murdoch children's research institute

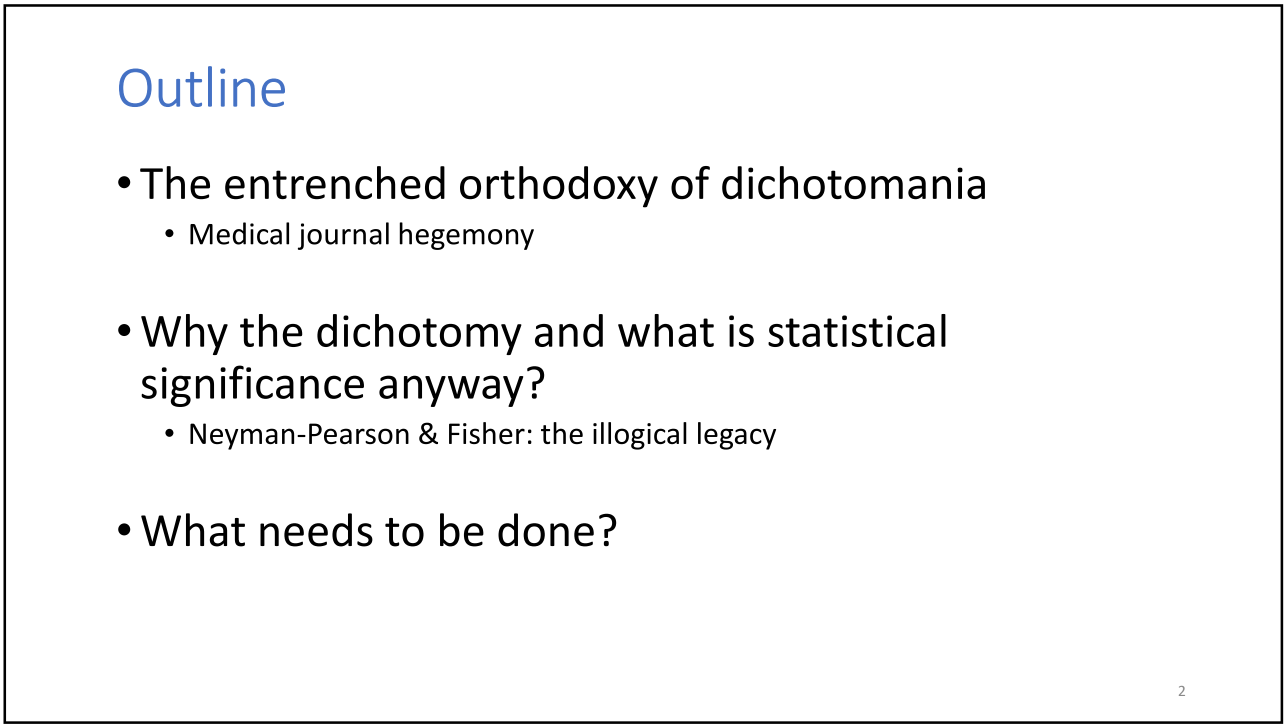
Beyond dichotomous thinking in clinical trials: why, how & who?

John Carlin

Murdoch Childrens Research Institute
& University of Melbourne

ISCB 2020
26-Aug-20

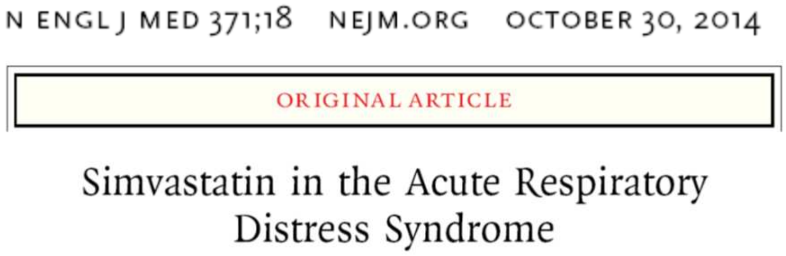
Melbourne Children's
Excellence in clinical care, research and education
The Royal Children's Hospital Melbourne
murdoch children's research institute
The University of Melbourne
ViCBiostat



Outline

- The entrenched orthodoxy of dichotomania
  - Medical journal hegemony
- Why the dichotomy and what is statistical significance anyway?
  - Neyman-Pearson & Fisher: the illogical legacy
- What needs to be done?

An example from *New England Journal of Medicine*



N ENGL J MED 371;18 NEJM.ORG OCTOBER 30, 2014

ORIGINAL ARTICLE

Simvastatin in the Acute Respiratory Distress Syndrome

Results:

"There was no significant difference between the study groups in the mean (±SD) number of ventilator-free days (12.6±9.9 with simvastatin and 11.5±10.4 with placebo, P = 0.21)…"

Conclusions:

"Simvastatin therapy, although safe and associated with minimal adverse effects, did not improve clinical outcomes…"

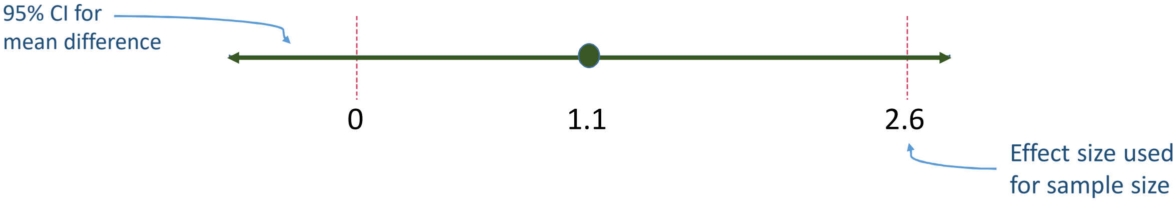
"… did not improve clinical outcomes…"

- Null hypothesis not rejected, therefore concluded "no effect"…
- BUT WAIT:

"Sample-size assumptions […] a sample of 524 patients […]
in order for the study to have 80% power, at a two-tailed significance level of 0.05, to detect a mean between-group difference of 2.6 ventilator-free days…"

"… did not improve clinical outcomes…"

- Let's look at the results again:



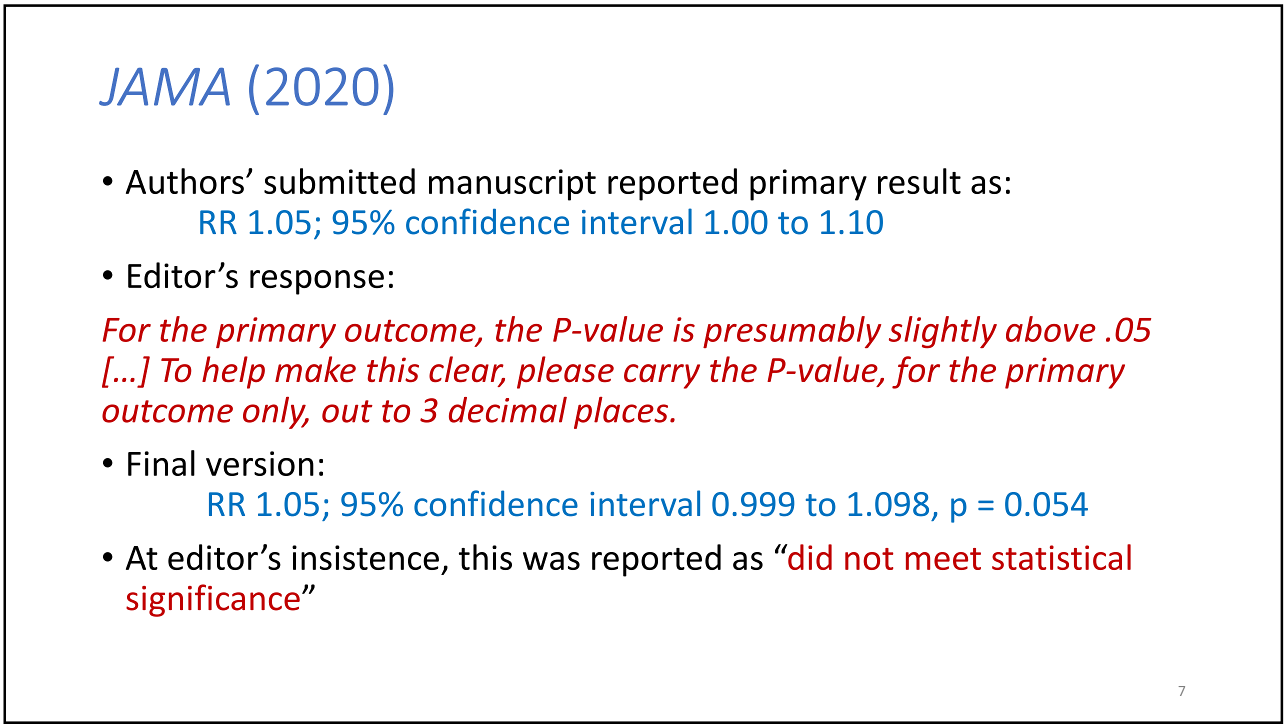
95% CI for mean difference

Effect size used for sample size

- 95% CI for mean difference: -0.6 to 2.8
- The alternative hypothesis should not be rejected either!!!

"… did not improve clinical outcomes…"

- Clearly a false dichotomy!
  - In fact, there was uncertain evidence for potential benefit
- Consequences:
  - Inevitable confusion about what the trial results mean:
    What happens next? Further research?
    Change clinical practice?
- Who is responsible?
  - Authors are usually following strict instructions from journals…



JAMA (2020)

- Authors’ submitted manuscript reported primary result as:

  RR 1.05; 95% confidence interval 1.00 to 1.10

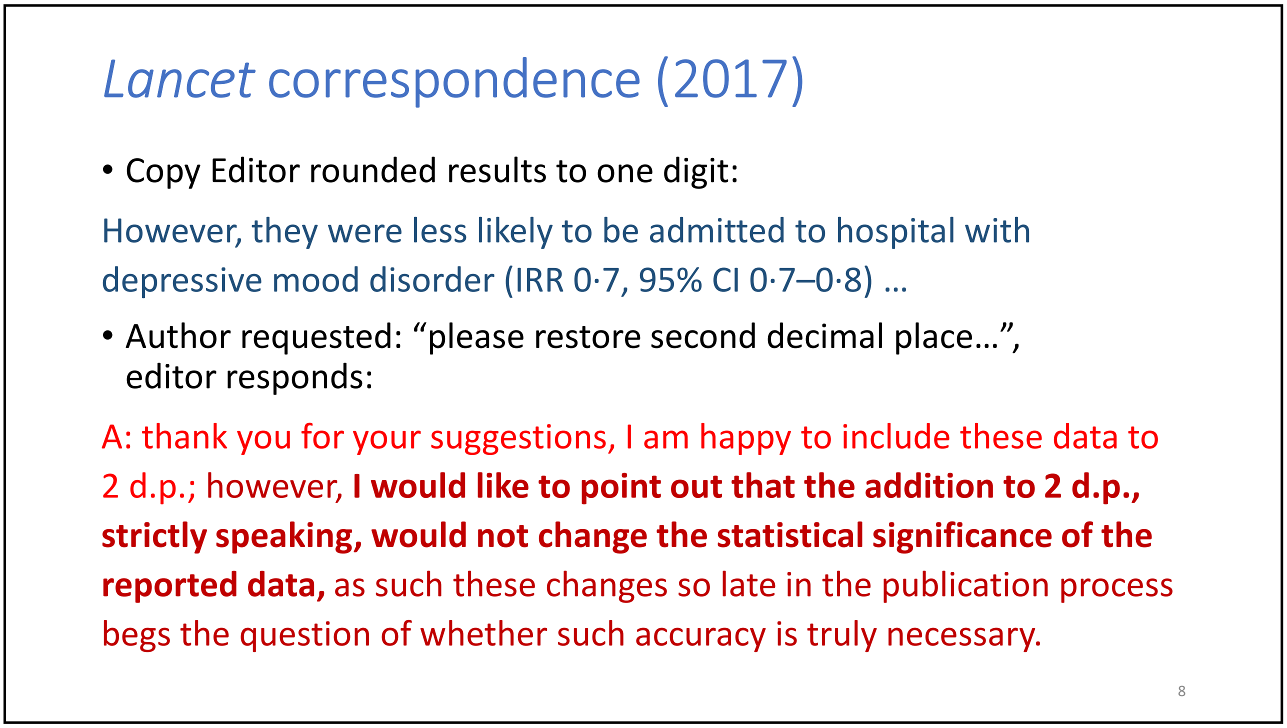
- Editor’s response:

For the primary outcome, the P-value is presumably slightly above .05 […] To help make this clear, please carry the P-value, for the primary outcome only, out to 3 decimal places.

- Final version:

  RR 1.05; 95% confidence interval 0.999 to 1.098, p = 0.054

- At editor’s insistence, this was reported as “did not meet statistical significance”



Lancet correspondence (2017)

- Copy Editor rounded results to one digit:

However, they were less likely to be admitted to hospital with depressive mood disorder (IRR 0·7, 95% CI 0·7–0·8) …

- Author requested: “please restore second decimal place…”, editor responds:

A: thank you for your suggestions, I am happy to include these data to 2 d.p.; however, **I would like to point out that the addition to 2 d.p., strictly speaking, would not change the statistical significance of the reported data,** as such these changes so late in the publication process begs the question of whether such accuracy is truly necessary.

JAMA again (2020)

Editorial correspondence to authors (at last stage before acceptance):

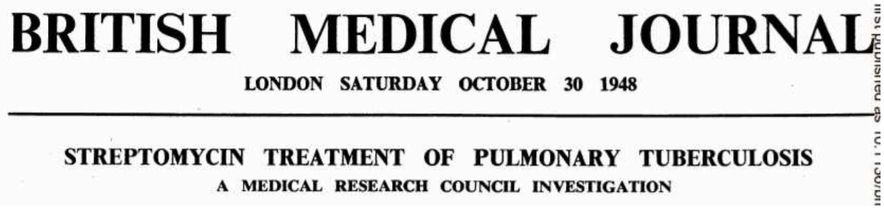
With regard to your request to change “no association” back to “not enough evidence,” I asked my supervisor and this was her response:

“If they want to say ‘evidence of association’ and ‘no evidence of association,’ I think that’s ok. If they want to report P values, they do need to define a level of significance.
In general, I think the issues around using P < .05 as a cutoff are pretty well understood by our readers. ”

🤔

Why the dichotomy?

- It goes back a long way…



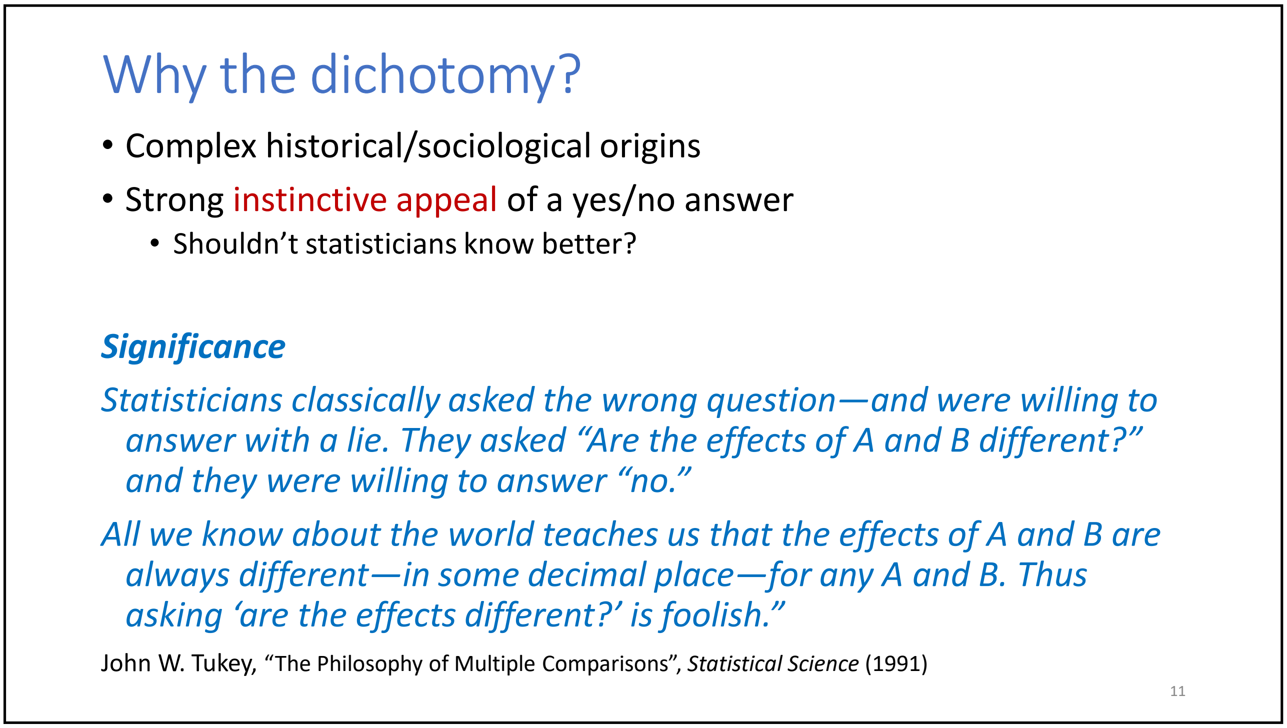
BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

“Four of the 55 S patients (7%) and 14 of the 52 C patients (27%) died before the end of six months. The difference between the two series is statistically significant; the probability of it occurring by chance is less than one in a hundred.”



Why the dichotomy?

- Complex historical/sociological origins
- Strong instinctive appeal of a yes/no answer
  - Shouldn't statisticians know better?

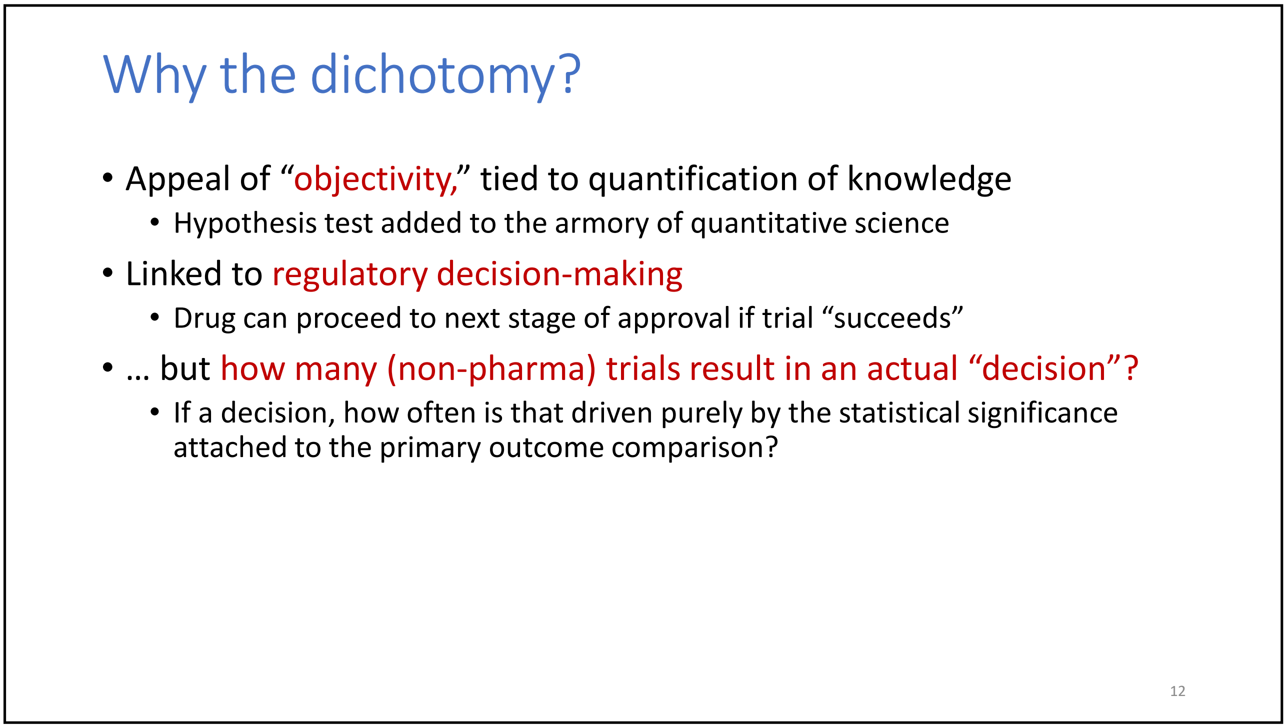
Significance

Statisticians classically asked the wrong question—and were willing to answer with a lie. They asked "Are the effects of A and B different?" and they were willing to answer "no."

All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'are the effects different?' is foolish."

John W. Tukey, "The Philosophy of Multiple Comparisons", *Statistical Science* (1991)

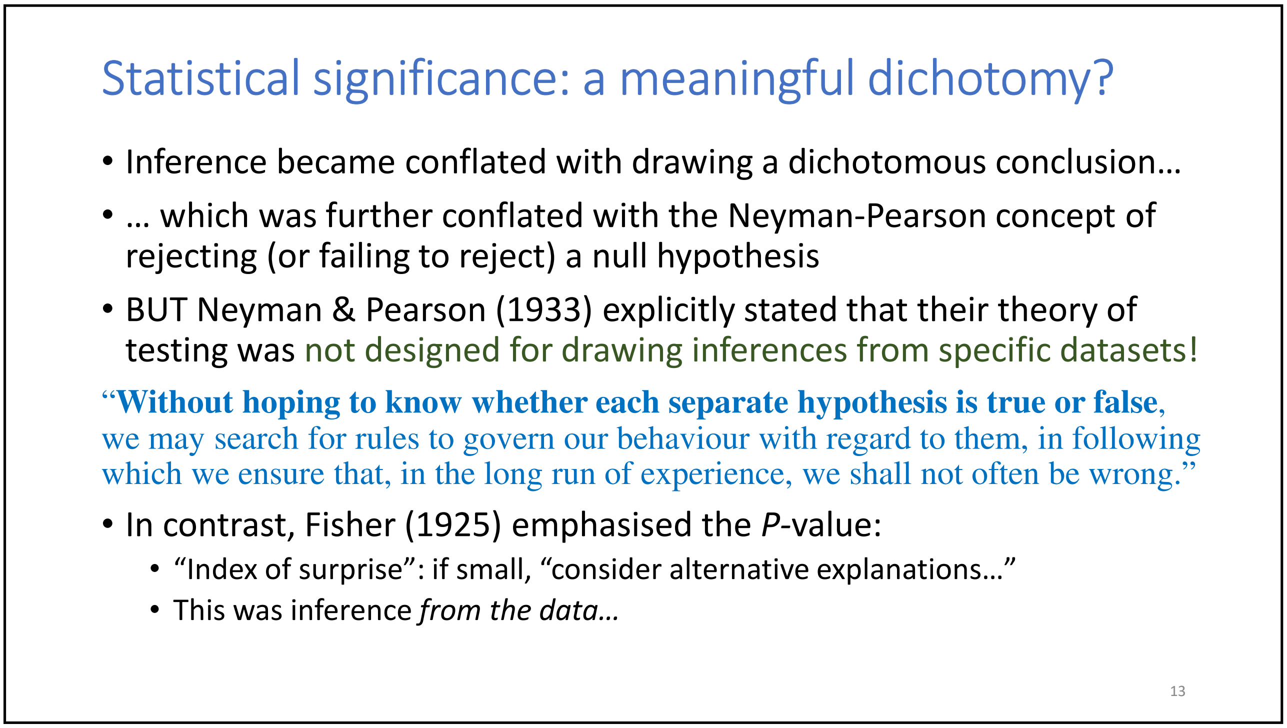
11



Why the dichotomy?

- Appeal of "objectivity," tied to quantification of knowledge
  - Hypothesis test added to the armory of quantitative science
- Linked to regulatory decision-making
  - Drug can proceed to next stage of approval if trial "succeeds"
- … but how many (non-pharma) trials result in an actual "decision"?
  - If a decision, how often is that driven purely by the statistical significance attached to the primary outcome comparison?

12

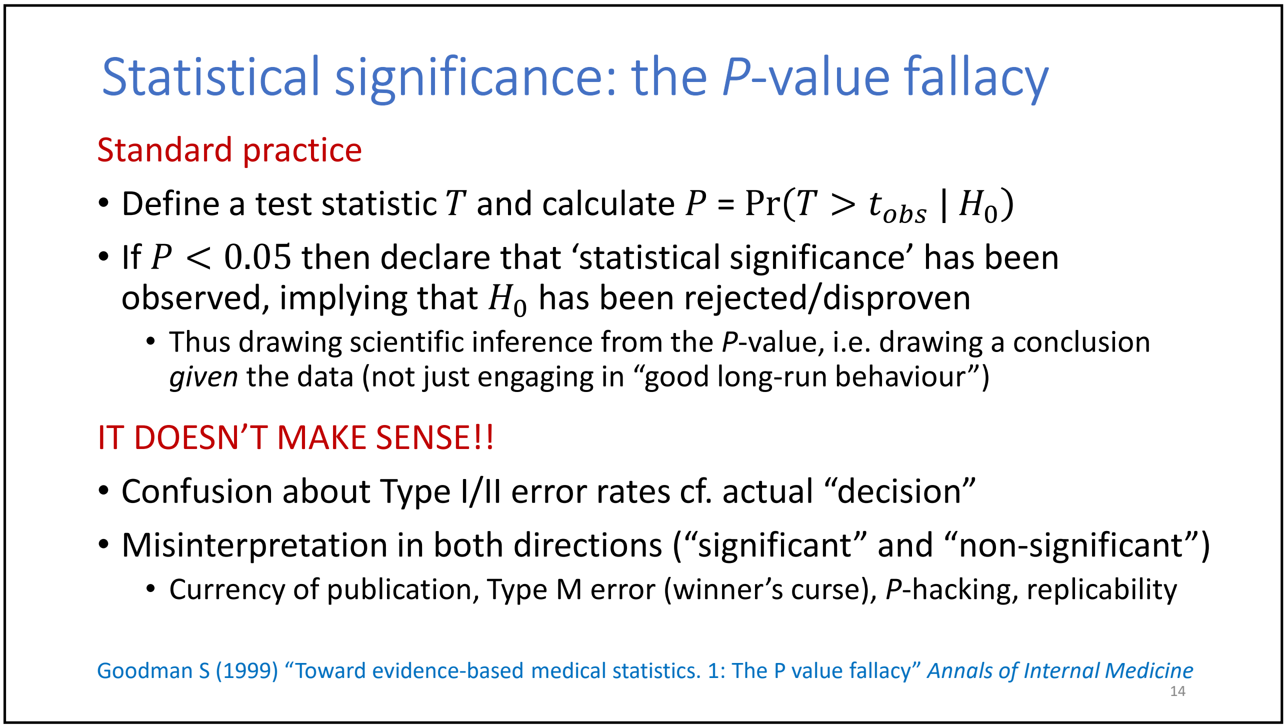


Statistical significance: a meaningful dichotomy?

- Inference became conflated with drawing a dichotomous conclusion…
- … which was further conflated with the Neyman-Pearson concept of rejecting (or failing to reject) a null hypothesis
- BUT Neyman & Pearson (1933) explicitly stated that their theory of testing was not designed for drawing inferences from specific datasets!

“**Without hoping to know whether each separate hypothesis is true or false**, we may search for rules to govern our behaviour with regard to them, in following which we ensure that, in the long run of experience, we shall not often be wrong.”

- In contrast, Fisher (1925) emphasised the *P*-value:
  - “Index of surprise”: if small, “consider alternative explanations…”
  - This was inference *from the data…*



Statistical significance: the *P*-value fallacy

Standard practice

- Define a test statistic T and calculate P = $\Pr(T > t_{obs} \mid H_0)$
- If $P < 0.05$ then declare that ‘statistical significance’ has been observed, implying that H_0 has been rejected/disproven
  - Thus drawing scientific inference from the *P*-value, i.e. drawing a conclusion *given* the data (not just engaging in “good long-run behaviour”)

IT DOESN’T MAKE SENSE!!

- Confusion about Type I/II error rates cf. actual “decision”
- Misinterpretation in both directions (“significant” and “non-significant”)
  - Currency of publication, Type M error (winner’s curse), *P*-hacking, replicability

Goodman S (1999) “Toward evidence-based medical statistics. 1: The P value fallacy” *Annals of Internal Medicine*

We know all these things?!

- *P* < .05 does NOT mean an effect is real
- *P* > .05 does NOT mean there is no effect

(And numerous other common misunderstandings!)

- So why do we allow such things to be said (and believed by non-statistical colleagues and general public)?

- Hypothesis tests in trials “not as bad” as elsewhere?
  - Stricter rules enforce some discipline re primary analysis, power, multiple comparisons etc
  - But still fundamentally illogical and misleading

We know all these things?
We’ve heard it all before!

However…

- In fact, not all statisticians are immune to the misunderstandings (are you?)
- We shouldn’t expect other scientists to understand a flawed theory: it’s our job to own and acknowledge the problems
- We need an action plan!

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

http://xkcd.com/892/

Reforms needed in practice

- De-emphasise hypothesis testing as the primary statistical activity
  - Encourage / teach interval estimation
  - … and probably more Bayesian inference

- Emphasise that inference in the face of random variation must be uncertain, not dichotomous
  - Insert qualifying words in conclusions to remind of the grey zones: “some evidence…”, “appears to…”, “suggests that…”
  - Teach reporting of data analysis results as incremental information not “findings”
  - Teach the facts, i.e. that hypothesis testing is a *flawed concept!* (both in training of statisticians and for consumers of statistical methods)

Example from my teaching

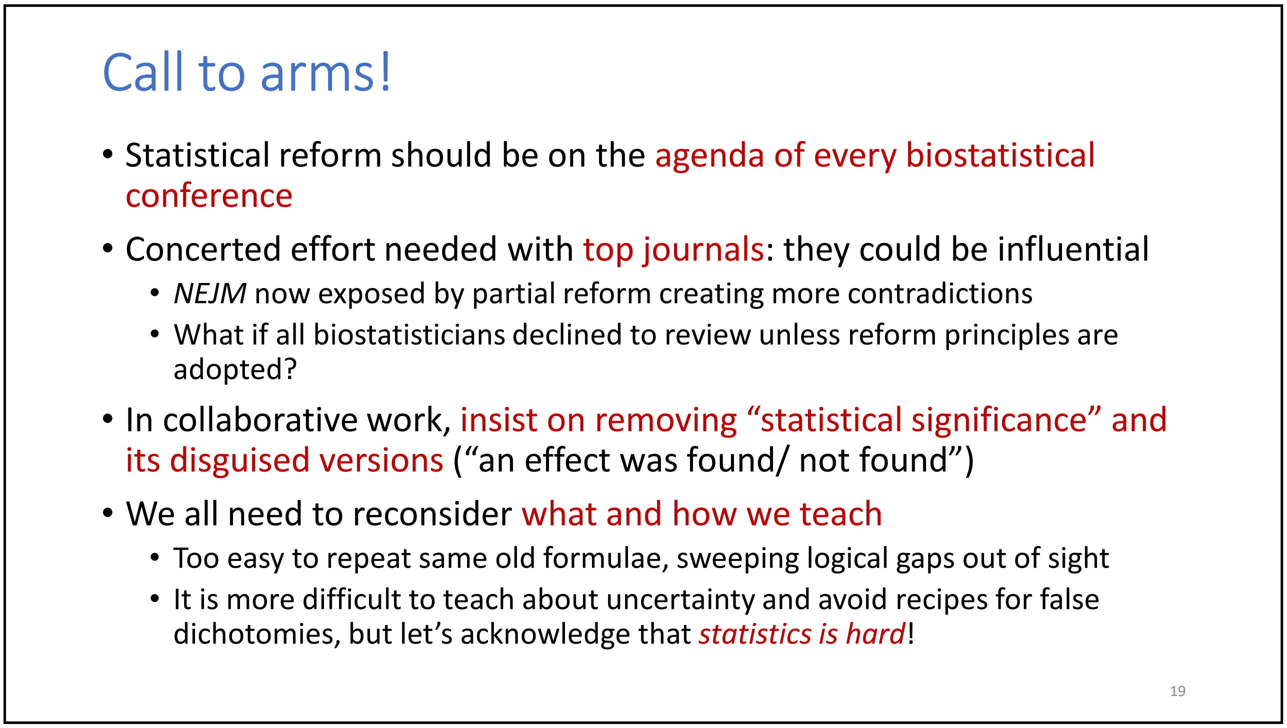
Inference Methods for Biostatistics

Week 7

Statistical inference in practice:
critical review and guidance

Outline

- Recap concepts & “classical” methods for hypothesis testing
  - Review example
- Disconnect between theory and reality
  - *P*-value almost universally misunderstood
- The common practice that interprets statistical significance (“$p < 0.05$”) as equivalent to “scientific proof” has *many* problems
  - 25 misinterpretations of *P*-values (Greenland, 2016)
  - Type S and Type M errors (Gelman, 2014)
  - Multiplicity and “*P*-hacking”



Call to arms!

- Statistical reform should be on the agenda of every biostatistical conference
- Concerted effort needed with top journals: they could be influential
  - *NEJM* now exposed by partial reform creating more contradictions
  - What if all biostatisticians declined to review unless reform principles are adopted?
- In collaborative work, insist on removing “statistical significance” and its disguised versions (“an effect was found/ not found”)
- We all need to reconsider what and how we teach
  - Too easy to repeat same old formulae, sweeping logical gaps out of sight
  - It is more difficult to teach about uncertainty and avoid recipes for false dichotomies, but let’s acknowledge that *statistics is hard*!