Syllabus OIDD 937 – Methods Stumblers: Pragmatic Solutions to Everyday Challenges in Behavioral Research

Instructor: Uri Simonsohn

This version: 2021 07 20
[currently being revised for Fall 2021]

Tentative schedule for Fall 2021 (all classes 9AM-12PM)

1. Tuesday, August 31, 2021
2. Wednesday, September 1, 2021
3. Monday, September 20, 2021
4. Wednesday, September 22, 2021
5. Monday, October 11, 2021
6. Wednesday, October 13, 2021

Course overview:
This PhD-level course is for students who have already completed at least a year of basic stats/methods training. It assumes students already received a solid theoretical foundation and seeks to pragmatically bridge the gap between standard textbook coverage of methodological and statistical issues and the complexities of everyday behavioral science research. It focuses on issues that (i) behavioral researchers are likely to encounter as they conduct research, but (ii) may struggle to figure out independently by consulting a textbook or published article. Topics meet this second criterion for one of four reasons:

1. They are technically challenging (e.g., When, and what does it mean, to cluster standard errors? What if data severely violate a test's assumption? How to bootstrap.)

2. There isn't yet consensus among methodologists about them and hence a behavioral researcher will encounter different recommendations on how to proceed depending on the source that's consulted (e.g., Bayesian vs frequentist inference for lab experiments, analyzing replication results.)

3. There is high degree of consensus among methodologists, but the ideas have not yet become mainstream among behavioral researchers (e.g., stimulus sampling)

4. There is high degree of consensus among methodologists but such consensus may be premature and behavioral scientists may be better off not following it (e.g., how to correct for publication bias in meta-analysis, testing for moderation in experiments, power analysis). This set will be obviously controversial.

Topics
Topic 1. Simulations and home-made distributions
Topic 2. P-curve & a skeptical view of meta-analysis
Topic 3. Bayesian hypothesis testing
Topic 4. Fixing and breaking things with regression
Topic 5. What are data for?

Readings
Below you will see a list of readings for each week, accompanied by a letter that indicates how deeply you are encouraged to read it:

A: Read the entire paper
R: Read key results only
S: Skim/browse

Papers without a [A/R/S] are for further reading if you are interested.

Topic 1. Simulations & Home-made distributions
(Permutation tests, bootstrapping & Monte-Carlo)

(A) Torfs & Brauer (2014) A (very) short introduction to R*,Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands,*
[if you have not used R before, this is a quick intro to the basics]
[online course on R: https://psyteachr.github.io/msc-data-skills/]

(A) Simonsohn (2013) Just post it: the lesson from two cases of fabricated data detected by statisics alone*,Psychological Science,* 24(10), 1875-1888

(S) Joseph P. Simmons, Nelson, & Simonsohn (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*,Psychological Science,* 22(11), 1359-1366
for today's discussion, just focus on the simulations)

(S) Simonsohn, Simmons, & Nelson (2020) Specification curve analysis*,Nature Human Behaviour,* 4(11), 1208-1214
For our purposes, the key issue is how to test the null that there is no effect across all specifications.

Other readings

Young (2019) Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results*,The Quarterly Journal of Economics,* 134(2), 557-598

Reruns tests in experimental econ using simulation and a share of results are no longer p<.05

Boos (2003) Introduction to the bootstrap world*,Statistical science,* 18(2), 168-174

Good 'introductory' reference, though I think it is best used once you already know a bit.

Boos & Brownie (1989) Bootstrap methods for testing homogeneity of variances*,Technometrics,* 69-82

You are unlikely to ever use this test, but going thought the logic of how it is designed it is useful for understanding just how many assumptions go and don't go into any bootstrap-based test in particular, and any hypothesis test more generally. In a nutshell, research questions are about general similarity of groups, but our tests are about one specific metric only (e.g., variance or mean).

Pitman (1937) Significance tests which may be applied to samples from any populations*,Journal of the Royal Statistical Society,* 4(1), 119-130

Seems to have independently invented permutation tests. Fun to browse for historical reasons.

Micceri (1989) The unicorn, the normal curve, and other improbable creatures*,Psychological bulletin,* 105(1), 156

Hey, distributions in real life are not normal(!)

Sawilowsky & Blair (1992) A more realistic look at the robustness and Type II error properties of the t-test to departures from population normality*,Psychological bulletin,* 111(2), 352

White (2003) A reality check for data snooping*,Econometrica,* 68(5), 1097-1126

What if you select post-hoc the best performing mutual fund, how do you correct for that?

Daniël Lakens (2015) The 20% Statistician - Always use Welch's t-test instead of Student's t-test

I am not persuaded, but a good exercise in simulating to answer a practical question.

Some recent attempts to cover some of the topics we will discuss

Bind & Rubin (2020) When possible, report a Fisher-exact P value and display its underlying null randomization distribution*,Proceedings of the National Academy of Sciences,* 117(32), 19151-19158

Morris, White, & Crowther (2019) Using simulation studies to evaluate statistical methods*,Statistics in medicine,* 38(11), 2074-2102

Rousselet, Pernet, & Wilcox (2021) The percentile bootstrap: a primer with step-by-step instructions in R*,Advances in Methods and Practices in Psychological Science,* 4(1), 2515245920911881

Topic 2. *P*-curve and a skeptical take on meta-analysis

(A) Simonsohn, Nelson, & Simmons (2014) *p*-curve: A Key to the File Drawer*,Journal of Experimental Psychology: General,* 143(2), 534-547
(only read *p*-curve for this topic)

(A) Vosgerau, Simonsohn, Nelson, & Simmons (2019) 99% impossible: A valid, or falsifiable, internal meta-analysis*,Journal of Experimental Psychology: General,* 148(9), 1628

Simonsohn (2015) Data Colada [41] - Falsely Reassuring: Analyses of ALL p-values
So ignore such analyses.

McShane, Böckenholt, & Hansen (2016) Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes*,Perspectives on Psychological Science,* 11(5), 730-749
They really *hate* *p*-curve

DataColada[24] *p*-curve vs excessive significance testing http://datacolada.org/24

DataColada[30] Trim-and-Fill is Fully of It (Bias) http://datacolada.org/30

DataColada[58] The Funnel Plot is Invalid Because of This Crazy Assumption: r(n,d)=0 http://datacolada.org/58

Lane & Dunlap (1978) Estimating effect size: Bias resulting from the significance criterion in editorial decisions*,British Journal of Mathematical and Statistical Psychology,* 31(107-112
Published papers over-estimate effect size. Many papers have made this same point later on, without properly citing it.

J. P. A. Ioannidis (2005) Why most published research findings are false,*Plos Medicine,* 2(8), 696-701
Classic paper. Bayesian calibration for p(H|D)

Pashler & Harris (2012) Is the Replicability Crisis Overblown? Three Arguments Examined,*Perspectives on Psychological Science,* 7(6), 531-536
No it is not overblown.

Francis (2013) Replication, statistical consistency, and publication bias,*Journal of Mathematical Psychology,*
It's surprising for all studies to be p<.05 (if we ignore that p>.05 are not published).

Simonsohn (2013) It Really Just Does Not Follow, Coments on Francis (2013),*Journal of Mathematical Psychology,*
Publication bias does not imply we should ignore published evidence.

Stanley & Doucouliagos (2014) Meta-regression approximations to reduce publication selection bias,*Research Synthesis Methods,* 5(1), 60-78
Wrong.
But good to know it exists.

David Card & Krueger (1995) Time-series minimum-wage studies: a meta-analysis,*The American Economic Review,* 85(2), 238-243
Seems like minimum-wage mattering is publication bias at work.

Leamer (1983) Let's take the con out of econometrics,*The American Economic Review,* 31-43
Selective reporting in economics, and an extreme solution

Sala-i-Martin (1997) I just ran two million regressions,*The American Economic Review,* 178-183
Tries to make Leamer 1983 less extreme.

Duval & Tweedie (2000) Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis,*Biometrics,* 56(2), 455-463
The most famous correction for publication bias.
It does not work. See www.datacolada.org/30

Rothstein, Sutton, & Borenstein (2005) Publication Bias in Meta-Analysis
If you want to know what else has been attempted.

Hedges & Vevea (1996) Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model*,Journal of Educational and Behavioral Statistics,* 21(4), 299-332

Makes Lane and Dunlap point, and comes one assumption short of developing *p*-curve 20 years earlier.

Sharpe (1997) Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away*,Clinical psychology review,* 17(8), 881-901

Not a fan of meta-analysis

Borenstein (2009) Criticisms of Meta-Analysis*,Introduction to Meta-Analysis,*

Some answers to concerns like those of Sharpe.

Wallis (1942) Compounding probabilities from independent significance tests*,Econometrica,* 229-248

This paper changed how I think of *p*-values.

Has wonderful metaphors to think about abstract ideas.

J. Ioannidis & Trikalinos (2007) An exploratory test for an excess of significant findings*,Clinical Trials,* 4(3), 245

Neat idea. Massively misused in follow up work.

In a way, a discrete version of *p*-curve that includes p>.05 results.

Topic 3 Bayesian hypothesis testing

Dienes (2019) How do I know what my theory predicts? ,*Advances in Methods and Practices in Psychological Science,* 2(4), 364-377

(A) Dienes (2011) Bayesian versus orthodox statistics: Which side are you on? ,*Perspectives on Psychological Science,* 6(3), 274-290

Rouder, Speckman, Sun, Morey, & Iverson (2009) Bayesian t tests for accepting and rejecting the null hypothesis,*Psychonomic Bulletin & Review,* 16(2), 225-237

Wagenmakers (2007) A practical solution to the pervasive problems ofp values,*Psychonomic bulletin & review,* 14(5), 779-804

Simonsohn (2014) Data Colada [13] - Posterior-Hacking

Simonsohn (2015) DataColada[35]: The Default Bayesian Test is Prejudiced Against Small Effects

Rouanet (1996) Bayesian methods for assessing importance of effects,*Psychological bulletin,* 119(1), 149-158

Sanborn & Hills (2014) The frequentist implications of optional stopping on Bayesian hypothesis tests,*Psychonomic bulletin & review,* 21(2), 283-300

Rouder (2014) Optional stopping: No problem for Bayesians,*Psychonomic bulletin & review,* 21(2), 301-308

Sanborn et al. (2014) Reply to Rouder (2014): Good frequentist properties raise confidence,*Psychonomic bulletin & review,* 21(2), 309-311

Topic 4. What regression can fix and break

Simonsohn (2016) Two Lines: The First Valid Test of U-Shaped Relationships

Median splits

Gelman & Park (2008) Splitting a predictor at the upper quarter or third and the lower quarter or third,*The American Statistician,* 62(4), 1-8

McClelland, Lynch, Irwin, Spiller, & Fitzsimons (2015) Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power,*Journal of Consumer Psychology,* 25(4), 679-689

Rucker, McShane, & Preacher (2015) A researcher's guide to regression, discretization, and median splits of continuous variables*,Journal of Consumer Psychology,* 25(4), 666-678

Iacobucci, Posavac, Kardes, Schneider, & Popovich (2015) Toward a More Nuanced Understanding of the Statistical Properties of a Median Split*,Journal of Consumer Psychology,* 25(4), 652-665

Interactions in non-linear models

Ai & Norton (2003) Interaction terms in logit and probit models*,Economics letters,* 80(1), 123-129

Karaca-Mandic, Norton, & Dowd (2012) Interaction Terms in Nonlinear Models*,Health Services Research,* 47(1pt1), 255-274

Greene (2010) Testing hypotheses about interaction terms in nonlinear models*,Economics Letters,* 107(2), 291-296

Controlling for things

Bhargava, Kassam, & Loewenstein (2014) A reassessment of the defense of parenthood*,Psychological science,* 25(1), 299-302

D. Card & Dahl (2011) Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior**,Quarterly Journal of Economics,* 126(1), 103

Spotlight and floodlight

Spiller, Fitzsimons, Lynch Jr, & McClelland (2013) Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression*,Journal of Marketing Research,* 50(2), 277-288

Westfall & Yarkoni (2016) Statistically Controlling for Confounding Constructs Is Harder than You Think*,PLOS ONE,* 11(3), e0152719

Preacher, Curran, & Bauer (2006) Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis*,Journal of educational and behavioral statistics,* 31(4), 437-448

Other

Lee & Lemieux (2010) Regression Discontinuity Designs in Economics*,Journal of Economic Literature,* 48(281-355

Gelman & Stern (2006) The difference between "significant" and "not significant" is not itself statistically significant*,The American Statistician,* 60(4), 328-331

Gelman & Imbens (2014) Why high-order polynomials should not be used in regression discontinuity designs

An interesting debate

A. J. Healy, Malhotra, Mo, & Laitin (2010) Irrelevant events affect voters' evaluations of government performance*,Proceedings of the National Academy of Sciences of the United States of America,* 107(29), 12804-12809

Fowler & Montagnes (2015) College football, elections, and false-positive results in observational research*,Proceedings of the National Academy of Sciences,* 112(45), 13800

A. Healy, Malhotra, & Mo (2015) Determining false-positives requires considering the totality of evidence*,Proceedings of the National Academy of Sciences,* 112(48), E6591

Topic 5 – What are data for?

Alogna et al. (2014) Registered Replication Reports: Schooler and Engstler-Schooler (1990)*,Perspectives on Psychological Science,* 9(5), 556-578

Classic study replicates, kind of, very narrowly. What does that mean?

Deaton (2010) Instruments, randomization, and learning about development*,Journal of economic literature,* 424-455

Imbens (2010) Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)*,Journal of Economic Literature,* 48(2), 399-423

Possibly my favorite title.

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

Classic critique of the use of *p*-values. Does not focus on the most important things. But a classic.

Bond et al. (2012) A 61-million-person experiment in social influence and political mobilization*,Nature,* 489(7415), 295-298

Large N study, referred in our slides couple of times.

McCloskey & Ziliak (1996) The standard error of regressions*,Journal of Economic Literature,* 97-114

Discusses practical vs statistical significance and reviews AER papers to make the point economists do not sufficiently focus on the former. 20 years later things are much better. To some extent this is the "New Statistics" for econ.

Mahoney (1979) Review Paper: Psychology of the Scientist: An Evaluative Review,*Social studies of Science,* 9(3), 349-375

In the 70s they already realized researchers are anything but neutral observers. Cool read.

Daniel Lakens (2016) So you banned p-values

Response to journal that banned *p*-values

Faul, Erdfelder, Lang, & Buchner (2007) G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,*Behavior research methods,* 39(2), 175-191

Effectively a manual for GPower. Good luck.

Cohen (1992) A power primer,*Psychological bulletin,* 112(1), 155-159

Some basic stuff and procedures. Very mechanistic.

Joseph P Simmons, Nelson, & Simonsohn (2013) Life after p-hacking,*Meeting of the Society for Personality and Social Psychology, New Orleans, LA,* 17-19

You need larger sample sizes than you think to detect obvious effects.

Button et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience,*Nature Reviews Neuroscience,*

Many neuroscience studies with small samples.

Cumming & Finch (2001) A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions,*Educational and Psychological Measurement,* 61(4), 532-574

Long article. It explains something that is very important, that courses in stats do not seem to cover, and that unfortunately, I do not know a better source for. The topic is the relationship between confidence intervals and hypothesis testing, especially applied to standardized effect size, and explains how noncentral distributions are related to both (non-central is how the test-statistic is distributed when the null is false).

Sedlmeier & Gigerenzer (1989) Do studies of statistical power have an effect on the power of studies? ,*Psychological Bulletin; Psychological Bulletin,* 105(2), 309

No. And yet, we keep doing them

Cohen (1962) The statistical power of abnormal-social psychological research: A review,*Journal of Abnormal and Social Psychology,* 65(3), 145-153

A older assessment of power in the published literature.

Mayo & Spanos (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction*, The British Journal for the Philosophy of Science,* 57(2), 323-357

Important notion of severity. Wished a version not written for philosophers existed.

Cumming (2014) The New Statistics Why and How*, Psychological Science,* 25(1), 7-29

p-values=bad. Confidence intervals=good.