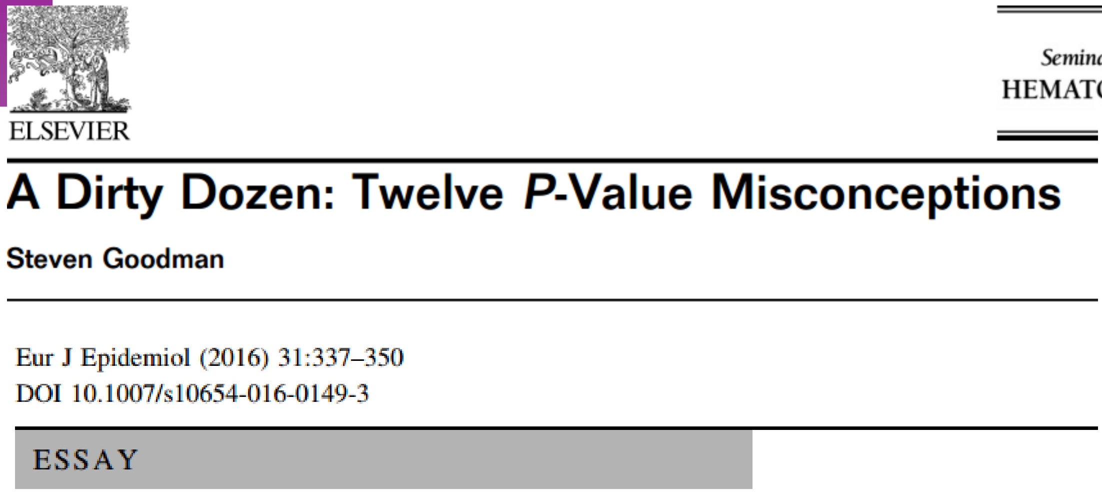
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, 885–908, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1289846

Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane[a] and David Gal[b]

  - “ ...The 0.05 (or any other) threshold used to dichotomize results into statistically significant and not statistically significant is **arbitrary**
  - One consequence of this dichotomization is that it facilitates confounding **statistical significance** with **practical importance**

And this is not the only problem!



ELSEVIER

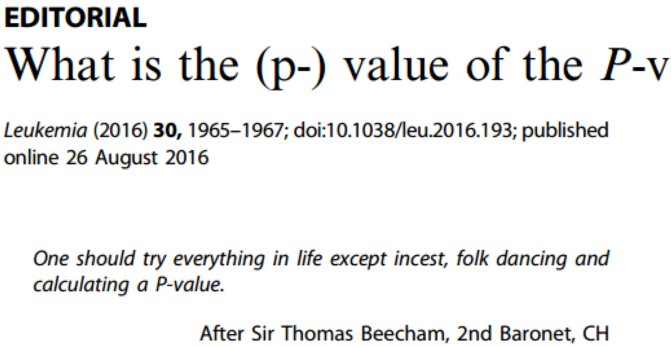
A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

ESSAY

Seminars in
HEMATOLOGY



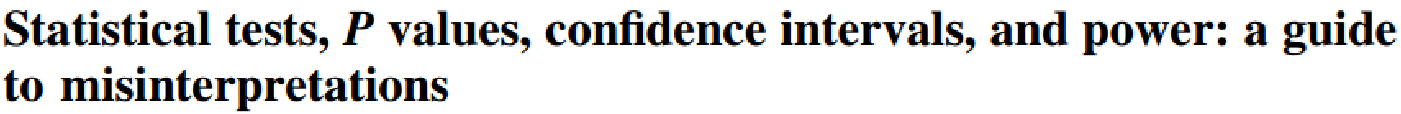
EDITORIAL

What is the (p-) value of the *P*-value?

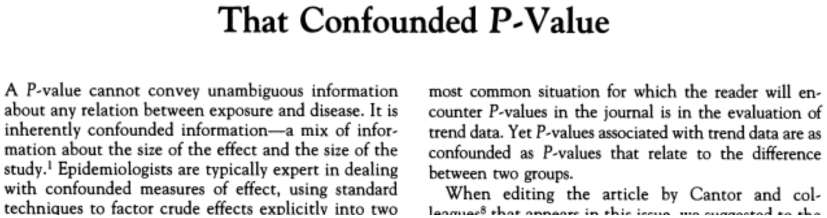
Leukemia (2016) **30,** 1965–1967; doi:10.1038/leu.2016.193; published online 26 August 2016

One should try everything in life except incest, folk dancing and calculating a P-value.

After Sir Thomas Beecham, 2nd Baronet, CH



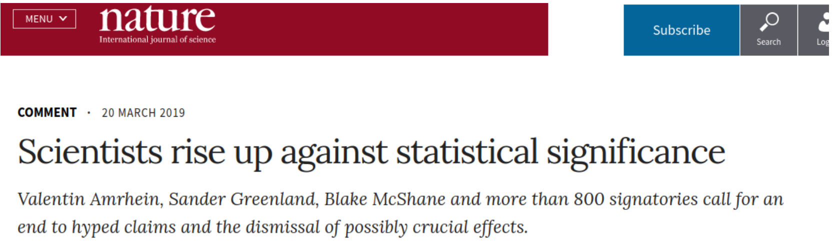
Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations



That Confounded *P*-Value

A P-value cannot convey unambiguous information about any relation between exposure and disease. It is inherently confounded information—a mix of information about the size of the effect and the size of the study.[1] Epidemiologists are typically expert in dealing with confounded measures of effect, using standard techniques to factor crude effects explicitly into two most common situation for which the reader will encounter *P*-values in the journal is in the evaluation of trend data. Yet P-values associated with trend data are as confounded as *P*-values that relate to the difference between two groups.

When editing the article by Cantor and colleagues[8] that appears in this issue, we suggested to the



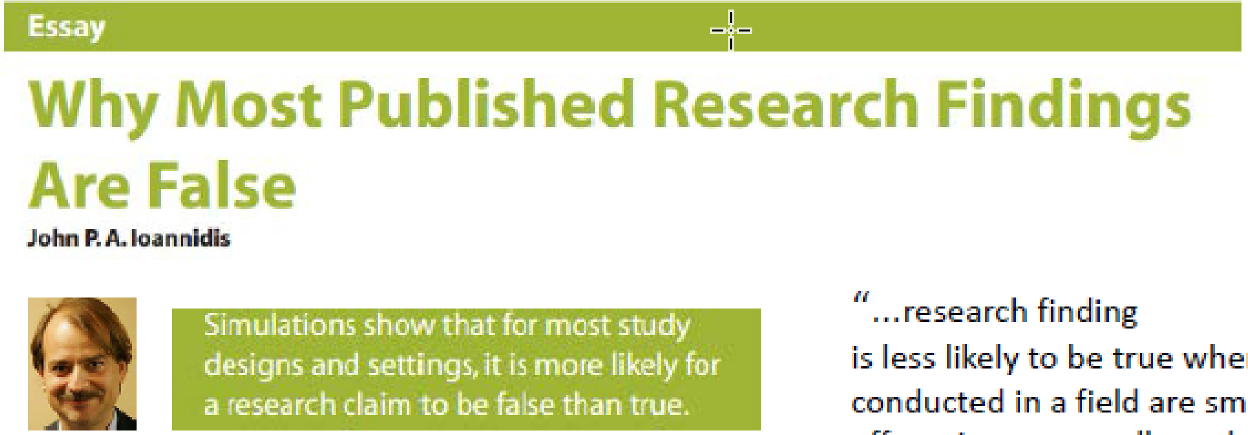
COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Scientific Reproducibility

Ioannidis, PloSmedicine, 2005 *Why most published research findings are false*



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

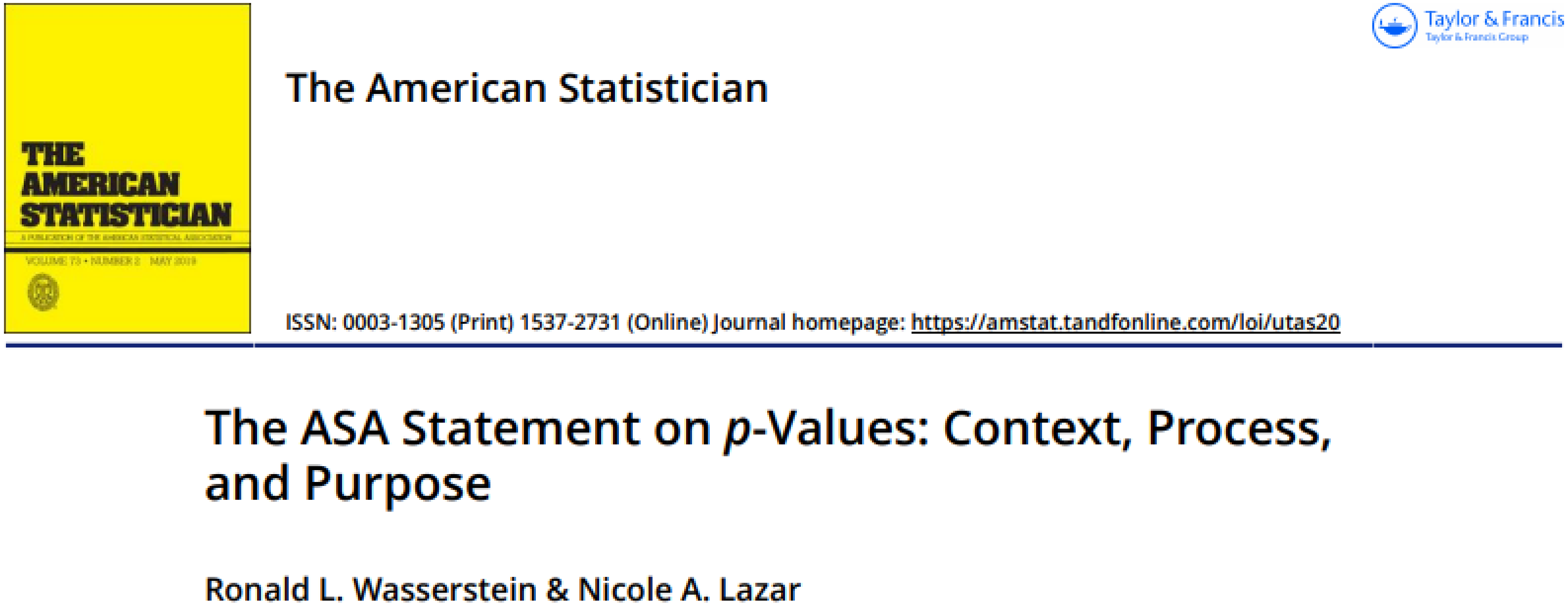
Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true.

+3400 citations

"...research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field *in chase of statistical significance."*

Scientists and Statisticians adopt a position

- The American Statistical Association decided to make a step forward, get involved and make a claim to make it better or don’t do it at all!


THE
AMERICAN
STATISTICIAN
A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION
VOLUME 73 • NUMBER 2 • MAY 2019
Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

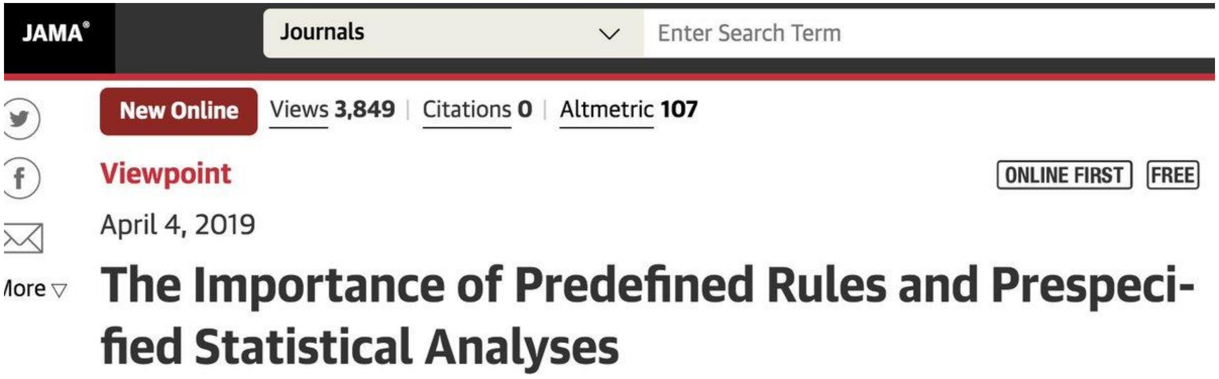
The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

The ASA six principles

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.

Although not everyone agreeded ...



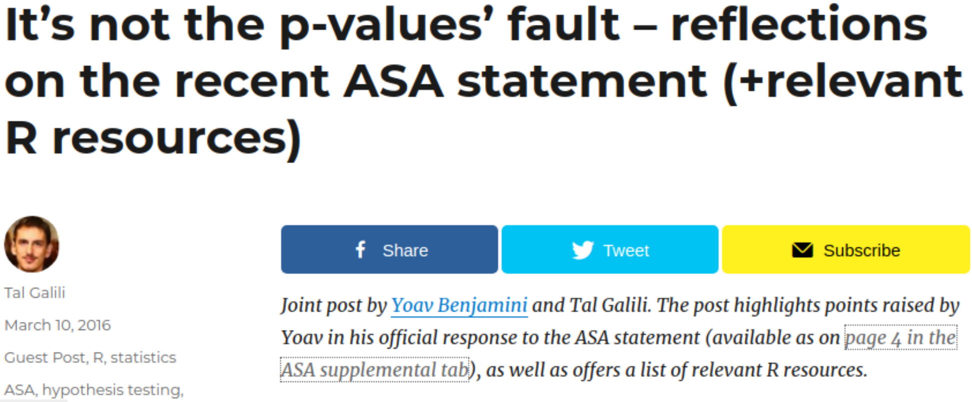
Viewpoint

April 4, 2019

The Importance of Predefined Rules and Prespecified Statistical Analyses

Do Not Abandon Significance

John P. A. Ioannidis, MD, DSc[1,2]



It’s not the p-values’ fault – reflections on the recent ASA statement (+relevant R resources)

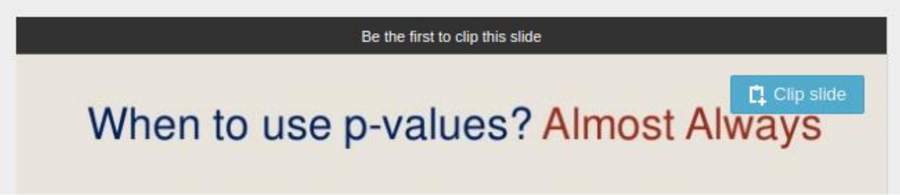
Tal Galili

March 10, 2016

Guest Post, R, statistics

ASA, hypothesis testing,

Joint post by Yoav Benjamini and Tal Galili. The post highlights points raised by Yoav in his official response to the ASA statement (available as on page 4 in the ASA supplemental tab), as well as offers a list of relevant R resources.



When to use p-values? Almost Always

Our roadmap from here

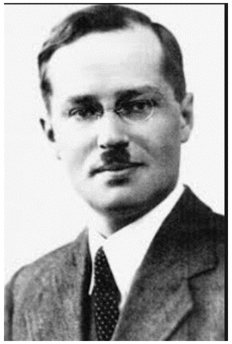
- Review basic concepts
- Review some of the things that p-values and significance are blamed of
- Try to answer the question:
  - ***What could we do if we decided not to use p-values***

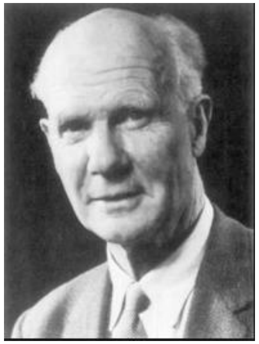
What is a p –value?

R.A. Fisher 1925

Introduced Significance Testing

"…we can calculate the standard deviation of the mean of a random sample of any size, and so test whether or not it differs significantly from any fixed value. If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance ; this is roughly equivalent to the corresponding limit P=.05"





J. Neyman E Pearson
1928-34

Established Hypothesis Testing Theory

- *H_0: Null Hypothesis*
- *H_1: Alternative Hypothesis*
- *Statistical test for Decision Criteria*
- *Type I error (α) and type II error (β)*
- *Power of a test. ($1-\beta$)*

Formal definition of p-value

“**p-value**. …to test the conformity of the particular data under analysis with H_0 in some respect:

…we find a function $T = t(\mathbf{y})$ of the data, to be called the **test statistic**, such that

- the larger the value of T the more inconsistent are the data with H_0;
- The random variable $T = t(\mathbf{Y})$ has a (numerically) known probability distribution when H_0 is true.

…the p-value corresponding to any t_{0bs} as

$$p = Pr(t) = Pr(T \geq t_{0bs};\ H_0)$$”

(Mayo and Cox 2006, p. 81)

Formal definition of p-value

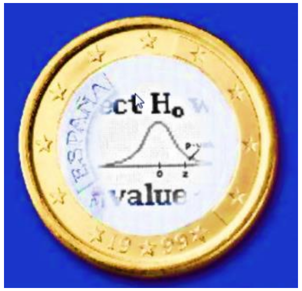
Probability under a certain model that you set up (null-hypothesis) that a certain data summary (e.g. a mean/difference of means) would be equal to or more extreme than what we get.

The P value is then the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis.

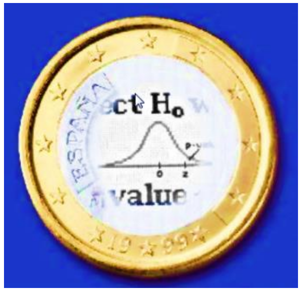
Supose we toss a coin 10 times

H_0:P(

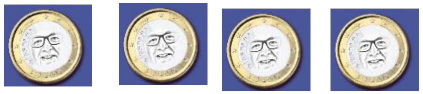
ESPAÑA
1999

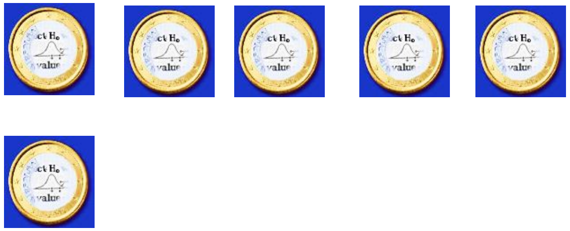
) = P()=0.5

Supose we toss a coin 10 times

H_0:P( ) = P()=0.5

Experiment 1

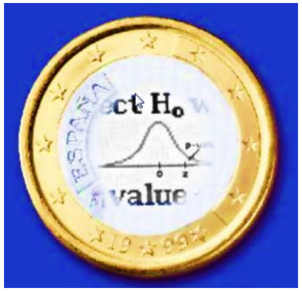




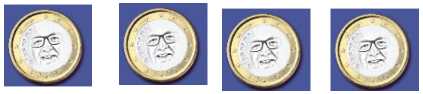
Supose we toss a coin 10 times

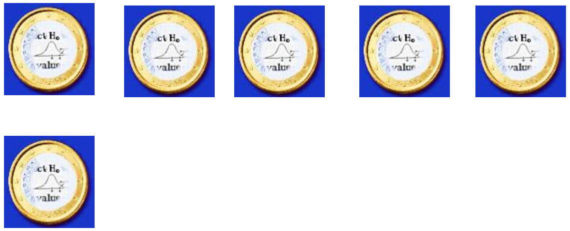
H_0:P(

ESPAÑA
1999

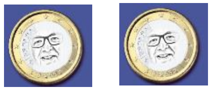
) = P()=0.5

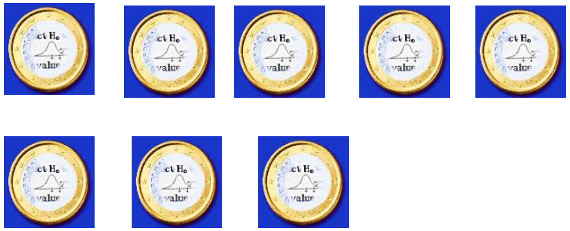
Experiment 1



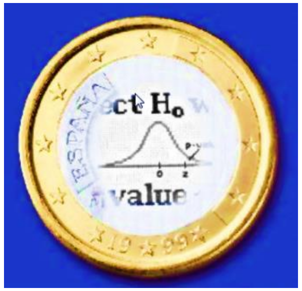


Experiment 2

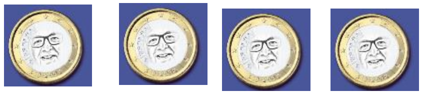
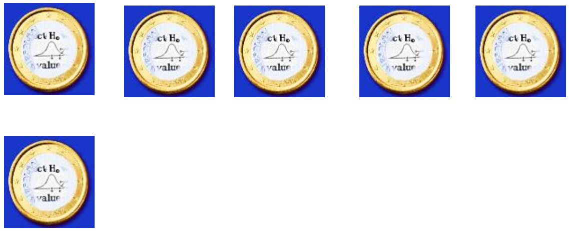


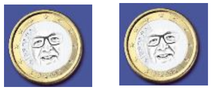
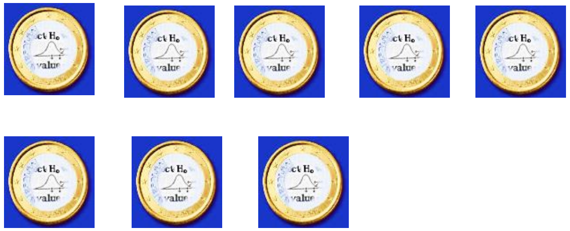


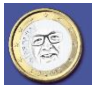
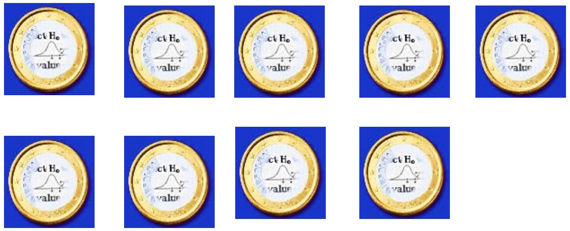
Supose we toss a coin 10 times

H_0:P(
ESPAÑA
1999
) = P(

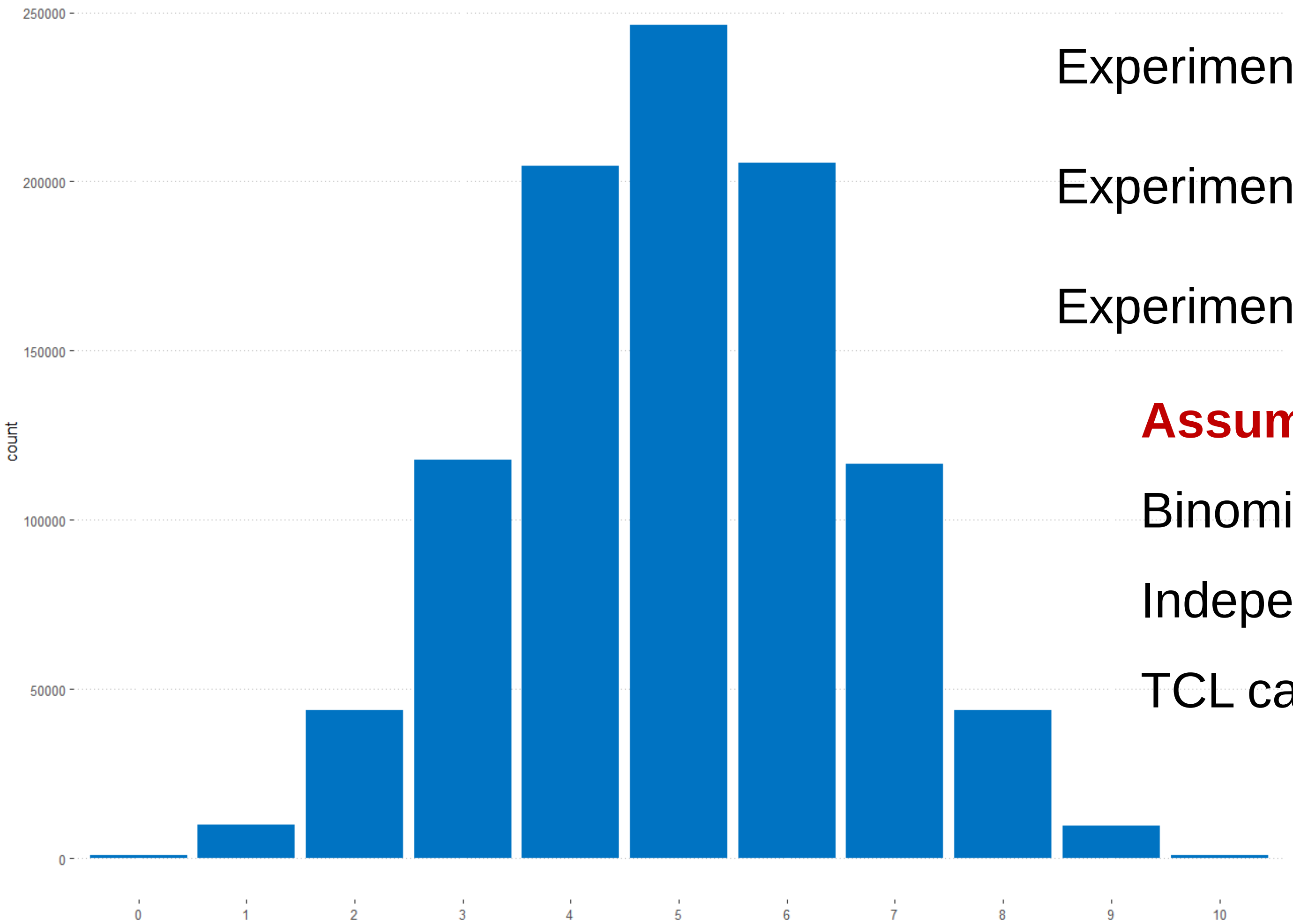
)=0.5

Experiment 1  

Experiment 2  

Experiment 3  

If we toss many many times…..



Probability of result or worst

Experiment 1: 0.5271

Experiment 2: 0.0578

Experiment 3: 0.0114

Assumptions

Binomial data

Independent observations

TCL can be applied

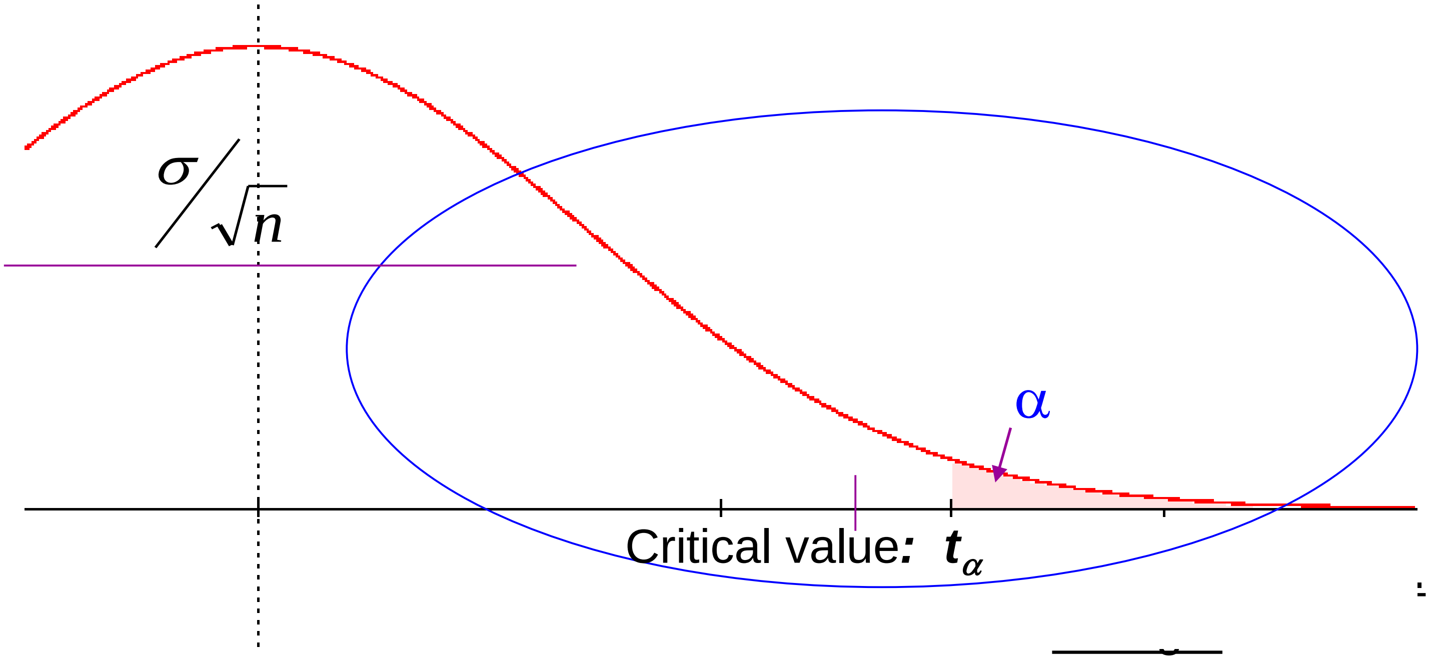
Hypothesis Testing

Test population hypothesis from samples

1. Establish Null Hypothesis(H_0)
2. Establish Alternative Hypothesis (H_α)
3. Select statistical test $T=t(\mathbf{y})$, to calculate probability ***under Null Hypothesis***
4. Obtain a sample $\boldsymbol{x_{ob}}$ and calculate test value $\boldsymbol{t_{ob}}$
5. Decide after comparing test value with **a critical value** $\boldsymbol{t_\alpha}$ or probability under null hypothesis.

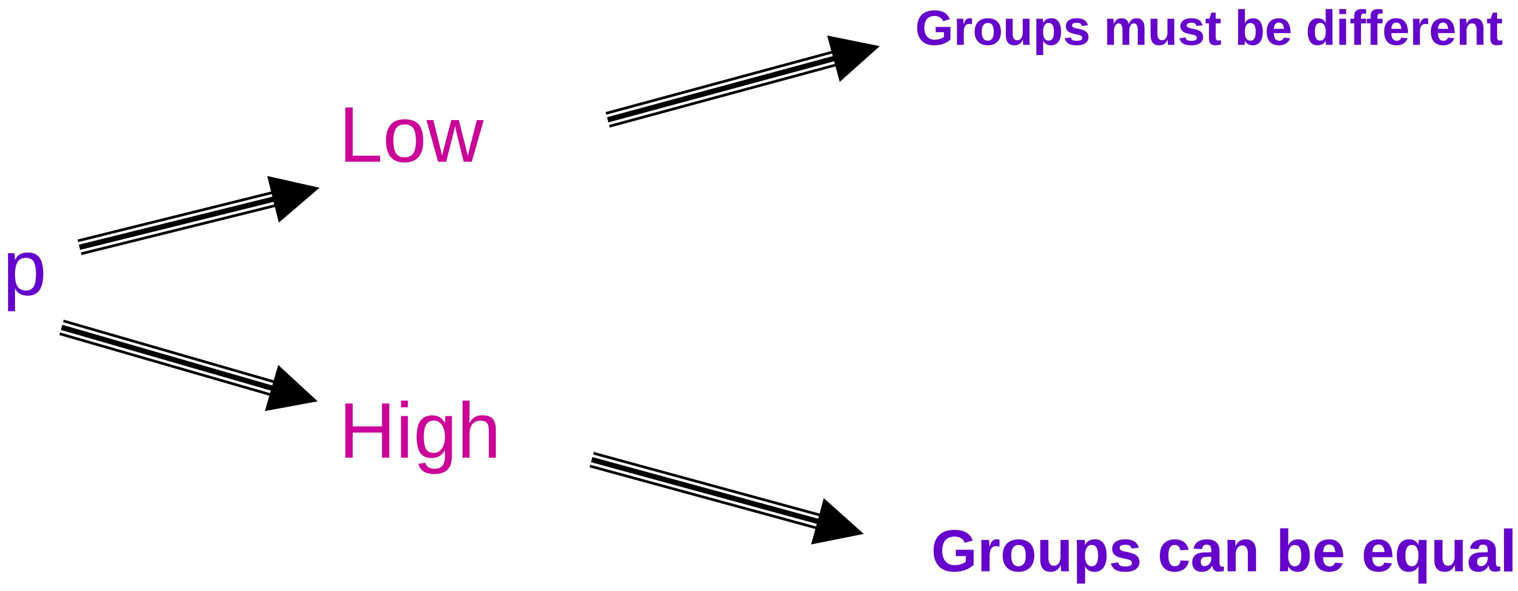
Critical Value

- At which value of the sample mean does one change from non-rejecting to rejecting the null hypothesis?
  - A value is selected such that the probability that the sample mean exceeds it, if the null hypothesis is true, is “small”, (for example 5%).
  - This value is called “Critical Value” $\boldsymbol{t}_\alpha$ and
  - the probability is called “significance level (α)”



¿How to decide which hypothesis is more likely?

Calculate probability (p) to observe differences between both groups under the hypothesis of no differences



p

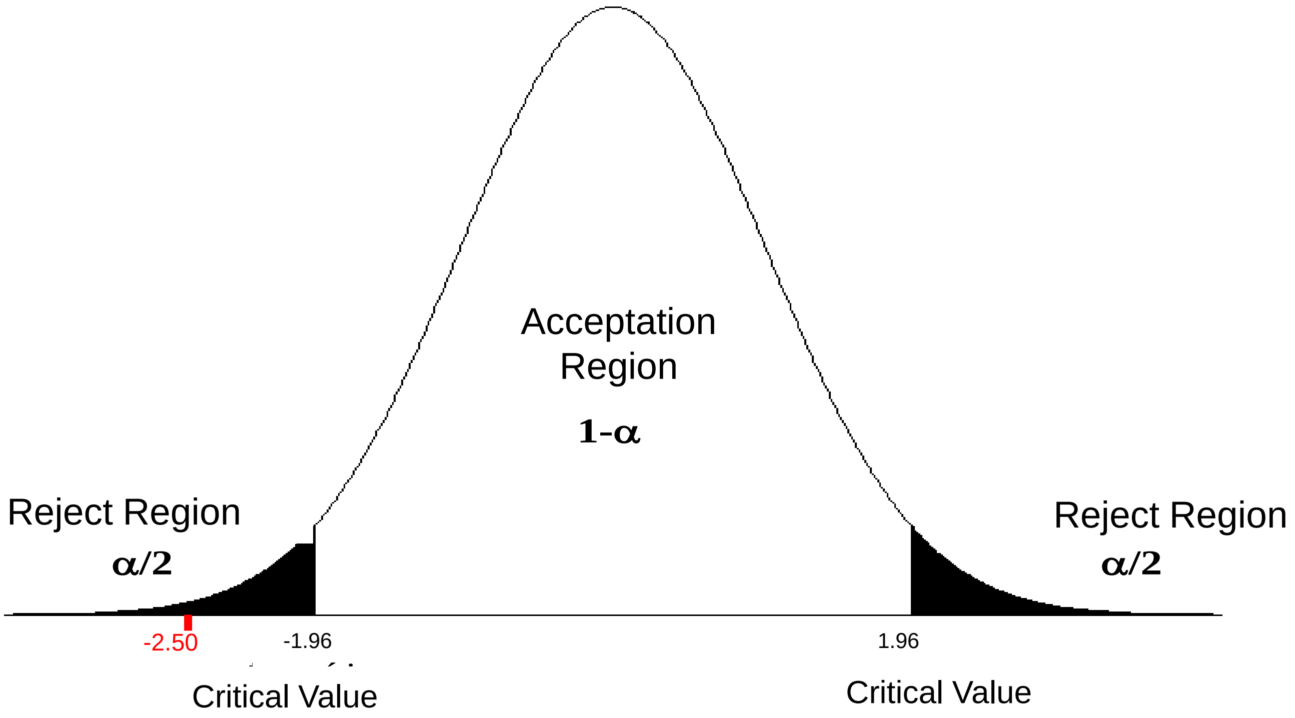
Low

Groups must be different

High

Groups can be equal

Decision Rule



Acceptation Region

$1\text{-}\alpha$

Reject Region

$\alpha/2$

Reject Region

$\alpha/2$

-2.50

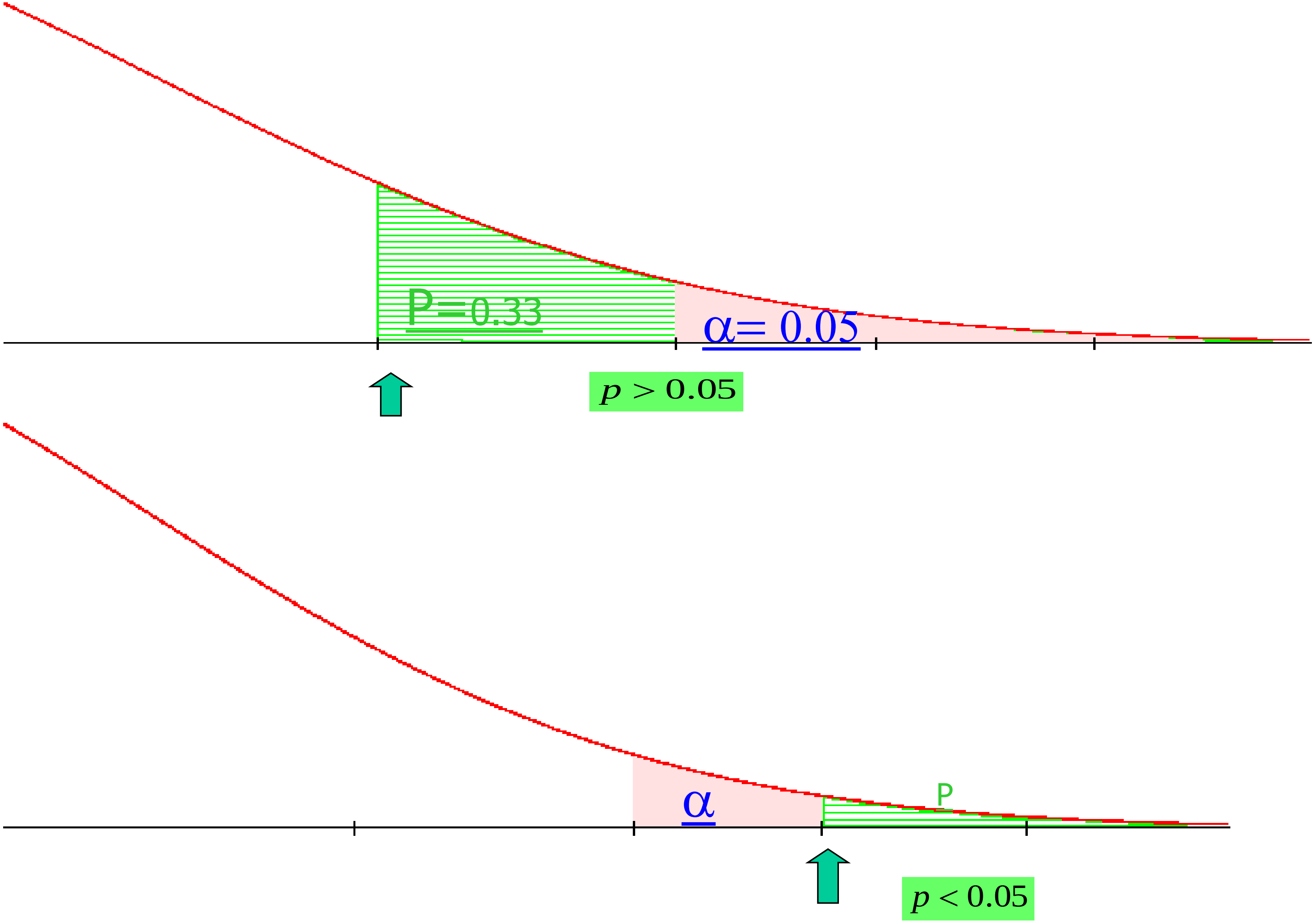
-1.96

1.96

Critical Value

Critical Value

P-value vs critical value



What is left

- Review some of the things that p-values and significance are blamed of
- Try to answer the question:
  - *What could we do if we decided not to use p-values*

P values depend on sample size

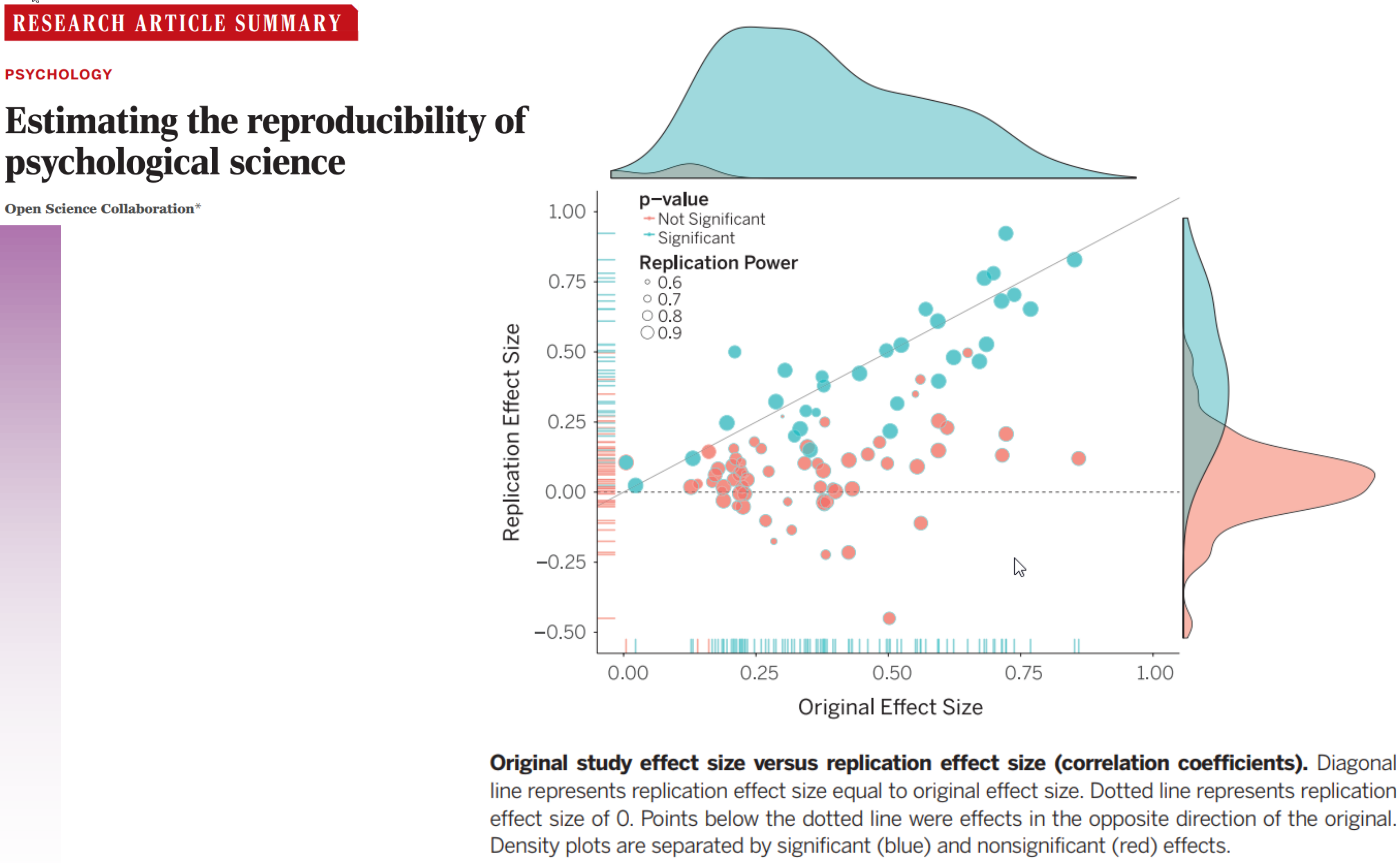
```
|
           |          col
       row |          1           2 |     Total
-----------+----------------------+----------
         1 |          2          98 |       100
           |       2.00       98.00 |    100.00
-----------+----------------------+----------
         2 |          4          96 |       100
           |       4.00       96.00 |    100.00
-----------+----------------------+----------
     Total |          6         194 |       200
           |       3.00       97.00 |    100.00

          Valor p Pearson  0.407
```

```
           |          col
       row |          1           2 |     Total
-----------+----------------------+----------
         1 |          4         196 |       200
           |       2.00       98.00 |    100.00
-----------+----------------------+----------
         2 |          8         192 |       200
           |       4.00       96.00 |    100.00
-----------+----------------------+----------
     Total |         12         388 |       400
           |       3.00       97.00 |    100.00

          Valor p Pearson 0.241
```

```
           |          col
       row |          1           2 |     Total
-----------+----------------------+----------
         1 |         20         980 |     1,000
           |       2.00       98.00 |    100.00
-----------+----------------------+----------
         2 |         40         960 |     1,000
           |       4.00       96.00 |    100.00
-----------+----------------------+----------
     Total |         60       1,940 |     2,000
           |       3.00       97.00 |    100.00

             Valor p Pearson 0.009
```

P-values are unstable

Reproducibility: P values depend on data samples



RESEARCH ARTICLE SUMMARY

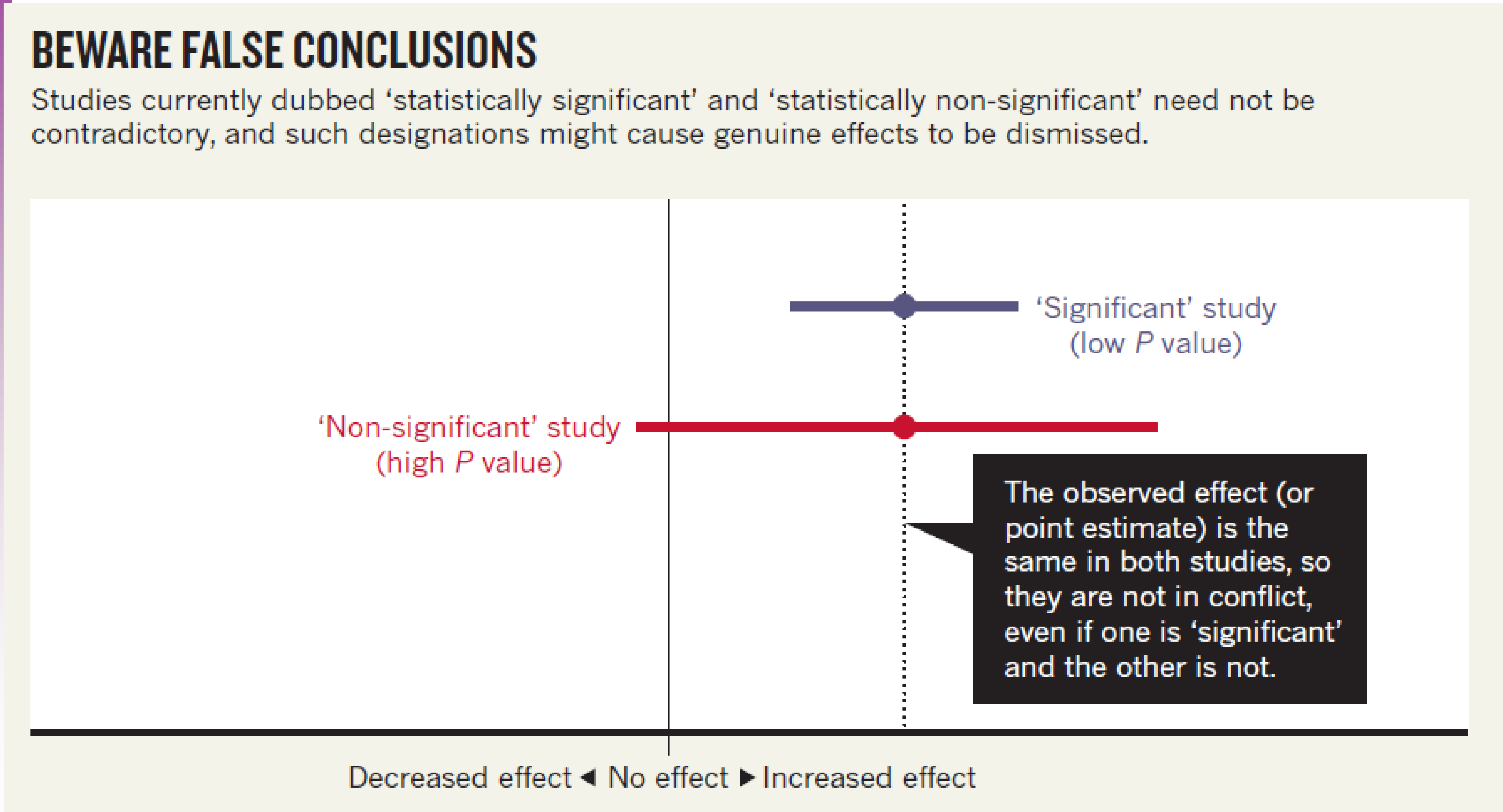
PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

False Conclusions



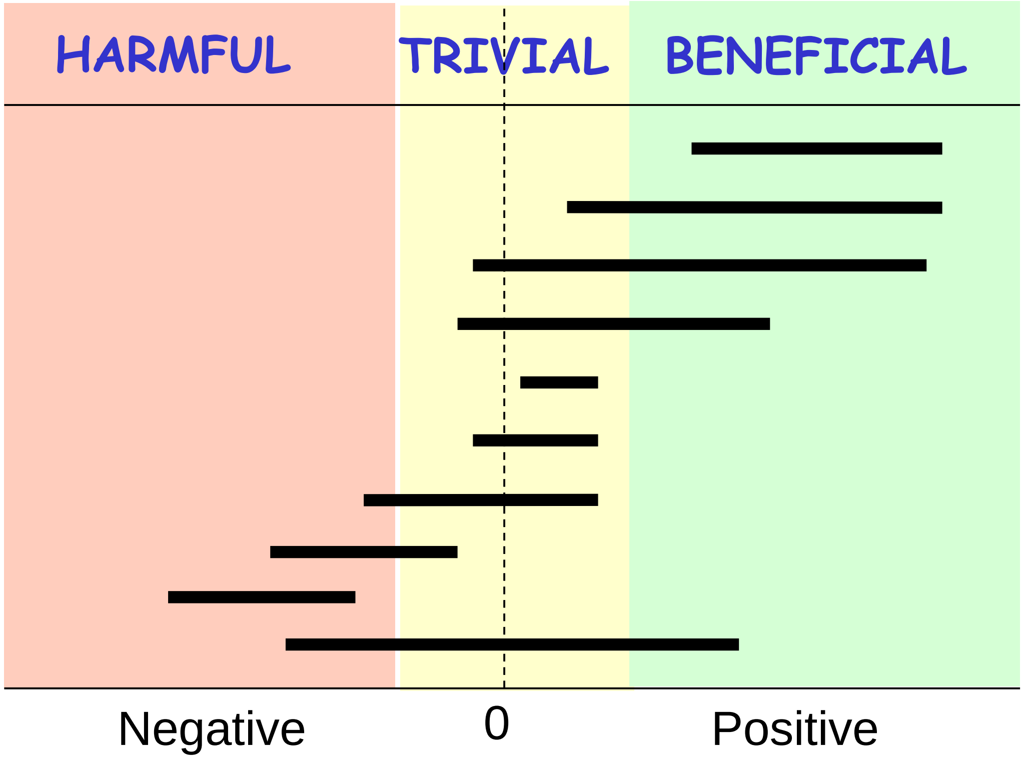
BEWARE FALSE CONCLUSIONS

Studies currently dubbed ‘statistically significant’ and ‘statistically non-significant’ need not be contradictory, and such designations might cause genuine effects to be dismissed.

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is ‘significant’ and the other is not.

Statistical vs Clinical Significance

Statistical Significance <> Scientific Significance:
Statistical Signicance , p < 0;05



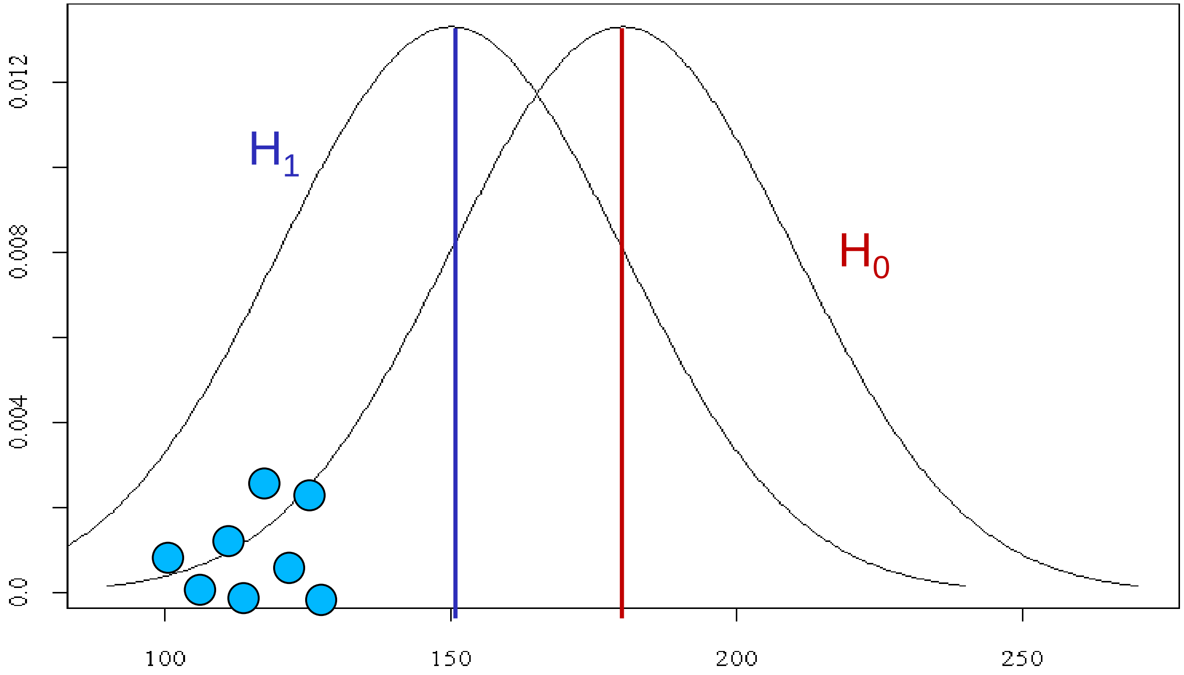
Clinical Significance	Statistical Significance
YES	YES
YES	YES
YES	NO
Depends	NO
Depends	YES
Depends	NO
Depends	NO
Depends	YES
YES(not use)	YES
NO	NO

https://www.researchgate.net/post/Whats_the_difference_between_statistical_and_practical_significance

The ASA Six Principles

1.- P-values can indicate how incompatible the data are with a specified statistical model.

every method of statistical inference relies on a **web of assumptions** which together can be viewed as a ‘statistical model’



P=0.045 if model assumptions are true and H0 is true

The ASA Six Principles

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..

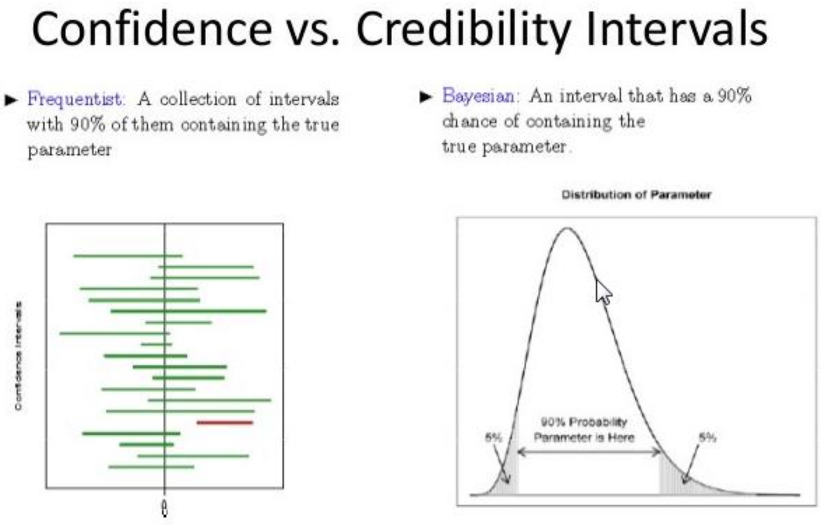
Probability (data observed /Ho) ≠Probability(Ho/ data observed)

What we get

What we want

$$P(H_0|Data) = \frac{P(Data|H_0)P(H_0)}{P(Data)}$$

reverend Thomas Bayes
(1702-1761)



Confidence vs. Credibility Intervals

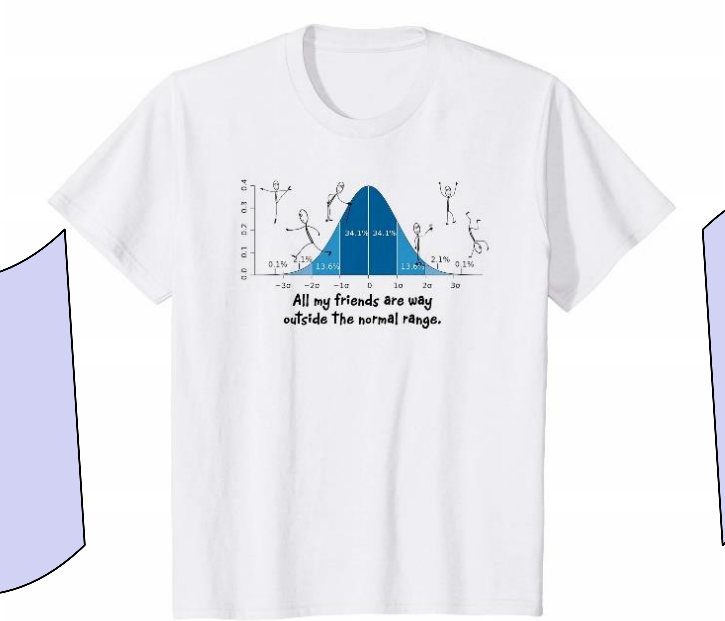
- Frequentist: A collection of intervals with 90% of them containing the true parameter
- Bayesian: An interval that has a 90% chance of containing the true parameter.

Courtesy: https://www.slideshare.net/ASQwebinars/bayesian-methods-in-reliability-engineering-15204318

The ASA Six Principles

3.- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold

If p=0.045 you cannot say you reject the hypothesis just because is under 0.05)



Price 19.999 €
Sure I will buy.
It is less than 20€

Price 20.001 €
Sure I will not buy.
It is over 20€

Absence of evidence is not evidence of absence

Douglas G Altman, J Martin Bland

BMJ VOLUME 311 19 AUGUST 1995

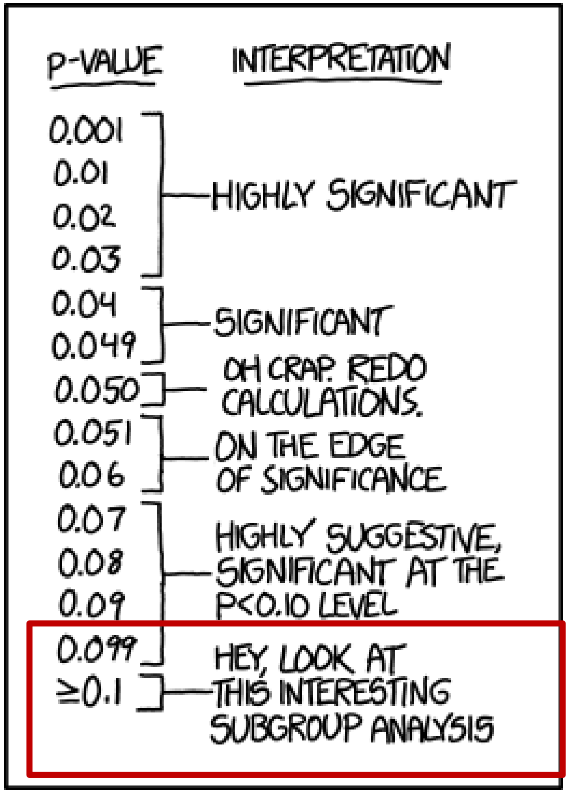
The ASA Six Principles

4.- Proper inference requires full reporting and transparency

- *P*-values and related analyses should not be reported selectively
- Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting

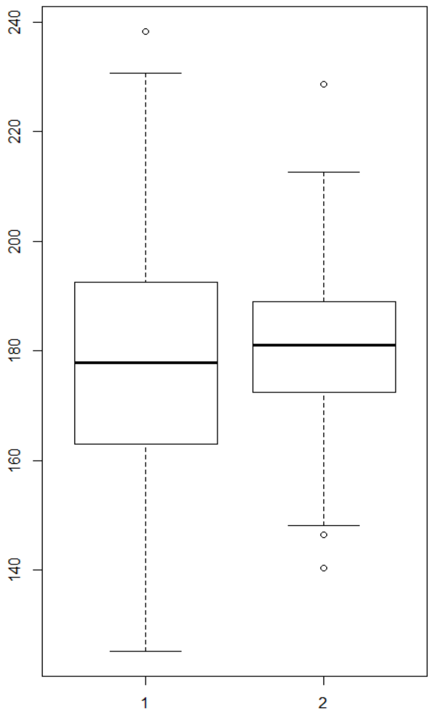
Be aware of:

- P-hacking
- “Fishing Expedition”
- Data dredgint
- Multiple Testing
- Multiplicity
- Significance chasing
- Significance questing
- Selective inference
- Etc.

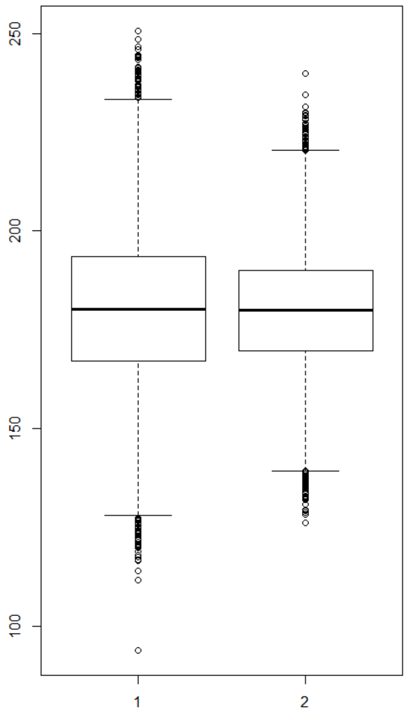

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

The ASA Six Principles

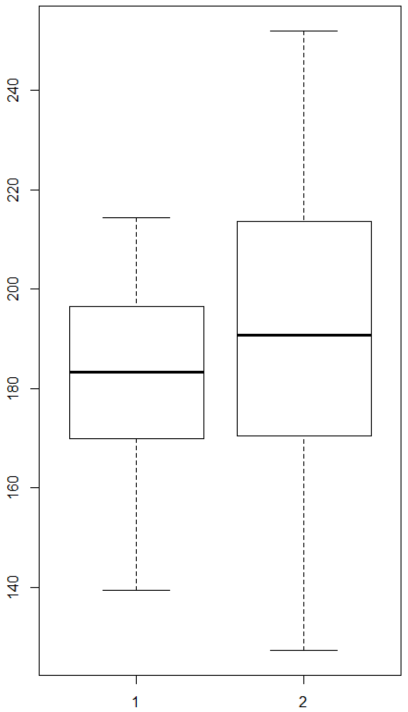
5.-A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.



N=100
Mean=180
Difference= -6.023
Value=0.6892



N=10000
Mean=180
Difference= 0.062
Value=0.02715



N=50
Mean=180 & 190
Difference= -16.906
Value=0.1045

The ASA Six Principles

6.-By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- Researchers should recognize that a *p*-value without context or other evidence provides limited information.
  - For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis.
  - Likewise, a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data.
- For these reasons, **data analysis should not end with the calculation of a *p*-value** when other approaches are appropriate and feasible”.

The ASA Six Principles

From a practical Point of view

1. Think about the underlying assumptions of your model
2. Avoid statements about the truth of tested hypotesis
3. Don’t do statements about the effect based on p value lower or higher 0.05
4. Don’t do sequence analyses reports and slicing results. Avoid “Data Torture”
5. Avoid statements of the intensity of effects based on differences on p-values
6. Use additional information than inferential results if feasible.

Finally

- Try to answer the question:
  - What could we do if we decided not to use p-values

Moving to a World Beyond “$p < 0.05$”

The American Statistician

Moving to a World Beyond “$p < 0.05$”

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

“Don’t” Is Not Enough

Use “less statistical significance” and more statistical Thinking

Don’t Say “Statistically Significant

Statistical inference is not—and never has been—equivalent to scientific inference

There Are Many Do’s

The statistical community has not yet converged on a simple paradigm for the use of statistical inference, But ther are solid principles for the use of statistics

ATOM Recomendations:

Accept Uncertainity

There is variation in effects. Confidence intervals , or better said "Compatibility intervals" should be the start point to seek for better measures , more sensitive desgins and large samples use

Be Thoughtful

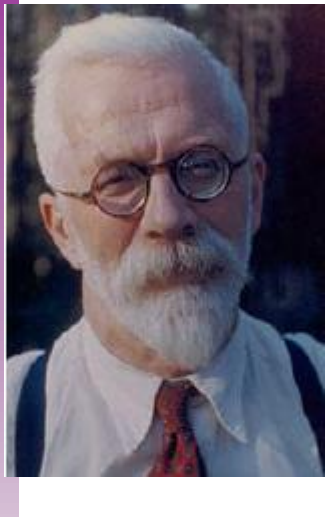
Think about Practical implications of the estimate, Precision in Estimates. Model correctly specified . Think before and Be flexible in conducting analysis

Be Open

In the development and presentation of research work. Be transparent and complete reporting results Provide exhaustive information in what, why and how you do it.

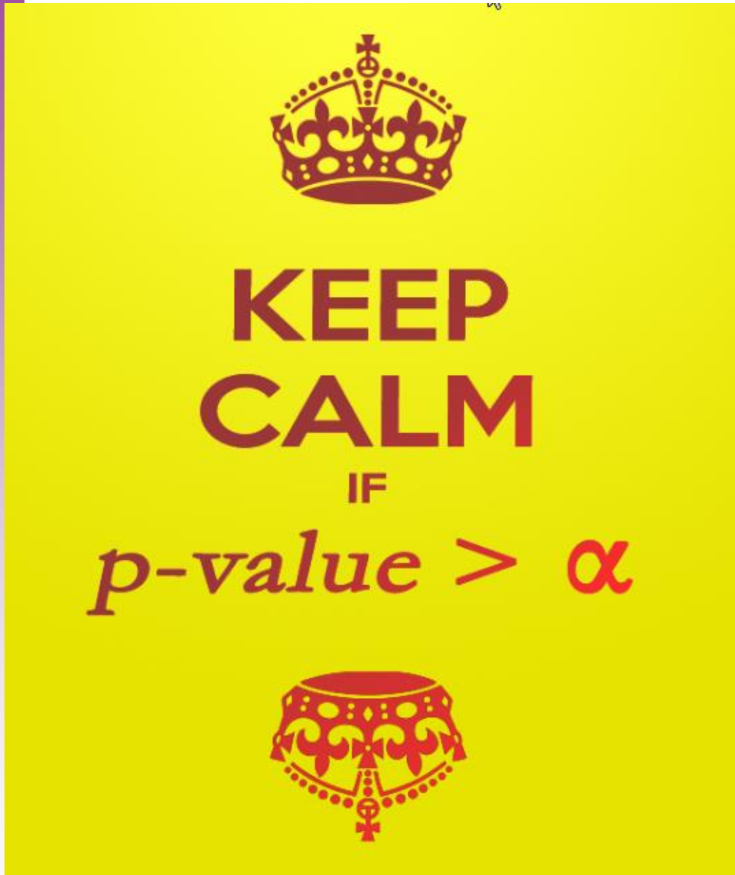
Be Modest

Express the limitations of your work , recognize that there are no true models. Statistics is not Reality.



"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."

Ronald Fisher



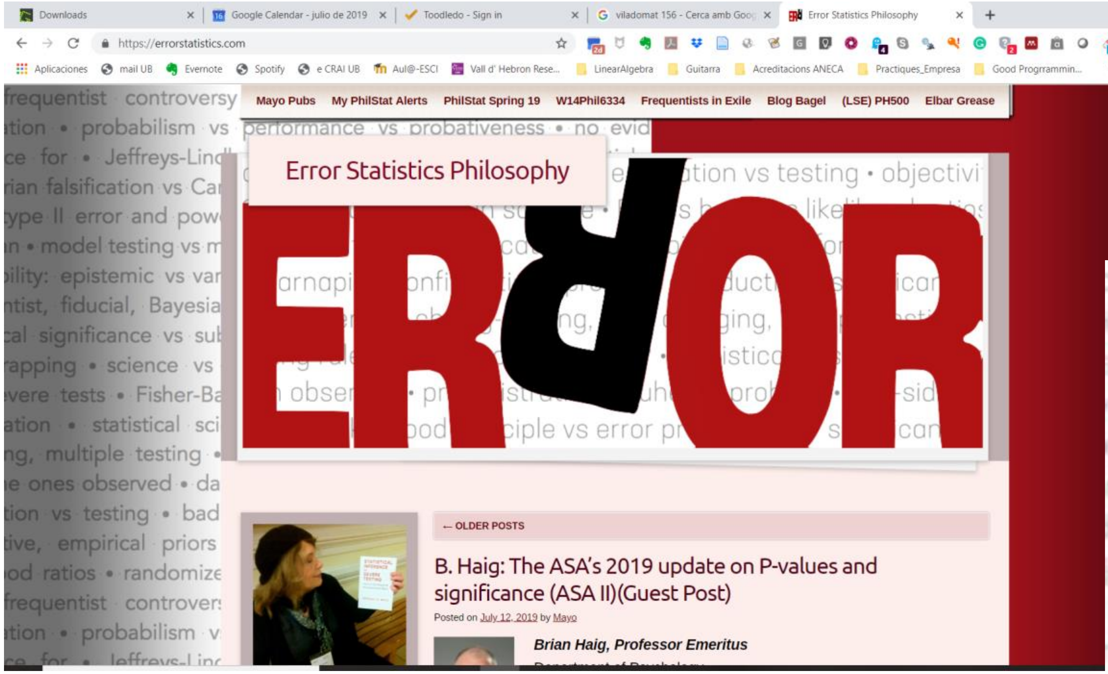
Thank you
Gracias
Gràcies

Find you in the next
Nos vemos en la pròxima
Ens veíem a la propera

Vall d'Hebron
Institut de Recerca
VHIR

ueb.vhir.org

Where to go for more …?

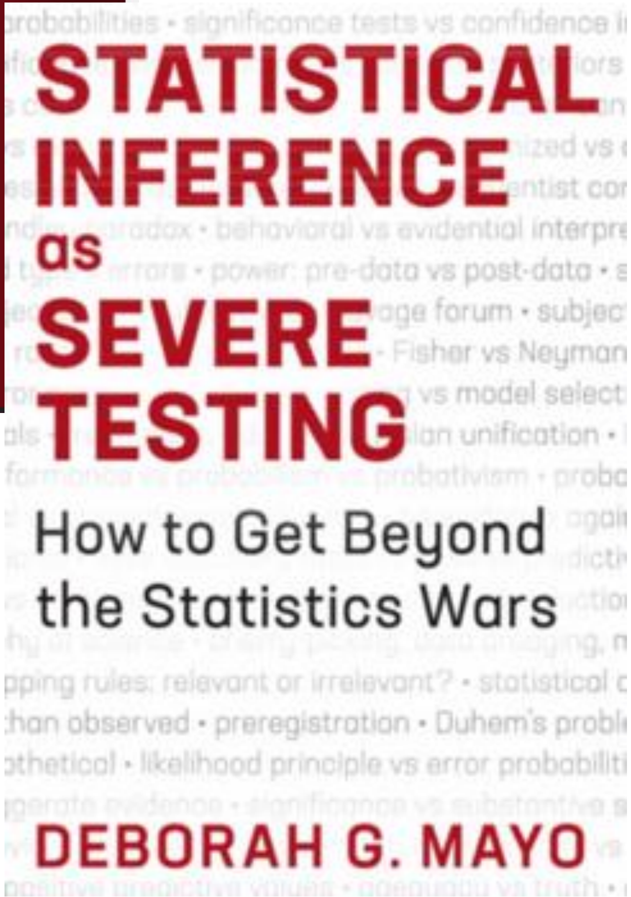


B. Haig: The ASA's 2019 update on P-values and significance (ASA II)(Guest Post)

Posted on July 12, 2019 by Mayo

Brian Haig, Professor Emeritus

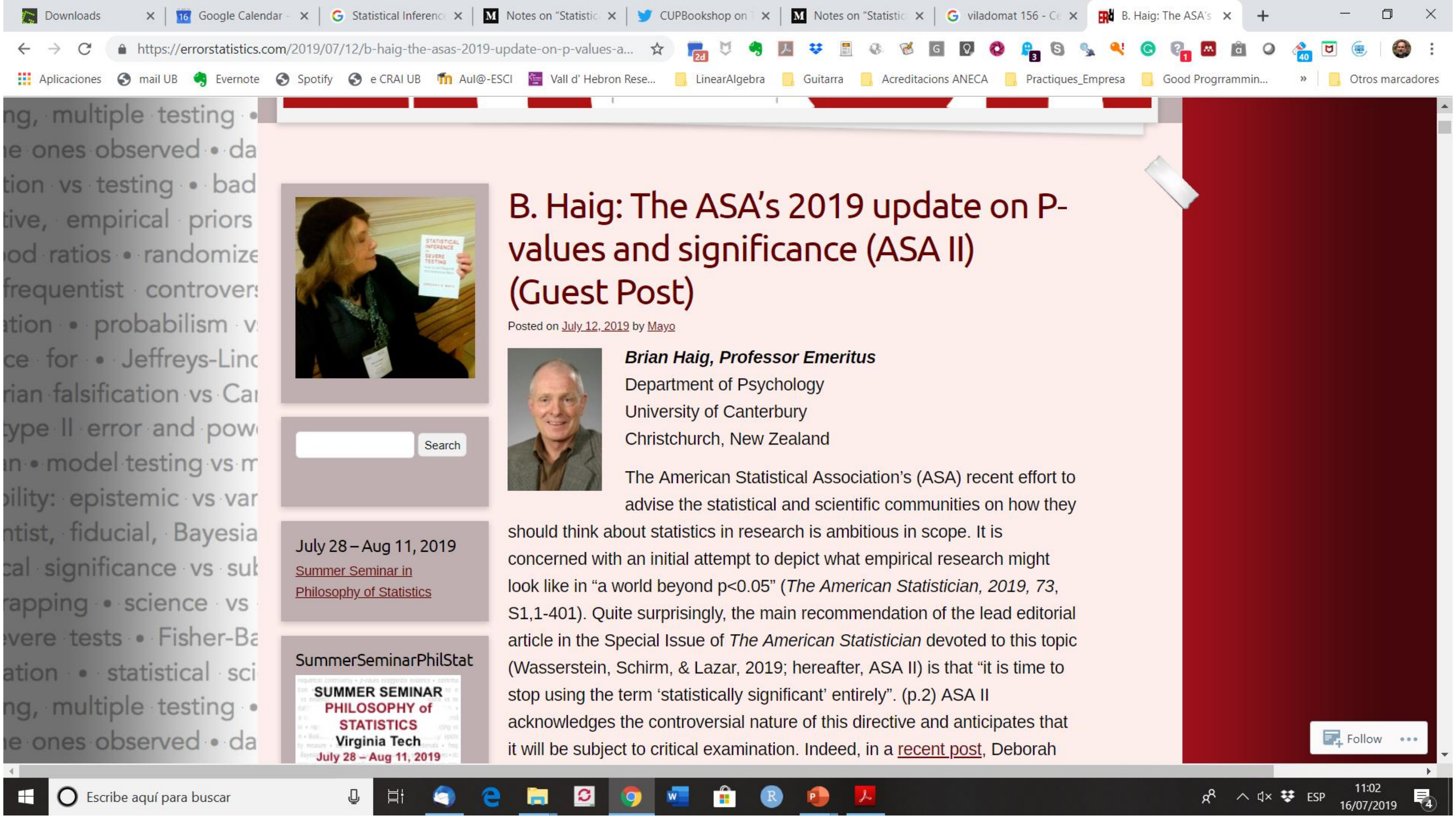
[Deborah Mayo: Error Statistics Philosophy](https://errorstatistics.com)


STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO

Where to go for more …?



July 28 – Aug 11, 2019

Summer Seminar in Philosophy of Statistics

SummerSeminarPhilStat

SUMMER SEMINAR PHILOSOPHY of STATISTICS Virginia Tech July 28 – Aug 11, 2019

B. Haig: The ASA's 2019 update on P-values and significance (ASA II) (Guest Post)

Posted on July 12, 2019 by Mayo

Brian Haig, Professor Emeritus
Department of Psychology
University of Canterbury
Christchurch, New Zealand

The American Statistical Association's (ASA) recent effort to advise the statistical and scientific communities on how they should think about statistics in research is ambitious in scope. It is concerned with an initial attempt to depict what empirical research might look like in "a world beyond p<0.05" (*The American Statistician, 2019, 73*, S1,1-401). Quite surprisingly, the main recommendation of the lead editorial article in the Special Issue of *The American Statistician* devoted to this topic (Wasserstein, Schirm, & Lazar, 2019; hereafter, ASA II) is that "it is time to stop using the term 'statistically significant' entirely". (p.2) ASA II acknowledges the controversial nature of this directive and anticipates that it will be subject to critical examination. Indeed, in a recent post, Deborah

Some references

‘It’s Time to Talk about Ditching Statistical Significance’. *Nature* 567, no. 7748 (20 March 2019): 283–283. https://doi.org/10.1038/d41586-019-00874-8.

Amrhein, Valentin, Sander Greenland, and Blake McShane. ‘Scientists Rise up against Statistical Significance’. *Nature* 567, no. 7748 (20 March 2019): 305–7. https://doi.org/10.1038/d41586-019-00857-9.

Goodman, Steven. ‘A Dirty Dozen: Twelve P-Value Misconceptions’. *Seminars in Hematology* 45, no. 3 (1 July 2008): 135–40. https://doi.org/10.1053/J.SEMINHEMATOL.2008.04.003.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. ‘Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations’. *European Journal of Epidemiology* 31, no. 4 (21 April 2016): 337–50. https://doi.org/10.1007/s10654-016-0149-3.

Laber, Eric B., and Kerby Shedden. ‘Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-Values for Statisticians’. *Journal of the American Statistical Association* 112, no. 519 (2017): 902. https://doi.org/10.1080/01621459.2017.1311265.

Mayo, Deborah G. *Statistical Inference as Severe Testing*. Cambridge University Press, 2018. https://doi.org/10.1017/9781107286184.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. ‘Moving to a World Beyond “ *p* &lt; 0.05”’. *The American Statistician* 73, no. sup1 (29 March 2019): 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wellek, Stefan. ‘A Critical Evaluation of the Current “ *p* -Value Controversy”’. *Biometrical Journal* 59, no. 5 (1 September 2017): 854–72. https://doi.org/10.1002/bimj.201700001.