p-values: beyond polemics to practical proposals

James Carpenter

London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk

Presented by: Carsten Schmidt

Wed 15th July, IBC Seoul

STRATOS INITIATIVE

Acknowledgements

Michal Abrahamowicz (McGill, Canada)

Els Goetghebeur (Ghent, Belgium)

Willi Sauerbrei (Freiburg, Germany)

Victor Kipnis (NIH, USA)

Carsten Schmidt (Greifswald, Germany)

on behalf of the STRATOS p-values working group

Overview

- Criticism of use and reporting of p-values
- Reproducibility and p-values
- Use of p-values in three broad modelling aims
  - Descriptive
  - Predictive
  - Explanatory / Causal
- Example: descriptive analysis
- Discussion

3

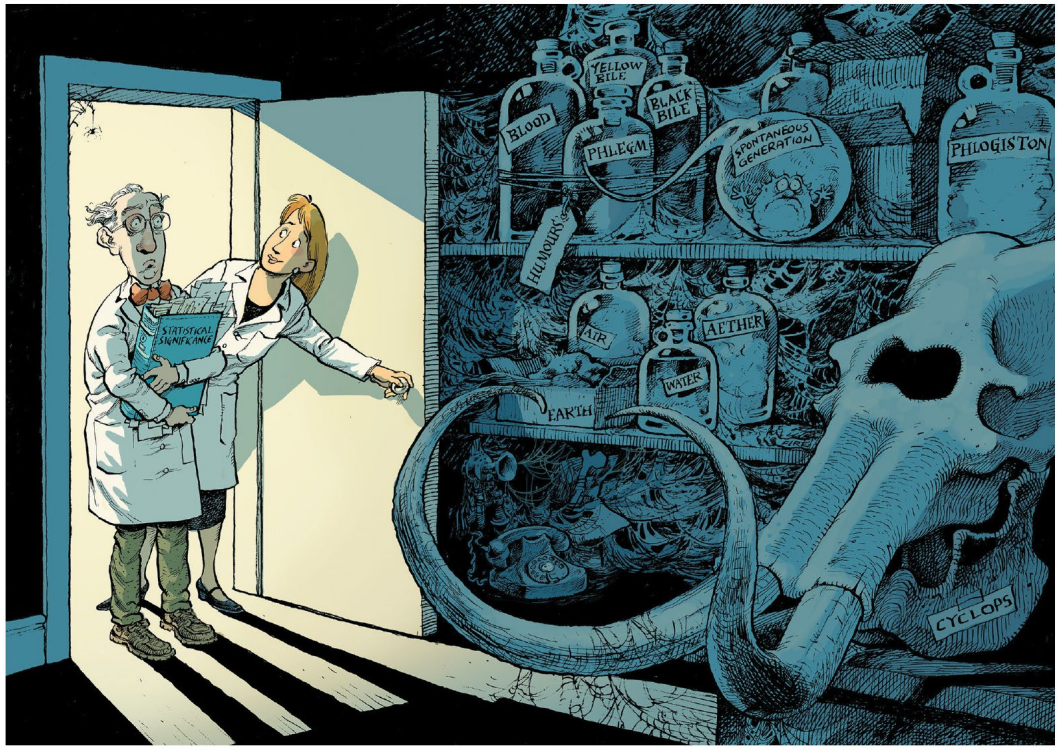
Critique of p-values

- Large literature on misuse of p-values
  - 2019 paper by Amrhein et al. [1] arguing for retiring ‘statistical significance’
  - ‘confidence interval’ should be renamed to ‘compatibility interval ’
  - Emphasize the need to focus on estimates and the uncertainty in them, including p-values

4

Critique of p-values

- Large literature on misuse of p-values
  - 2019 paper by AMR [1] a…e'
  - 'confidence interval' sh…val '
  - Emphasize the need to …in them, including p-values



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.

Critique of p-values

- Large literature on misuse of p-values
  - 2019 paper by AMR [1] arguing for retiring ‘statistical significance’
  - ‘confidence interval’ should be renamed to ‘compatibility interval ’
  - Emphasize the need to focus on estimates and the uncertainty in them, including p-values

- Well written, but more guidance needed
  - Principal issue is lazy use and (deliberately?) misleading reporting

6

Some consequences

- Over-interpretation of point estimates may be an issue:
  - E.g. [2], Methods: “To avoid placing too much emphasis on statistical significance, we emphasize the strength of associations… as well [cite AGM].”
  - Then, Results section explicitly interprets OR = 5.456, with 95% CI: 0.3–90.6 as “5 times” risk increase
  - Re-constructed p-value is about 0.24
- More driving change needed: Amrhein et al. cited but advice often ignored:
  - E.g. [3] cite AGM, but then report “significant associations” or “effects” in the conventional way:

    ‘… AI/AN women were found to be significantly more likely to have a high-risk (OR=1.28; 95% CI: 1.01-1.66)’

Can STRATOS help?

- STRATOS stands for Strengthening Analytical Thinking for Observational Studies
  - Hypothesis testing, inference and reporting are central to this
  - We should have something to say
- Discussion highlighted concern that the AMR paper could be used by those who wish to jettison more formal inference
  - The current situation may be bad, but it could be worse!
- Therefore, we agreed to develop a manuscript which aimed to provide more practical guidance
  - We concluded the most appropriate way was to illustrate the principled use of p-values, testing and statistical significance in carefully chosen examples

8

Reproducibility and p-values

Reproducibility is, rightly, an increasing focus of efforts to improve research. However, many researchers are not aware that the p-value is not a reproducible statistic.

For a sample mean from a population with known variance 1, the table shows the probability of the p-value from a second experiment being < 0.05, if the true mean is equal to that observed in the first experiment:

p-value from 1st experiment:	**0.05**	**0.01**	**0.005**	**0.001**
Pr(p<0.05) in repeat of experiment:	0.5	0.73	0.8	0.91

9

Use of p-values & significance testing in research

Large phase-III randomised trials are used to decide whether to adopt new treatments – in this setting, formal hypothesis testing is broadly agreed to be appropriate: reproducibility comes from the design which controls the type-1 error and power.

However, observational research has much broader aims, main aims are classified as [e.g., 4]:

- Descriptive
- Predicative modelling
- Explanatory / Causal inference

10

STRATOS INITIATIVE

p-values supporting reproducible research:

Follow reporting guidelines [e.g., 6], especially relating to

- Presenting p-values with parameter estimates and confidence intervals.
- Not interpreting $p \geq \alpha$ as evidence of the 'lack of effect' but, instead, as 'insufficient evidence to reject the null hypothesis'.
- Not interpreting parameter estimates, confidence intervals, and p-values isolated from previous / related findings
- Not comparing effect sizes between groups, or across studies, using p-values.

Focus on reporting both methodology (i.e. methodological decisions) and results [7] so that analyses are truly reproducible.

11

Exploratory modelling - Example

- Understand the relationship between a continuous outcome and ten continuous and binary variables in a large dataset
- Key tasks: select variables and select functional form
  - Use pre-defined choice of tuning parameters to control both the number of variables, and the complexity of the functional form (including interactions)
  - This can take the form of penalty functions (e.g. lasso); choice of number and position of knots in splines
  - In Multivariable Fractional Polynomial (MFP) modelling, p-values are the tuning parameters [8]

12

p-values as tuning parameters

- The MFP procedure uses two tuning-p-values:

    1. for variable selection: this controls how strong a variable's (potentially non-linear) effect needs to be for inclusion;

    2. To select functional form (functional selection procedure): are we interested to identify small deviations from linearity (large p-value) or would it be acceptable to oversee smaller deviations (small p-value) and specify a simpler model with a linear function.

For the multivariable analysis, MFP starts with the full model and combines backward elimination with the functional selection procedure for continuous variables.

13

Comments I:

- These p-values are not sacred ‘p-value thresholds’ but **model-complexity tuning parameters**.
- The selected model is not ‘the final truth’ but the primary selected MFP model.
- Sensitivity and stability analyses are important:
  - Sensitivity analysis: using different model-complexity p-values
  - Stability analyses: bootstrap procedures could be used
- It is important that all steps of this process are reported, ideally by using a structured reporting approach.

14

Comments II:

- Ideally scientific knowledge is used to pre-specify a model.
- However, in most analyses, data dependent decisions are still needed, although parts of the model may be ‘known’. For the ‘unknown’ part, tuning p-values play an important role for model building.
- It is often argued that p-values should be replaced by information criteria (eg AIC, BIC) but that does not solve the issue. For example, for a one degree of freedom term, 0.157 corresponds approximately to AIC. P-values allow suitable choices according to the main of an analysis.
- Emphasis needs to be on improved methodological reporting, rather than debates over the choice of p-values [e.g. 7].

15

Explanatory / Causal analysis

Useful to distinguish between
(a) identifying confounders and selecting their functional form versus
(b) testing for the causal effect of the exposure/intervention.

– For (a), since the aim is to balance the covariates in the dataset (e.g. using a propensity score), using p-values (viewed as a tuning parameter) for model selection is acceptable; pre-specification of the number of tests is not required.

– For (b), because the result may contribute to a decision-making process, akin to trials, pre-specification in the protocol is important, as is *a-priori* specification of the alternative hypothesis (minimally important clinical difference) and estimation of the power to detect this. Pilot studies may be helpful here [see, e.g., 9]

16

Explanatory / Causal analysis: pre-specify!

Some **RCTS** exploit p-values by combining co-or dual primary outcomes with **hierarchical testing**. They control the overall type I error by (re)assigning alpha portions over different (sub) hypotheses leading to boundaries for a network of p-values. Such **gatekeeping strategies**, yield additional evidence on a set of estimands while avoiding undue false positives

With **observed exposures** and focus on target trial emulation and **causal estimands, safe-guarding the causal null** is still critical, but now hinges on (untestable) causal assumptions, e.g. involving Instrumental Variables or no-unmeasured confounding.

To **defend conditional independence** in a DAG, or variable selection in adjusted models outcome or propensity model, simple p-values are less helpful. **Confounders** draw on their *joint* adjusted association with exposure & outcome.

p-values provide **an informal distance metric** when checking covariate balance and comparing their means among the treated and non-treated cohorts

Discussion

- Reproducibility is a key characteristic of good research. To support this we need
  - Increased focus on registration of studies
  - Publication of study protocols, including statistical analysis plans
  - Improved methodological reporting (as well as reporting of results)
- Within this context, point estimates, confidence intervals and p-values are all indispensable inferential tools : p-values are especially useful for joint tests
- P-values can also have a distinct use as tuning parameters in model building
  - Clear reporting is important to avoid confusion
- P-values are not reproducible: type-1 error and power are
- Template examples are a key tool to change practice
- Journals have a key role to play –
  papers with poorly reported methods should be rejected without review

18

Selected references

1. Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature 2019;567:305-307
2. Panikkar B, Lemmond B, Allen L, Dipirro C, Kasper S. Making the invisible visible: Results of a community-led health survey following PFAS contamination of drinking water in Merrimack, New Hampshire. Environmental Health: A Global Access Science Source 2019;18(1):79.
3. Marmor S, Longacre CF, Altman AM, Hui JYC, Jensen EH, Tuttle TM. Genomic expression assay testing among American Indian and Alaska Native women with breast cancer. Cancer 2020;126(24):5222-5229.
4. Carlin JB, Moreno-Betancur M. On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching. Stat Med. 2025 Jun;44(13-14):e10244. doi: 10.1002/sim.10244. PMID: 40553044; PMCID: PMC12186762.
5. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative.
6. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)statement: guidelines for reporting observational studies. Lancet. 2007 Oct 20;370(9596):1453-7. PMID: 18064739
7. Sauerbrei, W., Ambrogi, F., De Bin, R., Boulesteix, A. L., Goetghebeur, E., & Huebner, M. (2025). Commentary: Regression Models—Efforts Are Required to Improve Statistical Practice and Teaching. Statistics in Medicine, 44(13-14), e10341.
8. Sauerbrei, W., Royston, P., Kipruto, E. (2025). Multivariable Fractional Polynomial Models and Extensions. In: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-69359-9_397
9. Ahlem Fourati, Marion Vergonjeanne, Thibaut Gellé, Julien Magne, A practical guide to target trial emulation: connecting randomized trials and real-world data in cardiovascular research, European Journal of Preventive Cardiology, 2026;, zwag267, https://doi.org/10.1093/eurjpc/zwag267
10. Swanson S, Labrecque J, and Hernán M. Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable? European Journal Of Epidemiology 2018; 33(8): 723-728
11. Hansford HJ, Cashin AG, Jones MD, et al. Reporting of observational studies explicitly aiming to emulate randomized trials: a systematic review. JAMA Netw Open. 2023;6:e2336023. [PMID: 37755828] doi: 10.1001/jamanetworkopen.2023.36023

19