A Strong Case for Rethinking Causal Inference

JREE commentary: https://doi.org/10.1080/19345747.2023.2203683

John Deke

jdeke@mathematica-mpr.com

FCSM 2024

Two papers on inferential errors in education research

A Recipe for Disappointment: Policy, Effect Size, and the Winner's Curse

Simpson, A. (2022). https://doi.org/10.1080/19345747.2022.2066588

"Filtering a set on any quantity measured with error risks the 'winner's curse': conditional on selecting higher valued measures, the measurement likely overestimates the latent value."

Quantifying "Promising Trials Bias" in Randomized Controlled Trials in Education

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2022). https://doi.org/10.1080/19345747.2022.2090470

"…low powered trials tend to systematically exaggerate effect sizes among the subset of interventions that show promising results ($p<\alpha$). We conduct a retrospective design analysis to quantify this bias across 22 such promising trials, finding that the estimated effect sizes are exaggerated by an average of 52% or more."

Both are examples of Type M errors

Gelman, A., & Carlin, J. (2014). https://doi.org/10.1177/1745691614551642

Illustration of Type M error

- n.studies <- 1000
- truth <- rnorm(n.studies,mean=0,sd=0.2)
- se <- rep(sqrt(4/400),n.studies)
- est <- rnorm(n.studies,truth,se)
- tstat <- est/se
- stat.sig <- tstat > 1.96
- mean(truth[stat.sig])

  0.25
- mean(est[stat.sig])

  0.31

Mathematica

Four author-proposed solutions

Account for Type M errors in power analysis and conduct larger studies

Conduct more targeted studies

Use more non-experimental studies

Adjust impact estimates to account for ‘winner’s curse’

Illustration of bigger or more targeted studies

```
mean(truth[stat.sig])
    0.25
mean(est[stat.sig])
    0.31
```

```
mean(truth[stat.sig])
    0.21
mean(est[stat.sig])
    0.23
```

```
mean(truth[stat.sig])
    0.37
mean(est[stat.sig])
    0.40
```

Mathematica

Adjust for Winner's Curse

- truth <- rnorm(n.studies,mean=0,sd=0.2)
- se <- rep(sqrt(4/400),n.studies)
- est <- rnorm(n.studies,truth,se)
- tstat <- est/se
- est.sig <- 1.96*0.10
- wt.prior <- 1/(0.2^2)
- wt.data <- 1/(0.1^2)
- est.adj <- (wt.prior*0 + wt.data*est.sig)/(wt.prior+wt.data)

  0.16
- mean(truth[est > 0.19 & est < 0.21])

  0.16

A better solution:

Address the underlying problem

Underlying problem: confusion about statistical inference

Two types of statistical inference

First Type – how do the data look?

Confusingly/lamentably called "frequentist" inference

Inference about the probability distribution of estimates (aka data summaries)

Impact estimate, subgroup difference in sample means

Example statement: The impact estimate is unbiased and has a standard error of 3

Second Type – what do the data mean?

Confusingly/lamentably called "Bayesian" inference

Inference about the probability distribution of estimands (aka parameters)

True impact, subgroup difference in population means

Example statement: The probability of a favorable impact is 90 percent.

First type ≠ Second type

An estimate of a thing is not necessarily The Thing

The estimated difference is almost never equal to the true difference

An estimate is influenced by (at least) two factors

A genuine phenomenon (for example, a genuine difference between two groups)

Random errors (for example, random sampling error)

First type of inference is all about the random errors

For example, if we assume the true effect is zero, how likely is an impact estimate of the magnitude we observe or larger?

Second type also requires information about genuine phenomena

For example, how often do educational interventions increase test scores?

Widespread confusion: interpretation of probability

Probability is a mathematical construct

Similar to how "2+2 = 4" is a mathematical construct

To be useful, mathematical constructs need to be connected to the real world

Story problem: "Suppose you have two apples. Your friend gives you two more apples. How many apples do you have?"

So, we need a story to connect the mathematical construct of probability to the real world

Frequentist story – probability viewed as relative frequency

Bayesian story – probability viewed as intensity of personal belief

But – very confusingly! – you don't have to adopt the "Bayesian story" of probability to conduct "Bayesian inference"

We can be "Bayesian" (in terms of our inferential target) and "frequentist" (in terms of interpretation of probability) at the same time. See Bayesians are Frequentists

What to do instead

Recommendation 1: Instead of trying to cope with Type M errors, let's just stop making them!

Do not attempt to draw inferences about genuine phenomena using a framework focused on estimates of those phenomena

Recommendation 2: Use appropriate methods (Bayesian data analysis)

Setup a 'full probability model' of both parameters and data

Conditioning on observed data, calculate and interpret the posterior distribution of parameters

Assess sensitivity of conclusions to modeling assumptions

Recommendation 3: Continue to use non-Bayesian inference, but for the right reasons

Provide a transparent representation of data and raw materials for future meta-analysis

Recommendation 4: Interpret probability as a relative frequency, not personal belief

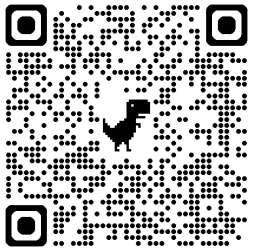
Final thoughts

BASIE (BAyeSian Interpretation of Estimates)

A framework I developed with Mariel Finucane that is aligned with my recommendations

https://www.acf.hhs.gov/sites/default/files/documents/opre/opre_brief_finucanedeke_042619_508_1.pdf

https://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE2022005



Recommendations apply to all statistical inference, not just causal inference

Other papers in our session provide examples

Is Bayes the only option?

People draw inferences about causal relationships all the time without Bayes

But – those aren't statistical inferences

Statistical inference about causal relationships (or model parameters more generally) is by definition Bayesian inference