The argument over p-values and the Null Hypothesis Significance Testing (NHST) paradigm

Matt Homer

m.s.homer@leeds.ac.uk

http://www.education.leeds.ac.uk/people/staff/academic/homer

This talk

- What prompted this talk - BASP?
- What is a p-value / NHSTP?
- What is the problem?
- Other quantitative methods problems
- What do the ‘experts’ say?
- What should we as researchers do?

What prompted this talk - BASP?

Editorial - Basic and Applied Social Psychology (BASP) - null hypothesis significance testing procedure (NHSTP) *'invalid*', and …

From now on, BASP is banning the NHSTP.

We hope and anticipate that banning the NHSTP will have the effect of ***increasing the quality*** *of submitted manuscripts by* ***liberating*** *authors from the* ***stultified*** *structure of NHSTP thinking thereby eliminating an important* ***obstacle to creative thinking.***

CSSME seminar – argument over p-values…

First

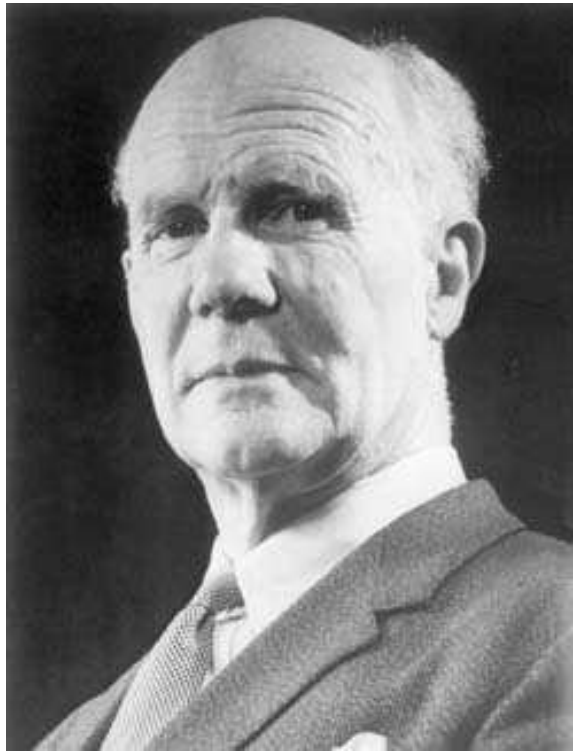
As you might expect….things are contested

NHSTP – what is it ? – at a glance

- Theory informed research question
- Null and alternative hypothesis – population
- Get some appropriate sample data
- Calculate test statistic
- Calculate associated p-value – likelihood of data under null
- Make a conclusion – p ‘small’ reject null, p ‘large’ accept – (will return to this later)

NHSTP – who started it all?

Ronald Fisher

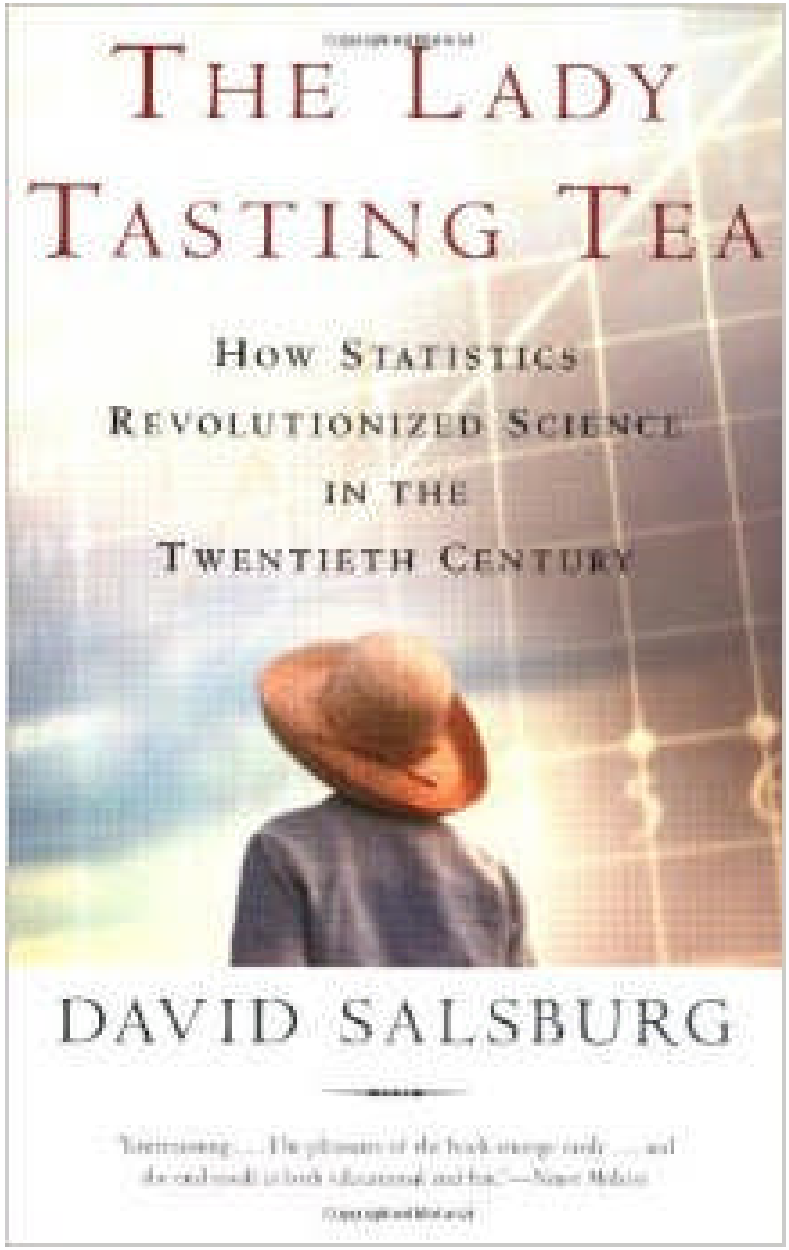


Jerzy Neyman and Egon Pearson

Early to mid 20th Century – foundations of statistical inference – big arguments about these…

A nice history of modern statistics

Why this title?



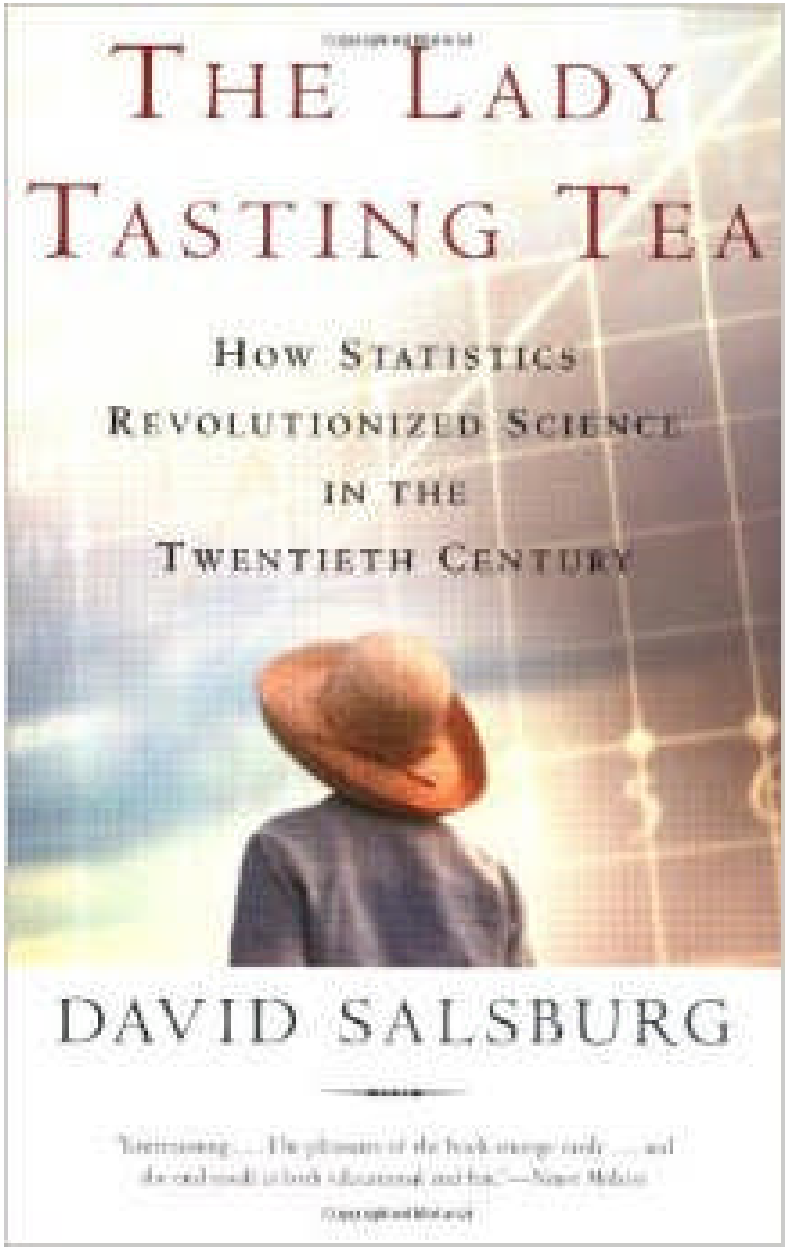
CSSME seminar – argument over p-values…

A nice history of modern statistics

Why this title?

- Randomized experiment – including the original Fisher **null hypothesis**.
- The ‘lady’ (Muriel Bristol) claimed to be able to tell whether the tea or the milk was added first to a cup.
- Eight cups, four of each variety, in random order - what is probability of her getting the number she got correct, but just by chance.

https://en.wikipedia.org/wiki/Lady_tasting_tea

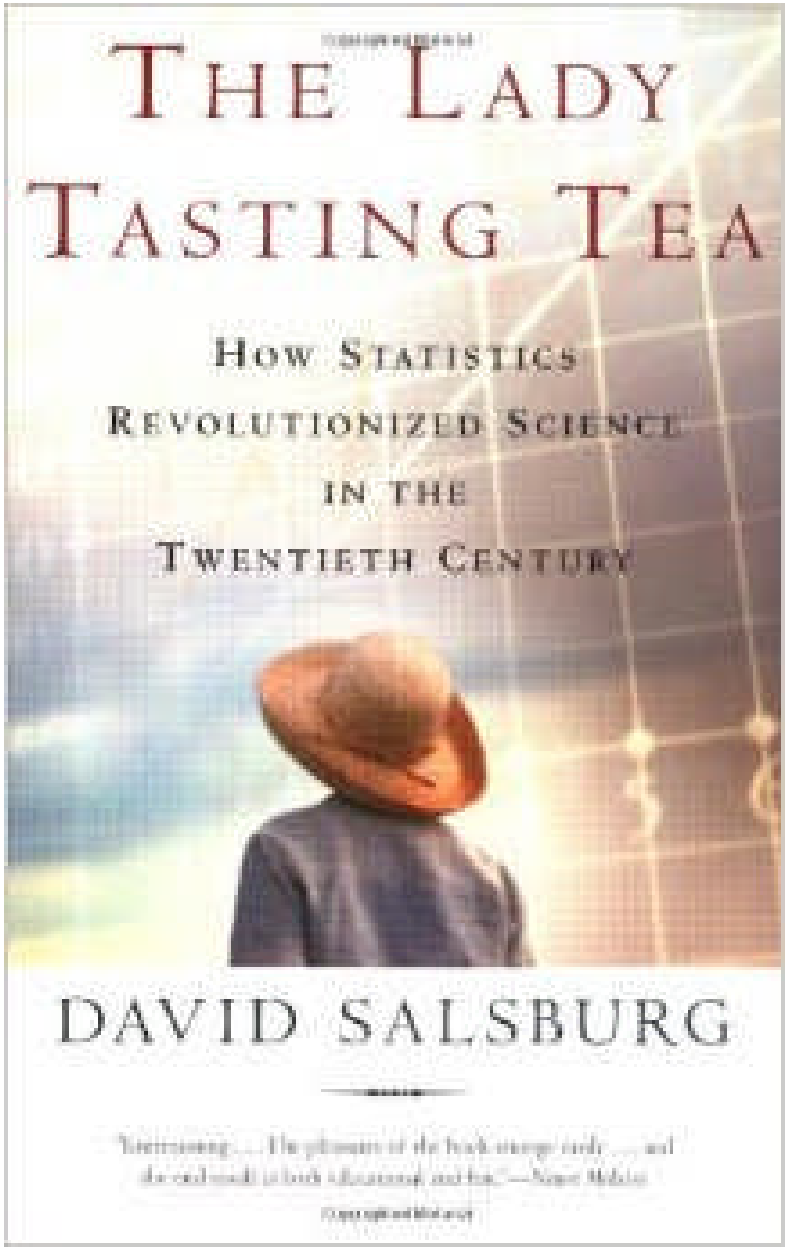


CSSME seminar – argument over p-values…

Where do you start to worry?

Correct 'guesses' in a row	p-value
1	0.500
2	0.250
3	0.125
4	0.063
5	0.031
6	0.016
7	0.008
8	0.004

Cut-off ~ 0.05???



NHSTP – how do you feel about it?

- Theory informed research question
- Null and alternative hypothesis
- Get some appropriate sample data
- Calculate test statistic
- Calculate associated p-value – likelihood of data under null
- Make a conclusion

- Are you confident that you understand it?
- Are you confident that you can use it properly?
- Where do your uncertainties lie?
- Why do you think BASP banned it?

5 minutes to discuss in groups and then feedback

The ban - some of the motivation?

- The technical details are widely misunderstood!
- The null is never(?) true
- Sample size issues
- It tells you about the data given the null, rather than the likelihood of the null given the data.
- Arbitrariness of cut-off (e.g. 5% level) – Cohen 1994
- There are logical problems – e.g. Gorard 2010, with response from Neale 2015

(Field, 4th edition, p60-62)

+ *General misuse – more later*

Other problems – poor research

- Over-emphasis on ‘significance’
- Data dredging/p-hacking – hunting for ‘significance’ – data mining, multiple testing, many models, picking ‘key’ outcomes, salami slicing (sub-groups) – all post hoc

  If you torture the data long enough, it will confess

  Ronald H. Coase

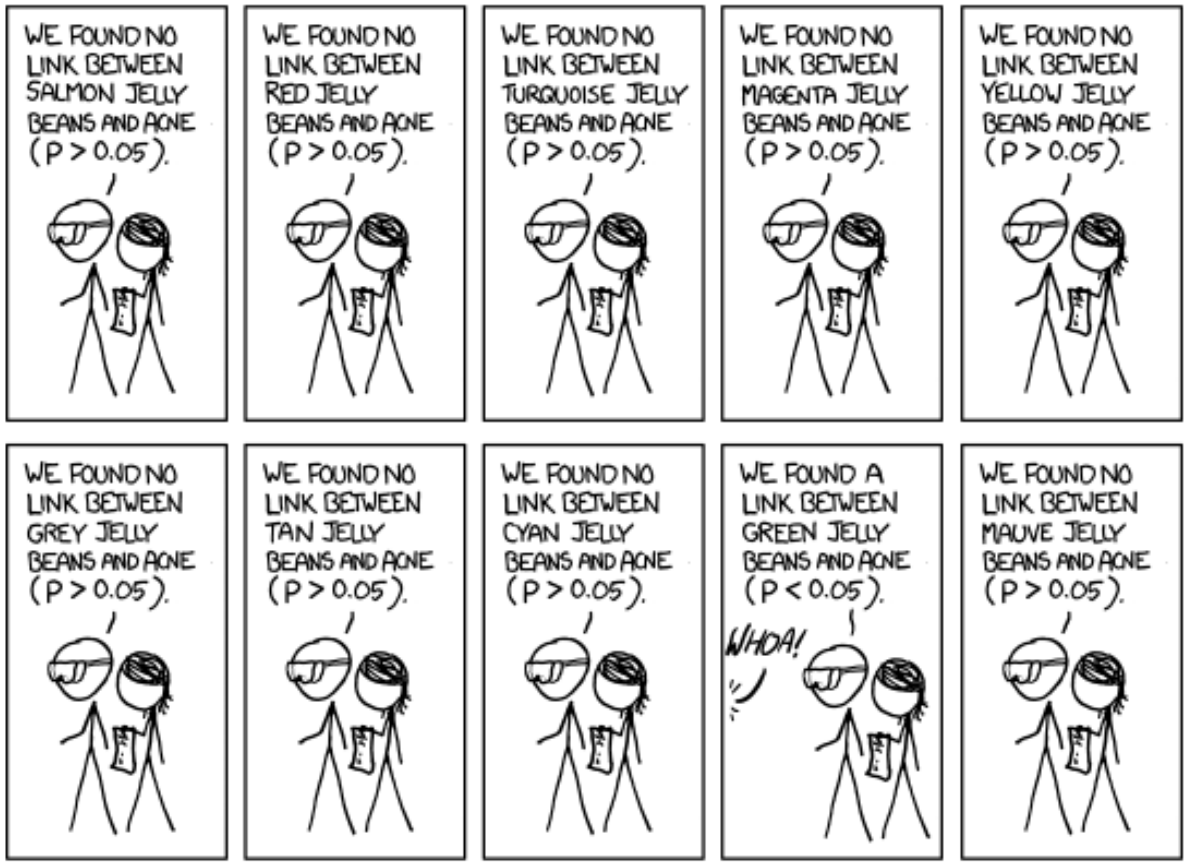
- Problems of replication – things don’t – false positives!

  Ioannidis 2005 – ‘most published findings are false’

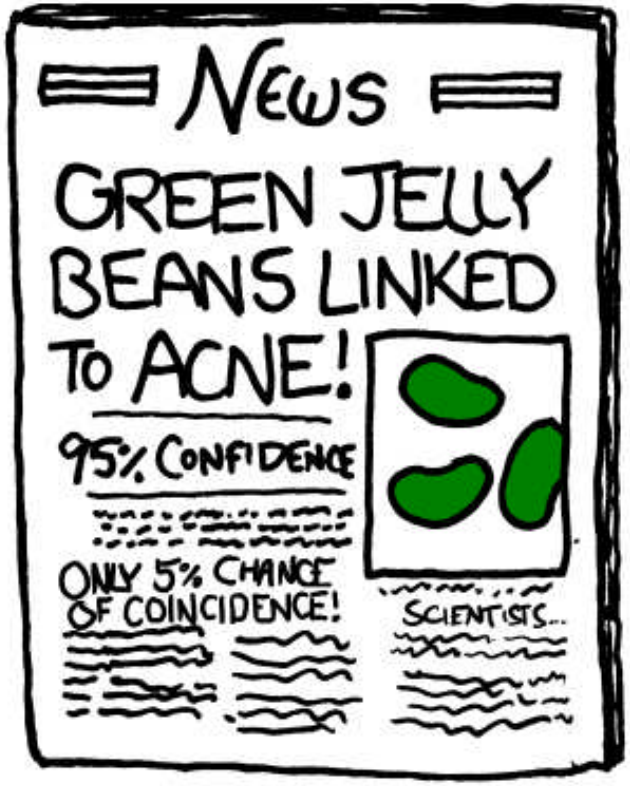
More general ‘meta-problems

- File draw problem – unpublished studies
- Under powered and/or biased studies - confounding

CSSME seminar – argument over p-values…



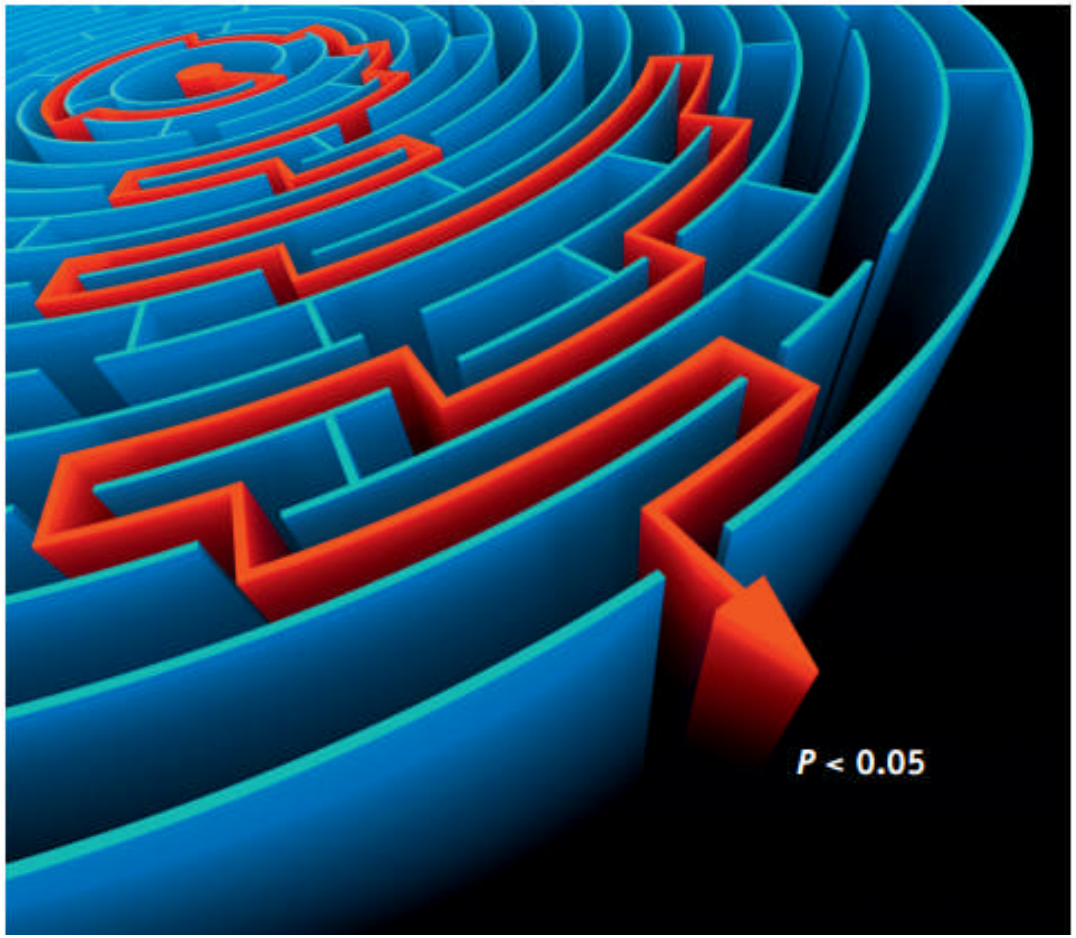
The fallacious link link between one variable with many levels and another: sub-group analysis

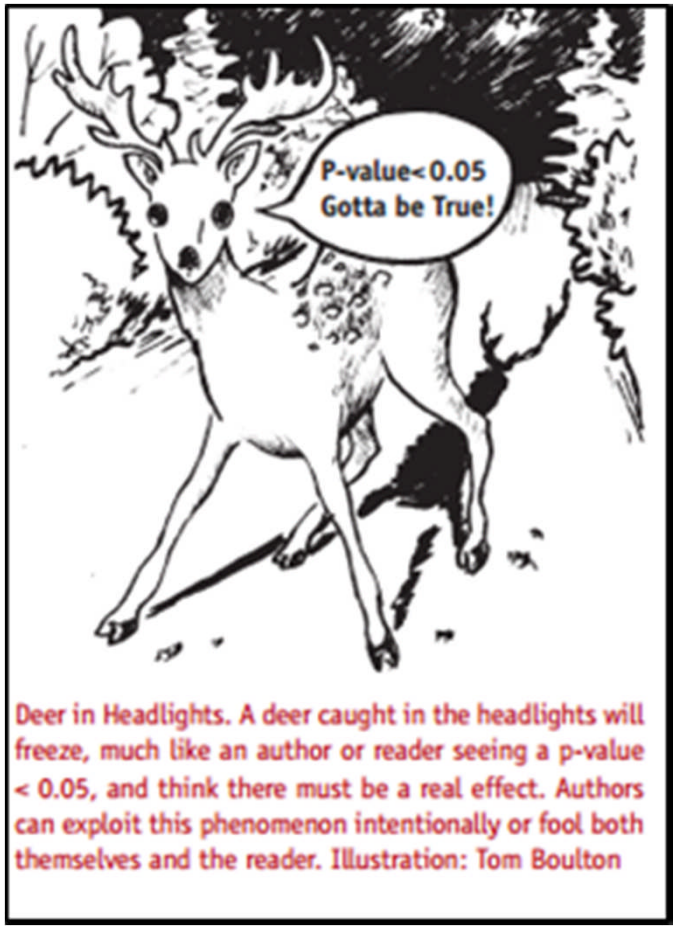


The pitfalls of multiple testing

http://xkcd.com/

CSSME seminar – argument over p-values…





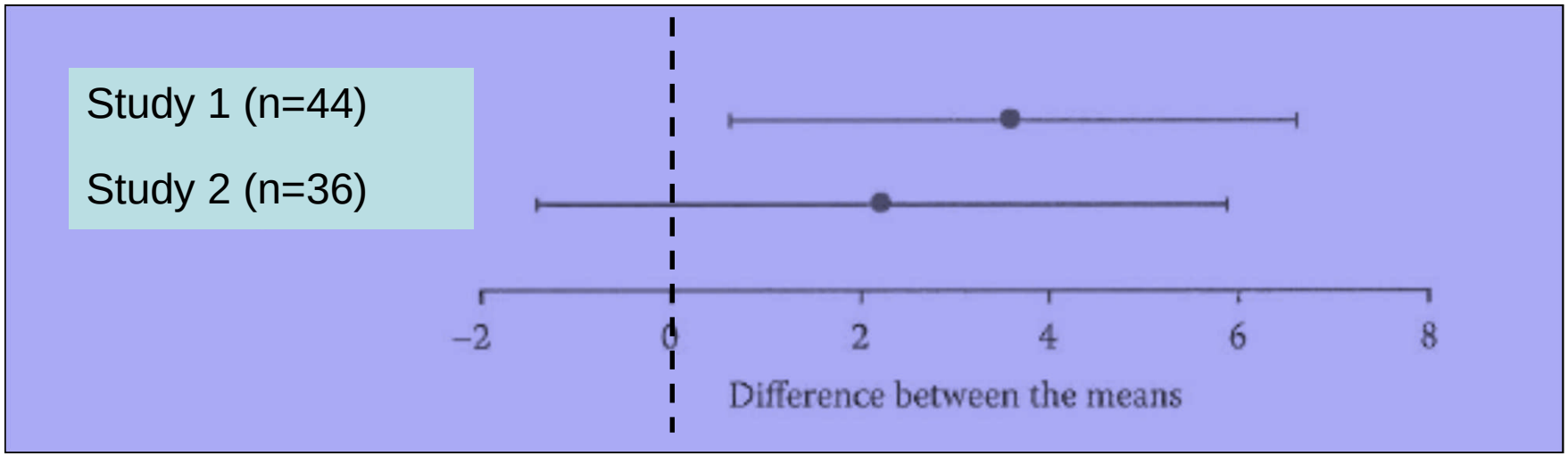
Post hoc manipulation to find ‘significance’ – torturing indeed

Fetishizing 5%’

Young, S.S. & Karr, A. (2011) Deming, data and observational studies. *Significance*, 8 (3), pp.116–120.

CSSME seminar – argument over p-values…

- Two studies of the same intervention are compared
- The 95% confidence intervals for the difference between control and intervention groups are shown for each



What would you conclude about the different results from the two studies?

Adapted from Cumming, G. (2011) *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York, Routledge.

What do the ‘experts’ say?

- Most say the BASP decision is silly
- Misused doesn’t mean wrong

I share the editors’ concerns that inferential statistical methods are ***open to mis-use and mis-interpretation****, but do not feel that a blanket ban on any particular inferential method is the most constructive response.*

Peter Diggle – President, Royal Statistical Society

CSSME seminar – argument over p-values…

Stephen Senn

If you don't make mistakes you don't learn. Attempting to eliminate ***false positives*** *in inference is to attempt scientific sterility and banning formal inferential methods won't even help to achieve this* ***foolish*** *aim.*

Head of Competence Center for Methodology and Statistics at the Luxembourg Institute of Health

CSSME seminar – argument over p-values…

Andrew Gellman

*I do like the idea of requiring that research claims **stand on their own** without requiring the (often spurious) support of p-values.*

Professor of statistics and political science and director of the Applied Statistics Center at Columbia University

http://andrewgelman.com/

Also author of: *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*

CSSME seminar – argument over p-values…

Robert Grant

*The trouble…is what **actions** the researcher takes on the basis of their many p-values. If we trained researchers to consider all **subjectivities** and **personal biases**, and to be open about them, in the way that **good qualitative researchers are**, far fewer errors would be made. A little dose of **philosophy of science** early on in training could help avoid common pitfalls later. A crude response like banning p-values serves as a fig leaf, because **the problem is in how researchers think.***

Senior lecturer in health and social care statistics at St George's, University of London and Kingston University

Try the exercises

- What do you think?

10 minutes to have a go and discuss in groups and then feedback

Feedback on the exercises

- Even(!) teachers get most of this wrong!

http://myweb.brooklyn.liu.edu/cortiz/PDF%20Files/Misinterpretations%20of%20Significance.pdf

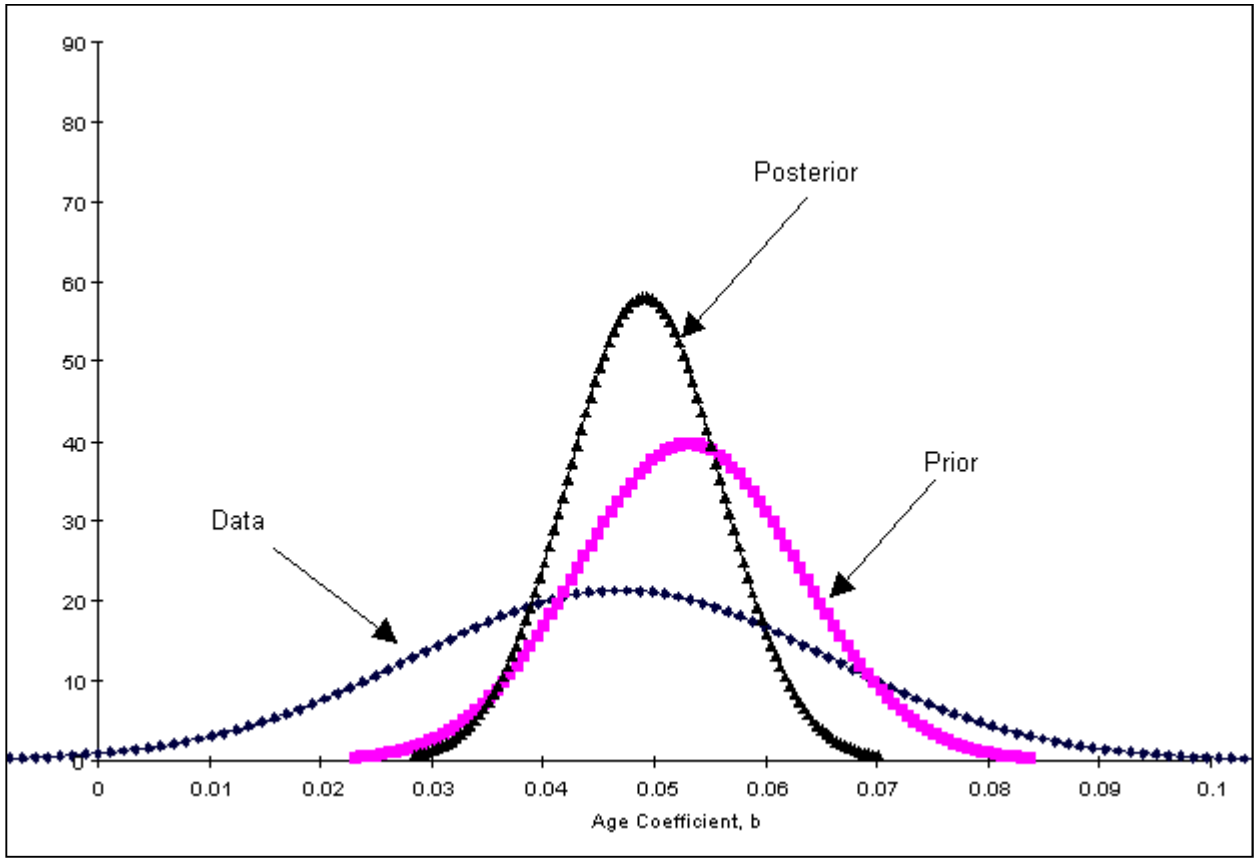
- Why is that?
- Should we worry?

My take

- P-values do tell you something useful
- But just part of story – not the focus
- Ban conflates two issues – p-values and bad research
- Arbitrary cut-offs are silly – ‘dance of the p-values’ – Cumming 2012
- Data visualisation, effect sizes, confidence intervals…much better
- ‘point estimates’ – hide the sampling error
- A process, not ‘the answer’ – evidence accumulation
- Bayesian methods ?????...

Bayesian approaches

- You can get the probability of hypothesis given data…
- But you need prior belief…
- And it's difficult software-wise/computationally



E.g. the effect of age on some outcome

Prior = what we believe age effect is prior to new data

Data = our sample

Posterior = Prior and data combined to update the age effect

Other non-statistical issues

- How much power should journal editors have?
- Should methods be part of the ‘rules’ of submission?
- Will the ban stick?
- Will it spread?
- What is the sociology of all this – are there bigger/hidden issues at play?
- What do you think on these issues?

5 minutes to discuss in groups and then feedback

What should we as researchers do?

Any ideas how we can make our quantitative research better?

Discuss – 5 minutes

Doing better quant. research

- Keep it simple
- Visualize
- Focus on effect sizes, confidence intervals
- Be theory driven
- Don’t ‘hunt’…null findings are findings (why?)
- Be selective – don’t test/try everything
- Include caveats and limitations
- Think! – validity – a process, not an end in itself – knowledge accumulating – fight isolationism

Select references

Cohen, J. (1994) The earth is round (p < .05). *American Psychologist*, 49 (12), pp.997–1003.

Cumming, G. (2011) *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York, Routledge.

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

Field, A.P. (2013) *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll*. London, Sage Publications.

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning. *Oxford Review of Education*, 36 (1), pp.63–77.

Neale, D. (2015) Defending the logic of significance testing: a response to Gorard. *Oxford Review of Education*, 41 (3), pp.334–345.

Norman, G. (2014) Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. *Advances in Health Sciences Education*, 19 (1), pp.1–5.

Salsburg, D. (2002) *Lady Tasting Tea*. 2 Reprint edition. New York, NY, Holt McDougal.

Trafimow, D. & Marks, M. (2015) Editorial. *Basic and Applied Social Psychology*, 37 (1), pp.1–2.

Young, S.S. & Karr, A. (2011) Deming, data and observational studies. *Significance*, 8 (3), pp.116–120.

Ziliak, S.T. & McCloskey, D. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, The University of Michigan Press.

Thanks…

Other thoughts??

Matt Homer: m.s.homer@leeds.ac.uk

http://www.education.leeds.ac.uk/people/staff/academic/homer