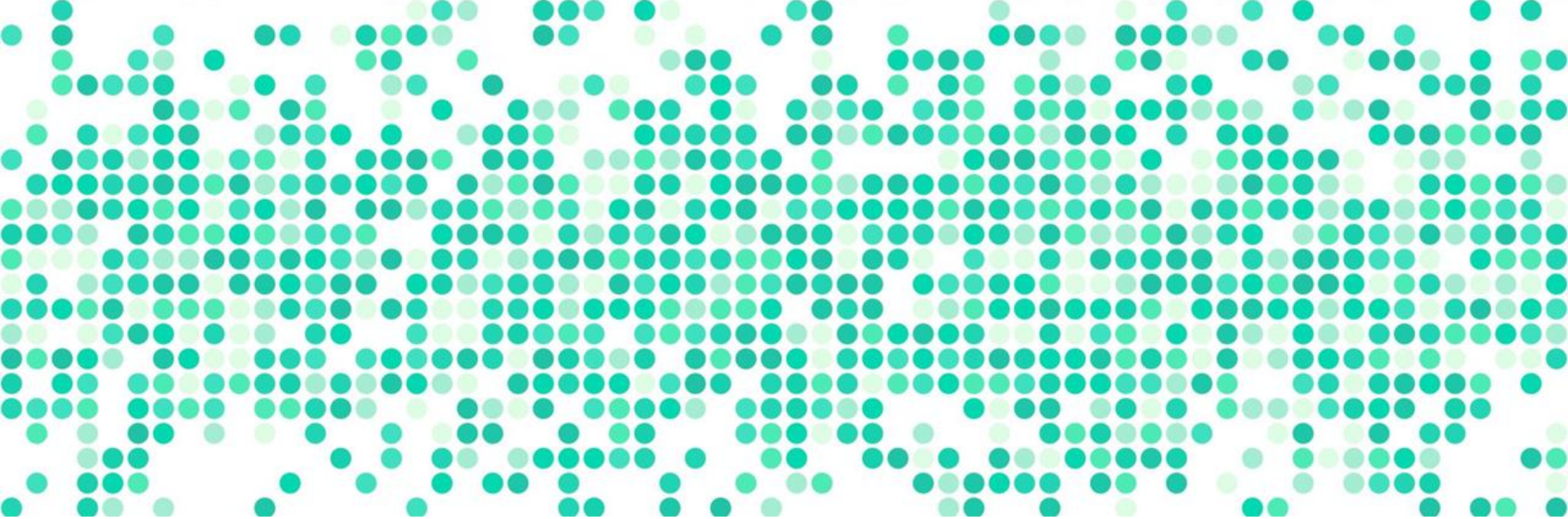


DATA ANALYSIS

Week 15: Bootstrapping + Review

logistics: points

- * indicates that scores could change based on upcoming assignments
- what's remaining
  - Week 15 quiz (1)
  - PS7 (5: maxes out at 30)
  - practice final (2)
  - final (20)
  - extra credit:
    - Surveys (0.5)
    - Qs +czar (1+1)
    - Analysis Ace (1)
    - Analyze a Research Paper (2)

*percentage based on 78 points (excluding extra credit)

View Rubric

Current Percentage Rubric

Criteria	Ratings	
Quizzes view longer description	*up to week 14, scaled to 10	/ 10 pts
Problem Sets view longer description	*up to PS6, total maxes out at 30	/ 30 pts
Midterm 1 view longer description	40% conceptual, 60% computational + bonus	/ 15 pts
Class Activities view longer description	2.5 for 90%, 2 for 80%, etc.	/ 2.5 pts
Video participation view longer description	2.5 for 80% videos watched	/ 2.5 pts
Practice Midterm 1 view longer description	1.5 for at least 50% score	/ 1.5 pts
Midterm 2	40% conceptual, 60% computational + bonus	/ 15 pts
Extra Credit Questions view longer description	*you'll have a comment here if you are close (8 weeks)	/ 1 pts
Practice Midterm 2 view longer description	1.5 for at least 50% score	/ 1.5 pts
Conceptual Czar view longer description	*1 if your question was on the conceptual exam	/ 1 pts
Analysis Ace view longer description	*highest computational score overall (TBD)	-- / 1 pts
Surveys view longer description	*0.5 (final, TBD)	/ 2 pts
Practice Final view longer description	2 for at least 50% score	/ 2 pts
Final Exam	* TBD: worth 25 if you opted out	-- / 20 pts
Extra: Analyze a Research Paper view longer description	* TBD	/ 2 pts

*your total accumulated points thus far

Total Points:

logistics: deadlines

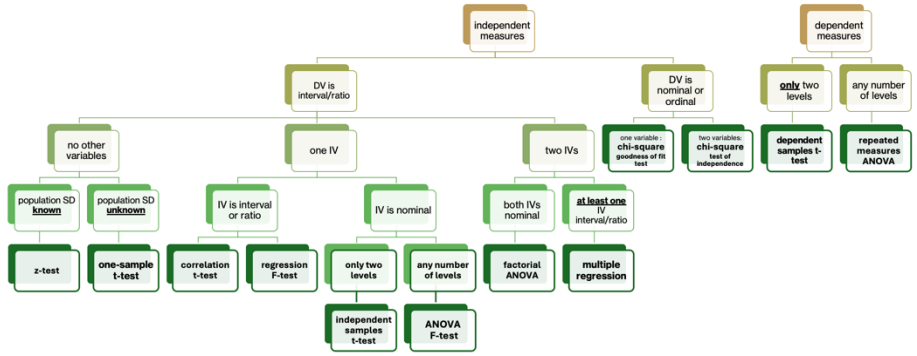
- no late submissions will be accepted or graded past 1 pm May 15
- this applies to ALL submissions (pending assignments, extra credit, late stuff, etc.)
- office hours
  - Prof. Kumar: calendly appointments
  - LAs: review session TBD

15	M: April 28, 2025	**PS5+ PS6 revision due**
15	T: April 29, 2025	W15: Miscellaneous Data
15	Th: May 1, 2025	W15 continued…
16	M: May 5, 2025	**PS7 due**
16	T: May 6, 2025	W16: Last Class / Final Exam review
17	Th: May 15, 2025	**PS7 revision + Computational Exam Due by 1.30 pm**
17	Th: May 15, 2025	**Conceptual Exam (1.30-3 pm, VAC South)**

logistics: final

- take-home computational: 60% of final exam points
  - **due at 1 pm May 15 – NO LATE SUBMISSIONS**
  - open book but **NOT open person**
- in-class conceptual (on Canvas): 40% of final exam points
  - **1.30 – 3 pm May 15 – VAC South (here)**
  - **closed** book (do NOT leave Canvas page once you begin)
- you can **bring**:
  - ONE handwritten help sheet
  - hypothesis flowchart
  - process sheet (packet)

hypothesis testing flowchart



Hypothesis Test Process

One interval/ratio variable

Ratio/interval level measurements
Independent observations

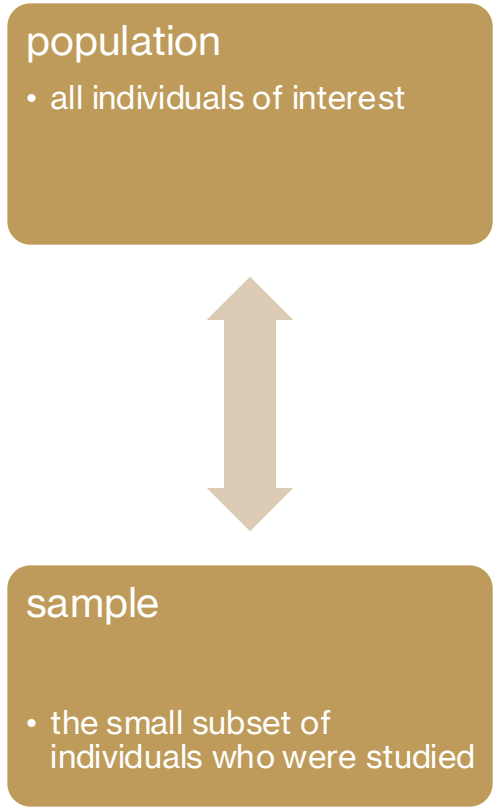
test	degrees of freedom	process
z-test population standard deviation **known**		• **Step 1: state the hypotheses:** o H_0: o H_1: • **Step 2: set criteria** o standard error = o critical value = • **Step 3: collect data** o test statistic:

today's agenda

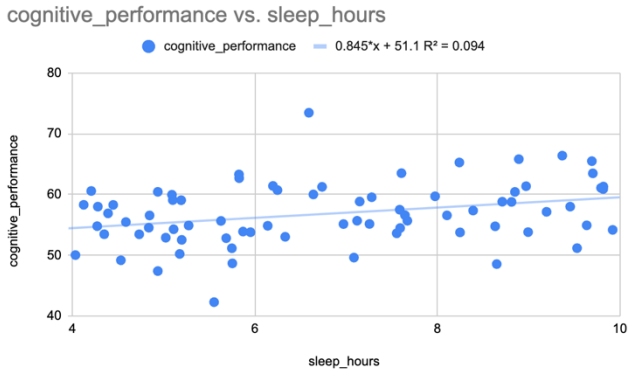
- revisiting sampling
- replication: crisis and safeguards
- reviewing statistical framework

revisit: from samples to populations

- we collect some data and obtain a sample statistic
- we want to know whether this sample statistic is close or far from our population parameter



revisit: sleep & performance



- recall that we wanted to explore the relationship between sleep and cognitive performance
- we began by asking: what if there was no true relationship between sleep and cognitive performance in the population?
- to create this “no relationship” null hypothesis, we shuffled the cognitive_performance column

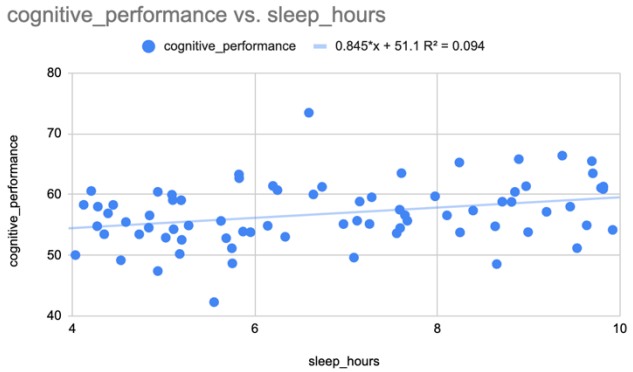
original

	sleep_hours	cognitive_performance
0	6.247241	60.718902
1	9.704286	63.479275
2	8.391964	57.356134
3	7.591951	54.474564
4	4.936112	47.395345

shuffled

	sleep_hours	cognitive_performance
0	6.247241	47.395345
1	9.704286	54.845377
2	8.391964	58.297484
3	7.591951	60.718902
4	4.936112	60.422949

revisit: sleep & performance

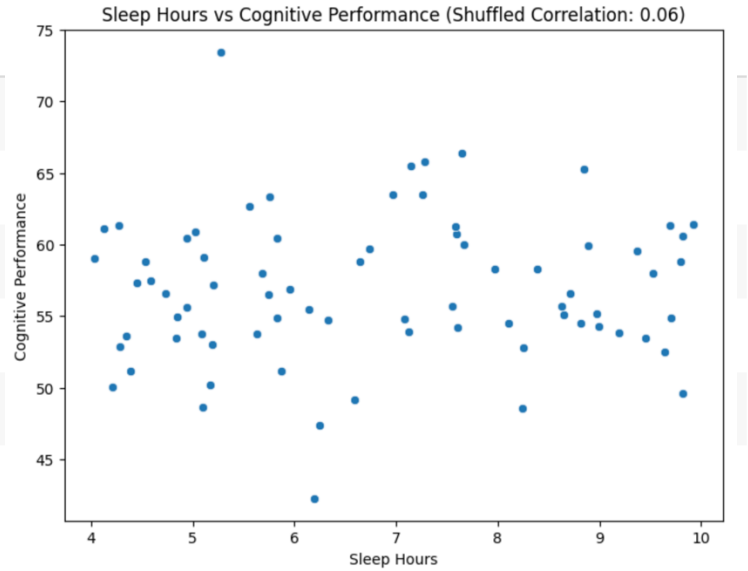


- recall that we wanted to explore the relationship between sleep and cognitive performance
- we began by asking: what if there was no true relationship between sleep and cognitive performance in the population?
- to create this “no relationship” null hypothesis, we shuffled the cognitive_performance column

original

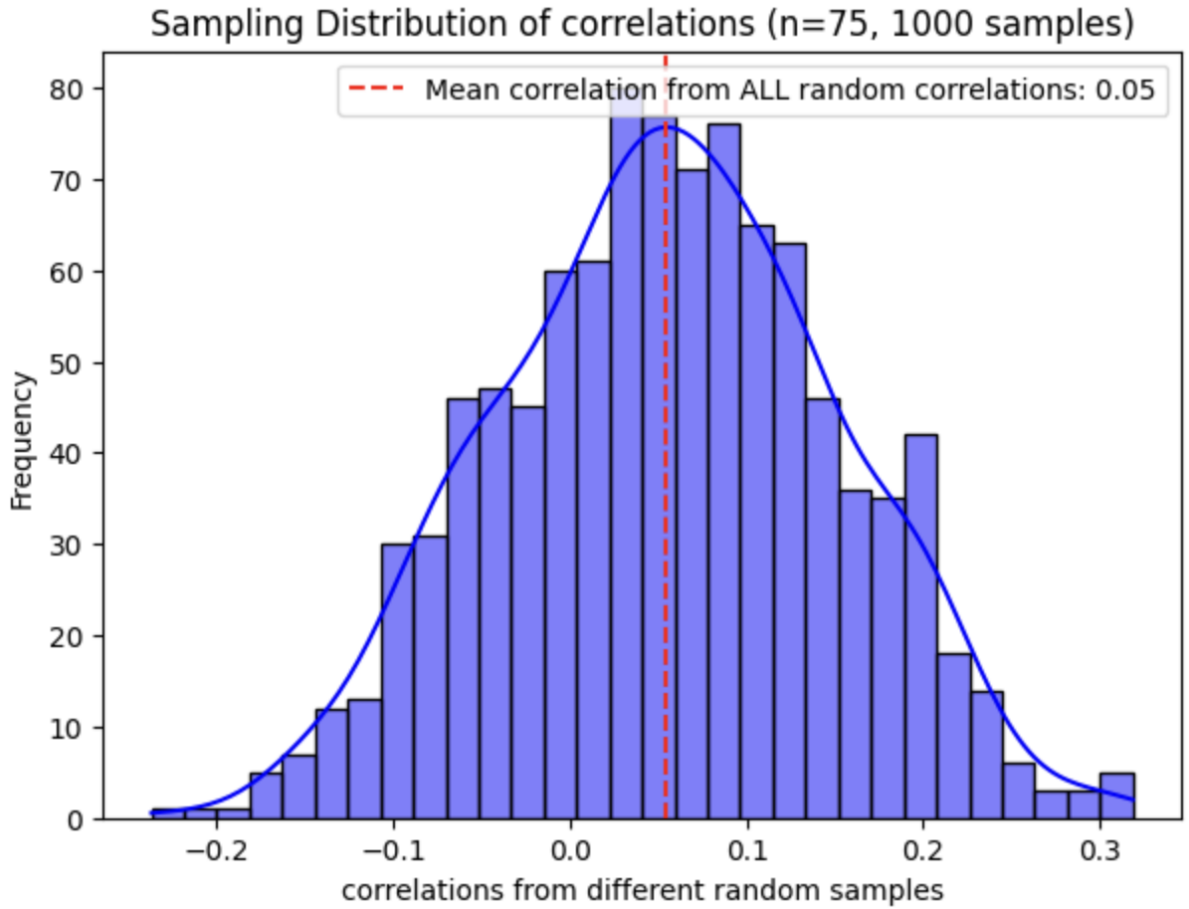
	sleep_hours	cognitive_performance
0	6.247241	60.718902
1	9.704286	63.479275
2	8.391964	57.356134
3	7.591951	54.474564
4	4.936112	47.395345

shuffled



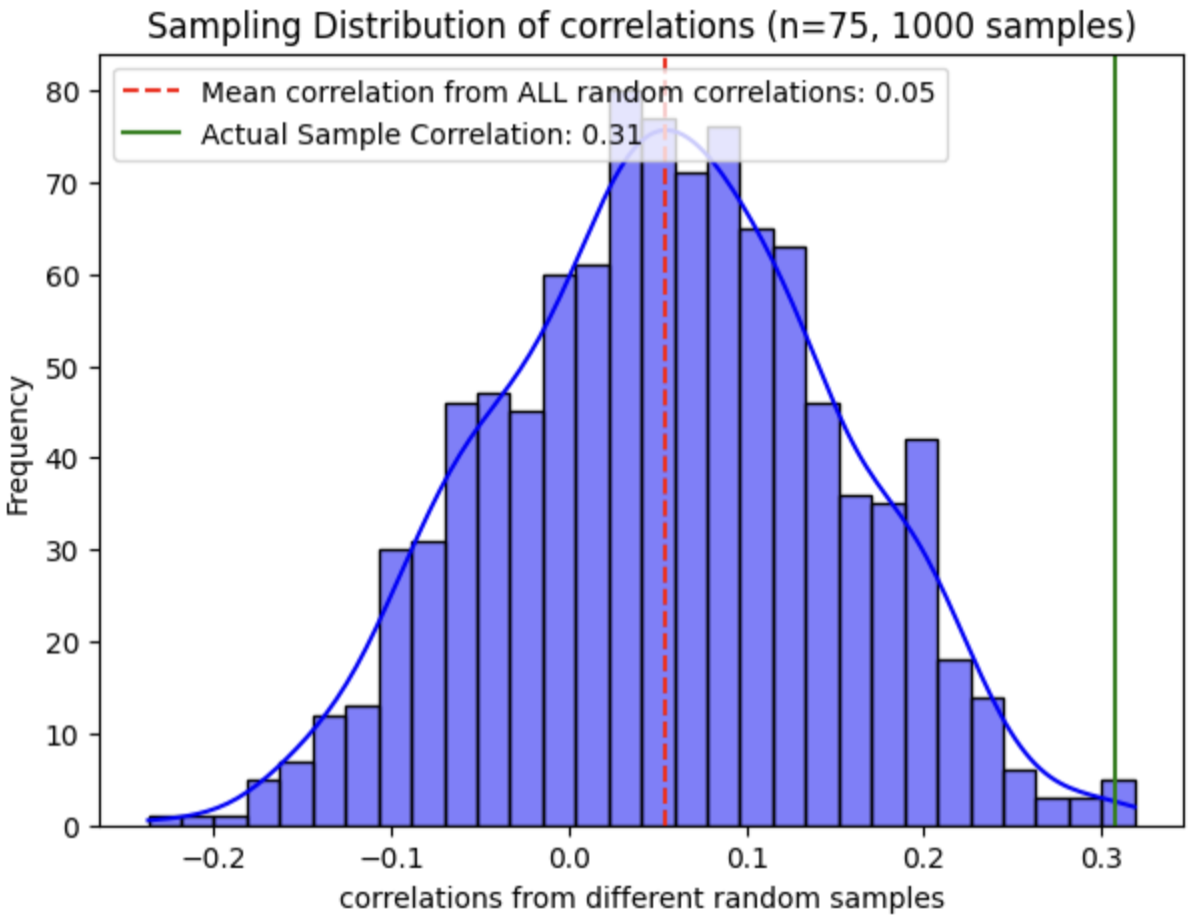
sampling distribution

- we took 1000 random samples with replacement from this null hypothesis distribution and calculated a correlation within each random sample
- what does the distribution of correlations look like for MANY such random samples?
- sampling distribution: distribution of all possible values of the **sample statistic** obtained from multiple samples of a given size



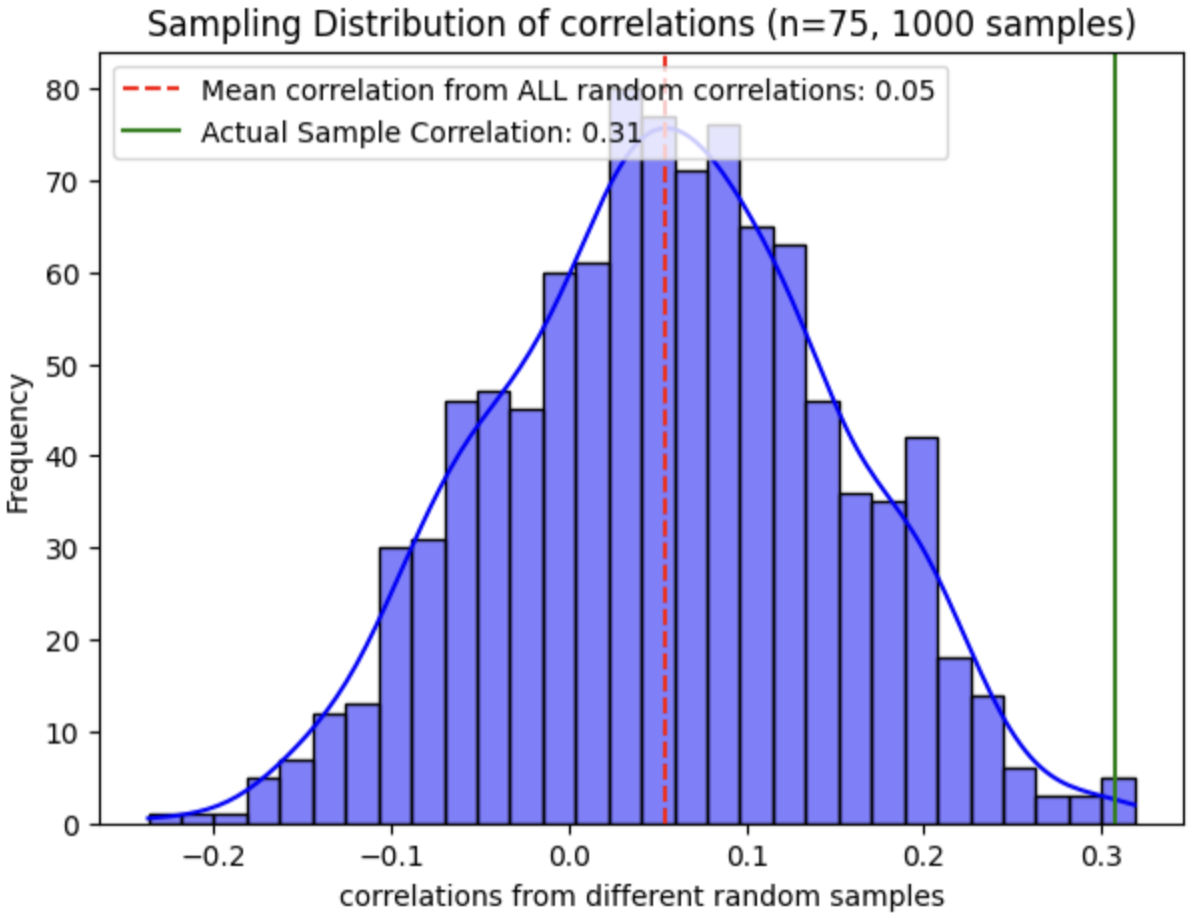
sampling distribution

- next, we compared this sampling distribution of random slopes from the shuffled dataset to the correlation in the actual sample
- we asked: **if there was no relationship between sleep and cognitive performance**: is obtaining a sample correlation of 0.31 typical?
- at this point, we shifted to assuming that the sampling distribution of correlations was t-distributed and proceeded with computing probabilities under that t-distribution



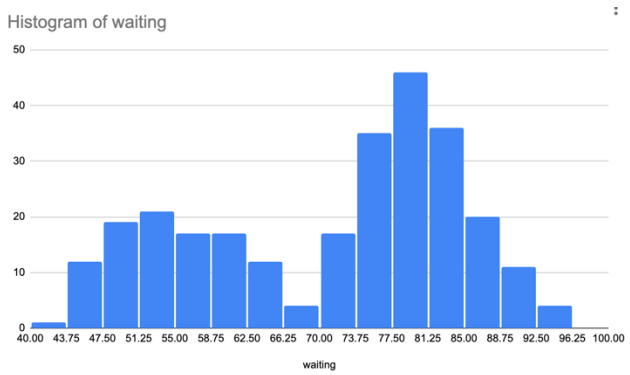
a bootstrapped distribution

- but, we could have simply looked at the probability of obtaining a correlation as extreme as ours under the hypothetical distribution we just generated
- without assuming an underlying distribution, you could still obtain a p-value
- this relaxes the many assumptions you make while conducting standard hypothesis tests
- in this case, p (*r* >= .31 | bootstrap) = .003



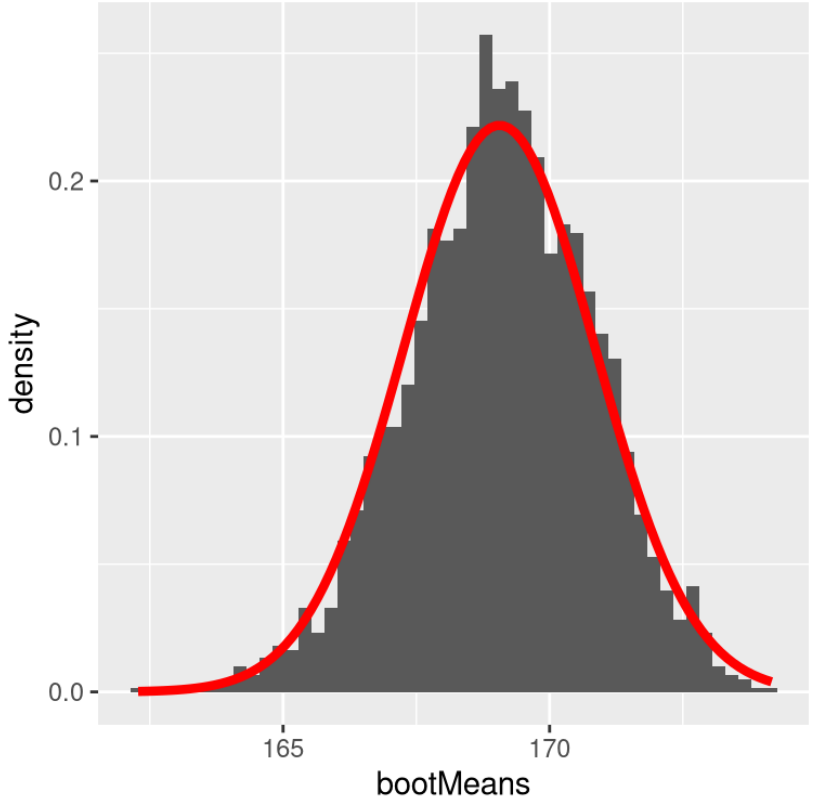
revisit: geyser eruptions dataset

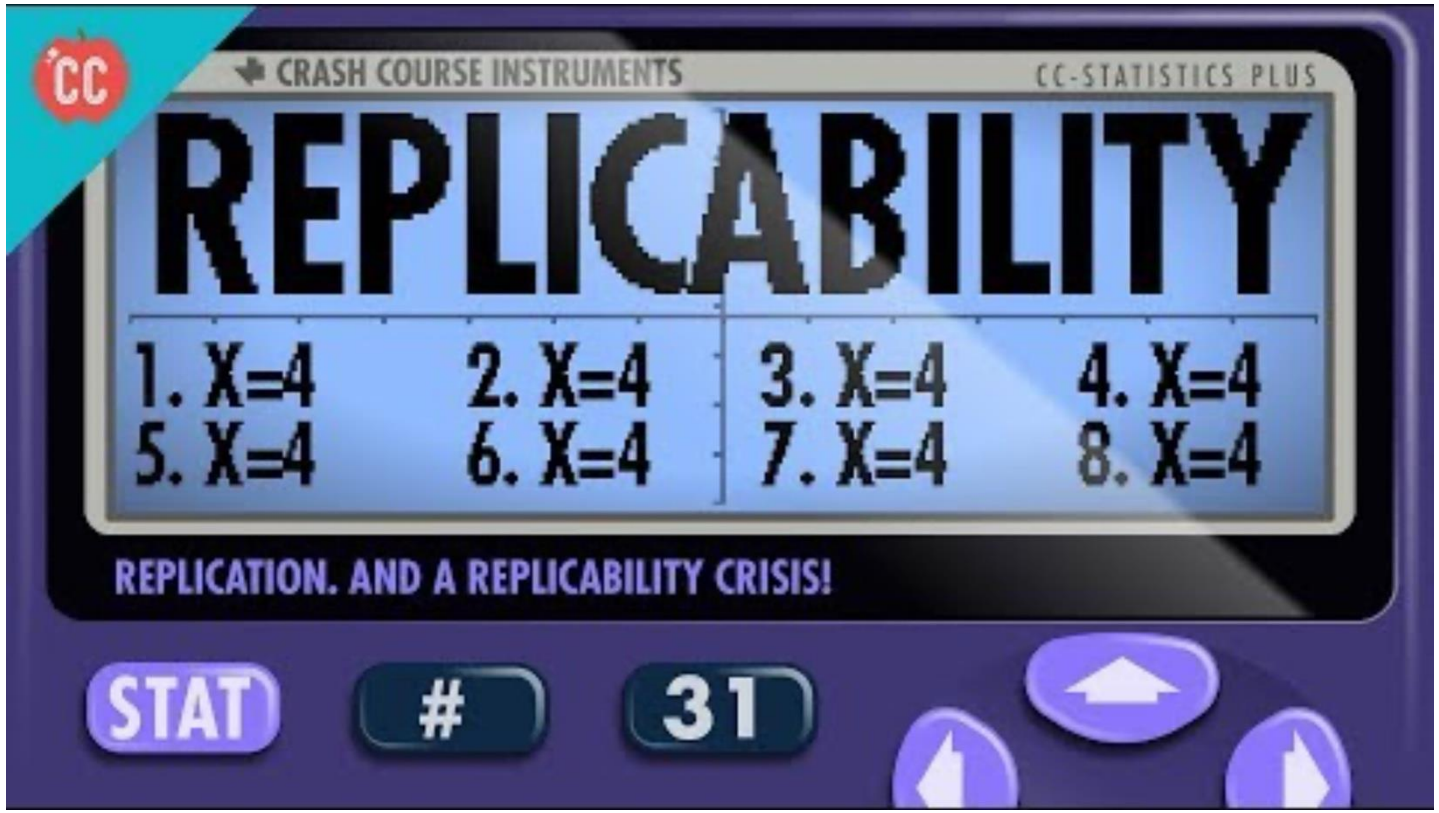
- Old Faithful geyser in Yellowstone National Park
- waiting time has a bimodal distribution
- to build any kind of model of waiting time, we would need to assume normality
  - waiting time ~ #earthquakes + pressure
- solution:
  - bootstrap! create a hypothetical “null” distribution with large N and proceed!



summary: bootstrapping

- when assumptions are violated, you can make adjustments OR consider an alternate strategy
- bootstrapping is a technique that allows you to generate several samples with replacement from the actual dataset to map out a hypothetical sampling distribution
- hypothesis testing can then be used by examining the probability of the data under this bootstrapped distribution: permutation test



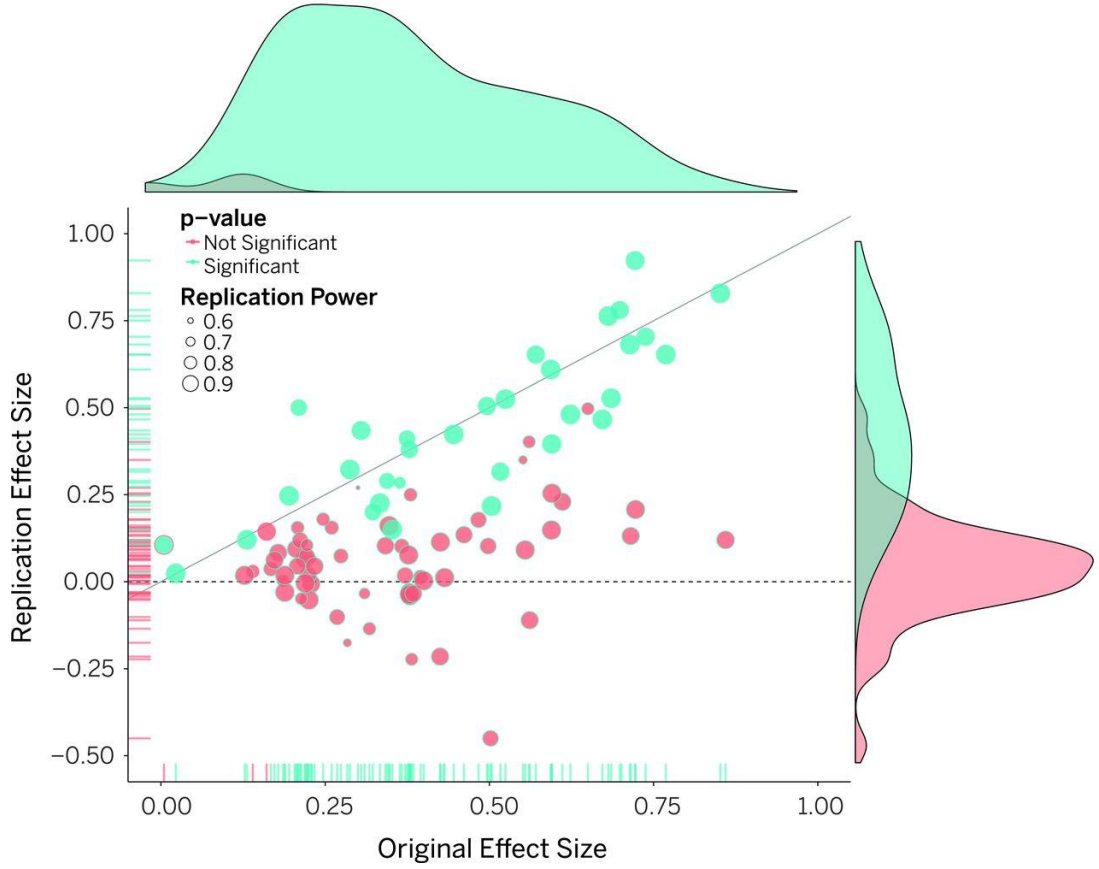

CC
CRASH COURSE INSTRUMENTS
CC-STATISTICS PLUS
REPLICABILITY
1. X=4
2. X=4
3. X=4
4. X=4
5. X=4
6. X=4
7. X=4
8. X=4
REPLICATION. AND A REPLICABILITY CRISIS!
STAT
#
31

video link

replication crisis

RESULTS

We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.



https://www.science.org/doi/10.1126/science.aac4716

replication crisis: why?

- type I error
- p-hacking / fraud
- violated assumptions
- overfitting

[109] Data Falsificada (Part 1): "Clusterfake"

Posted on June 17, 2023 by Uri, Joe, & Leif

This is the introduction to a four-part series of posts detailing evidence of fraud in four academic papers co-authored by Harvard Business School Professor Francesca Gino.

In 2021, we and a team of anonymous researchers examined a number of studies co-authored by Gino, because we had concerns that they contained fraudulent data. We discovered evidence of fraud in papers spanning over a decade, including papers published quite recently (in 2020).

Support Data Colada's Legal Defense

Simine Vazire is organising this fundraiser.

Created 1 day ago • Other

Data Colada Are Being Sued for Raising Scientific Concerns about Published Research: Support Their Legal Defense

replication crisis

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < 0.05$" (3 original nulls excluded; $n = 97$ studies); "mean original and replication effect sizes" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); "percent meta-analytic ($P < 0.05$)" (25 excluded; $n = 75$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

			Effect size comparison					Original and replication combined			
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Median original *df/N*	Mean (SD) replication effect size	Median replication *df/N*	Average replication power	Meta-analytic mean (SD) estimate	Percent meta-analytic ($P < 0.05$)	Percent original effect size within replication 95% CI	Percent subjective "yes" to "Did it replicate?"
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92	0.309 (0.223)	68	47	39
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91	0.138 (0.087)	43	34	25
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93	0.393 (0.209)	86	62	54
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92	0.286 (0.228)	58	40	32
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94	0.464 (0.221)	92	60	53

	Replications $P < 0.05$ in original direction
Original study characteristics	
Original *P* value	−0.327
Original effect size	0.304
Original *df/N*	−0.150
Importance of original result	−0.105
Surprising original result	−0.244
Experience and expertise of original team	−0.072

https://www.science.org/doi/10.1126/science.aac4716

correlates of replicability



Eleven years of student replication projects provide evidence on the correlates of replicability in psychology

Veronica Boyce[1,*], Maya Mathur[1], Michael C. Frank[1]

[1]Stanford University

Abstract

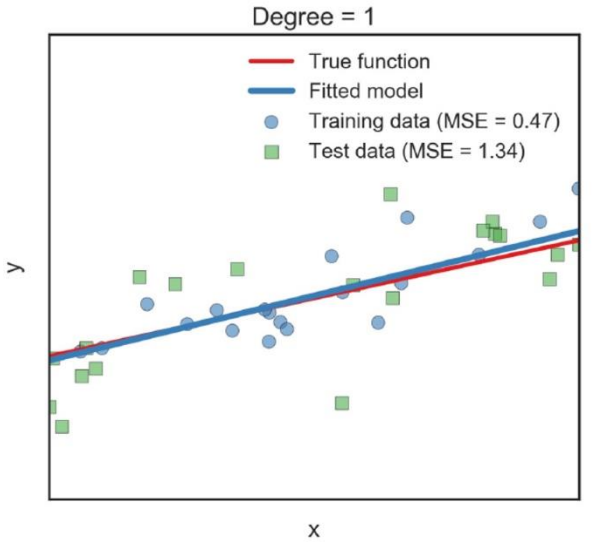
Cumulative scientific progress requires empirical results that are robust enough to support theory construction and extension. Yet in psychology, some prominent findings have failed to replicate, and large-scale studies suggest replicability issues are widespread. The identification of predictors of replication success is limited by the difficulty of conducting large samples of independent replication experiments, however: most investigations re-analyse the same set of ~170 replications. We introduce a new dataset of 176 replications from students in a graduate-level methods course. Replication results were judged to be successful in 49% of replications; of the 136 where effect sizes could be numerically compared, 46% had point estimates within the prediction interval of the original outcome (versus the expected 95%). Larger original effect sizes and within-participants designs were especially related to replication success. Our results indicate that, consistent with prior reports, the robustness of the psychology literature is low enough to limit cumulative progress by student investigators.

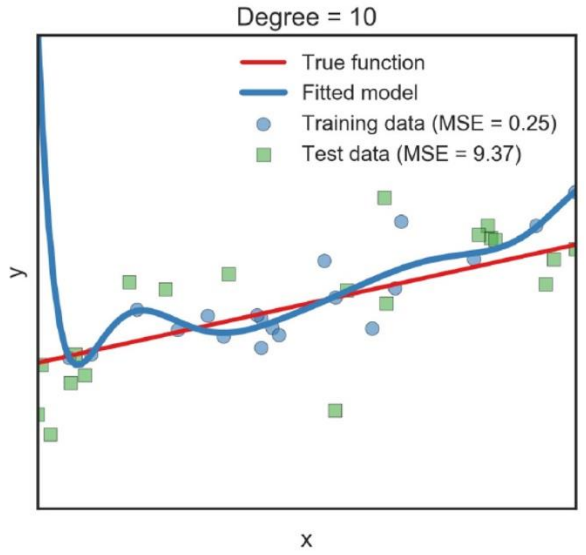
Table 1: The unadjusted Pearson correlations between each individual predictor and the subjective replication score. See Methods for how these variables were coded.

r	p	Predictors
0.333	0.000	Within participants design (versus between participants)
0.182	0.015	Log number of trials
0.150	0.047	Open data
0.080	0.294	Non psychology (versus cognitive psych)
0.075	0.322	Other psychology (versus cognitive psych)
0.064	0.399	Publication year
0.002	0.979	Open materials
-0.027	0.725	Stanford affiliation of original authors at time of replication
-0.047	0.536	Log ratio between replication and original sample sizes
-0.108	0.155	Log original sample size
-0.158	0.037	Switch to online for replication (versus same modality for original and replication)
-0.246	0.001	Social psychology (versus cognitive psych)
-0.267	0.000	Single vignette (versus multiple items/inductions per condition)

assessing model fit

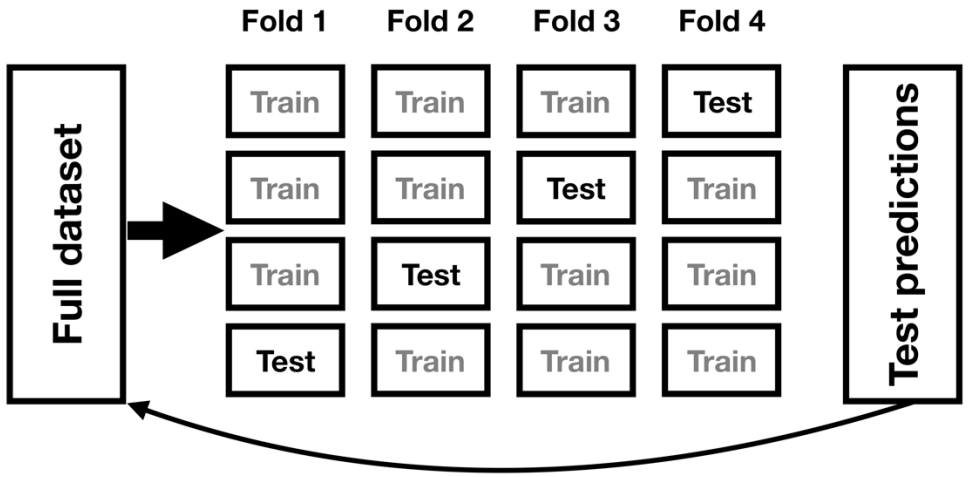
- if our goal is to reduce error, then we should be fitting a model with lots of parameters and variables
- BUT when our model fits the data too well, it can lead to overfitting
- “It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.” – Albert Einstein (1933)





cross validation

- common in machine learning contexts, new to psychological research
- existing data = training data + testing data
- fit the model repeatedly (k times) on training data, by leaving out a subset of the data and then test the model on the left-out dataset
  - “leave one out” cross validation (LOOCV)
  - k-fold cross validation
- also: regularization: lasso regression



Perspectives on Psychological Science
Volume 12, Issue 6, November 2017, Pages 1100-1122

https://doi.org/10.1177/1745691617693393

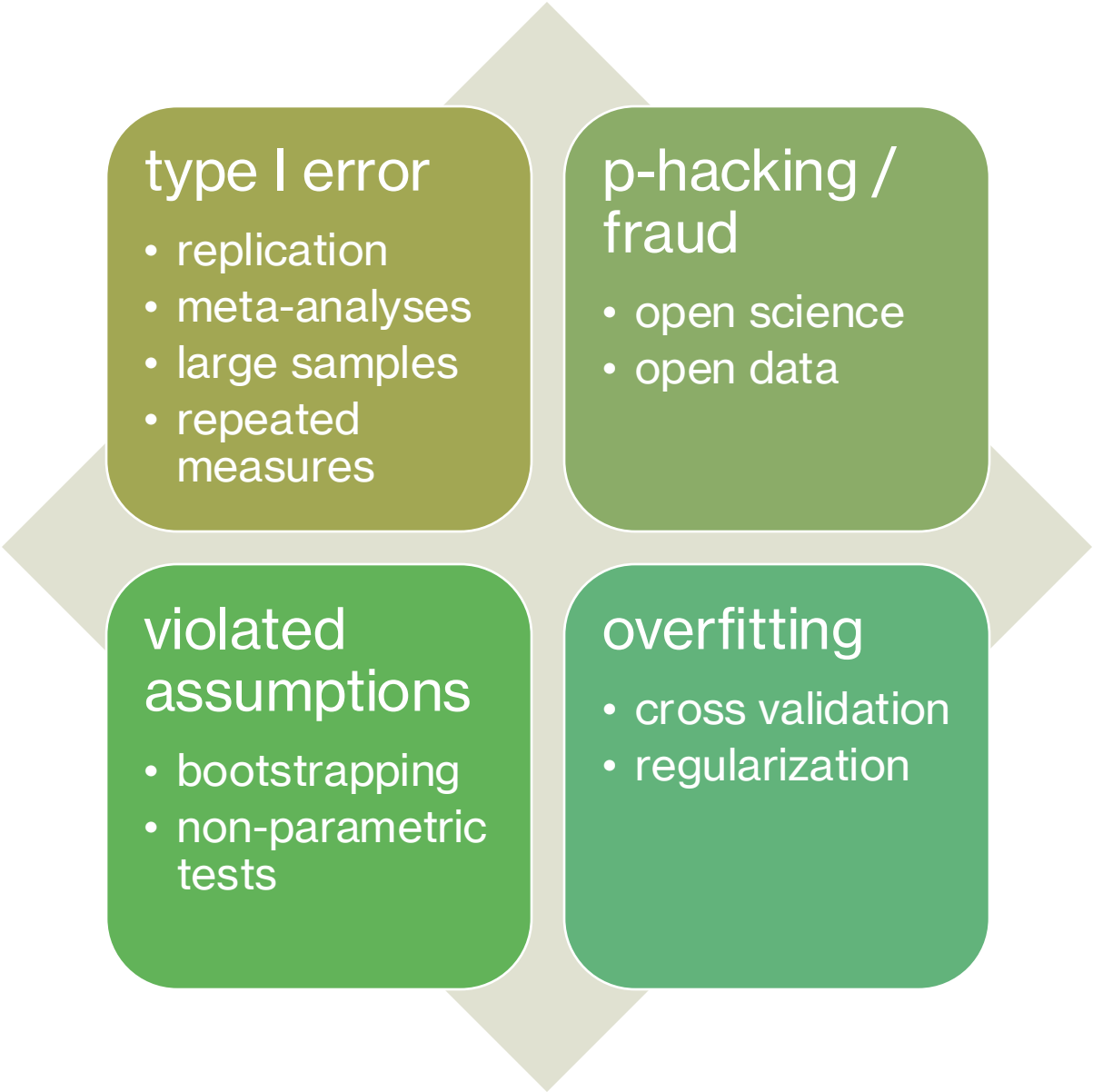
SAGE journals

Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning

Tal Yarkoni and Jacob Westfall

Abstract
Psychology has historically been concerned, first and foremost, with explaining the causal mechanisms that give rise to behavior. Randomized, tightly controlled experiments are enshrined as the gold standard of psychological research, and there are endless investigations of the various mediating and moderating variables that govern various behaviors. We argue that psychology’s near-total focus on explaining the causes of behavior has led much of the field to be populated by research programs that provide intricate theories of psychological mechanism but that have little (or unknown) ability to predict future behaviors with any appreciable accuracy. We propose that principles and techniques from the field of machine learning can help psychology become a more predictive science. We review some of the fundamental concepts and tools of machine learning and point out examples where these concepts have been used to conduct interesting and important psychological research that focuses on predictive research questions. We suggest that an increased focus on prediction, rather than explanation, can ultimately lead us to greater understanding of behavior.

replication crisis: safeguards



replication crisis: safeguards

Error Tight: Exercises for Lab Groups to Prevent Research Mistakes

Julia F. Strand
Department of Psychology, Carleton College

Abstract

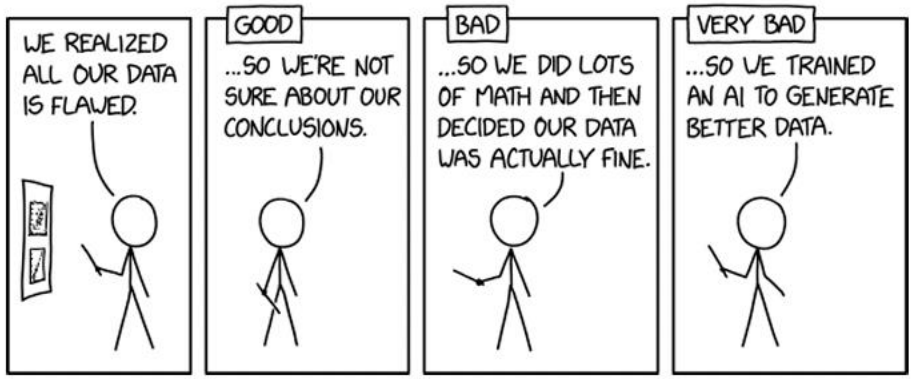
Scientists, being human, make mistakes. We transcribe things incorrectly, we make errors in our code, and we intend to do things and then forget. The consequences of errors in research may be as minor as wasted time and annoyance, but may be as severe as losing months of work or having to retract an article. The purpose of this tutorial is to help lab groups identify places in their research workflow where errors may occur and identify ways to avoid them. To do this, this article applies concepts from human factors research on how to create lab cultures and workflows that are intended to minimize errors. This article does not provide a one-size-fits-all set of guidelines for specific practices to use (e.g., one platform on which to backup data); instead, it gives examples of ways that mistakes can occur in research along with recommendations for systems that avoid and detect them. This tutorial is intended to be used as a discussion prompt prior to a lab meeting to help researchers reflect on their own processes and implement safeguards to avoid future errors.

Translational Abstract

Everyone makes mistakes. In science, mistakes can occur in many ways: Researchers may transcribe things incorrectly, make typos when writing code to analyze data, forget to do something they intended to, and so forth. These mistakes may simply waste time or require redoing work, but in more serious cases, they can ruin an experiment or lead to false conclusions. However, learning how to avoid errors in research isn't a standard part of training. This tutorial is intended to help lab groups identify places in the research process where errors may occur and identify ways to avoid them. To do so, this article draws on lessons from high-risk fields such as aviation, surgery, and construction, all of which have developed explicit, practical strategies to reduce mistakes on the job. This tutorial is intended to be used as a discussion prompt before a lab meeting to help researchers reflect on their own processes and implement safeguards to avoid future errors.

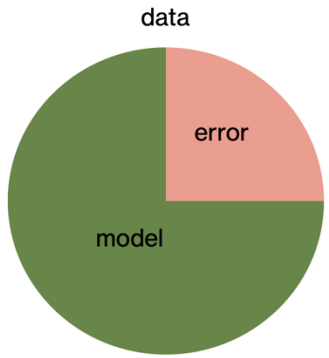
Keywords: error detection, independent verification, mistakes

final thoughts



- statistics is often taught from the framework of different-tests-for-different-data
- but...the same principles underlie most tests you encounter
- there ALSO exist methods of analysis that do not heavily rely on p-values (like frequentist statistics do) and account for prior information in making inferences (Bayesian statistics)
- keep an open mind and try to find connections between methods you read about and see around you!

review: data = model + error



- the goal of statistics is to find a simple explanation for the variation in the observed data, i.e., build a *model* of the data that approximates/explains it as well as possible
- over the course of the semester, we have encountered different kinds of data and models we can use to explain variation in those data
- review sheet

Reviewing Statistical Tests
data = model + error

Test	data	model	error
z-test			
one-sample t-test			

next time

- short review
- class reflections
- extra credit winner announcements
- BCQs

Before Class

- Do some ungraded practice

  -Weeks 1-6 Practice

  -Weeks 7-12 Practice

  -Weeks 13-15 Practice
- Submit any lingering questions here!
- Submit Final Class Survey

Here are the to-do's for this week:

- Submit Problem Set 7
- Submit any lingering questions here!
- Extra credit opportunities
  - Surveys
    - Final class survey
    - Vote in the Memer Contest
    - Complete Statistics Attitudes Survey
  - Analyze a Research Paper