Null-Hypothesis Significance Testing (NHST) critiques

EDUC 641: Unit 5

David D. Liebowitz

UNIVERSITY OF OREGON

Goals of the unit

- Articulate modern critiques of null-hypothesis significance testing framework
- Describe strategies to improve replicability and generalizability of quantitative research

Critiques of NHST

Two lines of criticism:

1. Concerns about generalizability and "statistical significance"
2. A different way of thinking about probability (Bayesian)

Generalizability

Research in WEIRD countries

- In 2015, 92 percent of all papers in developmental psychology featured participants from English-speaking countries and non-English-speaking Europe[1]
- In general, in social science much of our knowledge base comes from research conducting on participants living in Western, educated, industrialized, rich and democratic **(WEIRD)** nations

Non-representative samples

- Even in WEIRD countries, many samples are not representative of the national population
  - University students (at highly research-active universities)
  - Non-Hispanic, White and male patients more likely to be included in clinical trials
  - Convenience sampling (especially in intervention and survey research)
  - Non-stratified samples prevents conclusion being drawn about low-*n* groups
- Small samples are particularly at risk for returning idiosyncratic estimates

Learn more about all of this in EDUC 612 and 646!!

[1] Nielsen, M. Haun, D. Kartner, J & Legare, C. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology, 162*, 31-38.

Problems with "statistical significance"

Problems with "significance"

An explosion of academic and popular press attention to the "replication crisis" emerged in the early 2010s.

Failure to reproduce

Replication work often seen as un-original; weak incentives to conduct replications.

Many replications fail:

- New contexts
- Incomplete communication of treatment and conditions
  - Missing protocol, methodological or data management details
  - Incomplete specification of sequence of tests
- Disagreement as to what "counts" as a replication
  - **In design:** Exact replication, conceptual replication, replication + extension
  - **In outcome:**

Table 1
Six Replication Goals and Descriptions

No.	Goal	Recommended analysis	Success criterion
1	To infer the existence of a replication effect	Repeat analysis of original study	$p < .05$
2	To infer a null replication effect	Equivalence test	Confidence interval falls completely inside region of equivalence
3	To precisely estimate the replication effect size	AIPE, construct confidence interval for effect size	Effect size estimated with desired level of precision
4	To combine replication sample data with original results	Construct confidence interval for the average effect size of replication and original studies	Building on prior knowledge: more precise estimate of the effect of interest
5	To assess whether replication is clearly inconsistent with original	Construct confidence interval for the difference in effect sizes	Confidence interval for difference in effect sizes does not include 0
6	To assess whether replication is clearly consistent with original	Equivalence test, using confidence interval for the difference in effect sizes	Confidence interval for difference in effect sizes falls completely inside region of equivalence

Anderson, S. & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1-12.

Problems with significance

Some problems with the binary decision-making process of NHST include:

- Statistical significance as a condition for publication and a goal for researchers
  - $\rightarrow$ Publication bias: "successes" are published; "failures" end up in file drawers
- Novelty as a condition for publication in top-tier journals
  - If something is "unexpected" or "surprising" there's a decent chance it might not be true
- Frequent imprecise estimates of "differences" between groups, with little interest in quantifying the magnitude of the differences
- Weak theory driving analysis:
  - No non-null hypotheses
  - No prior belief about likelihood of findings
  - Hypothesis After the Results are Known (HARK-ing)
- Outright manufactured data/analysis

Problems with "significance"

If p<0.05 is the goal, there are many ways to get there.

This can be the product of intentional "p-hacking" or "researcher degrees of freedom" (explore your data, test many different models, try this variable, etc.)

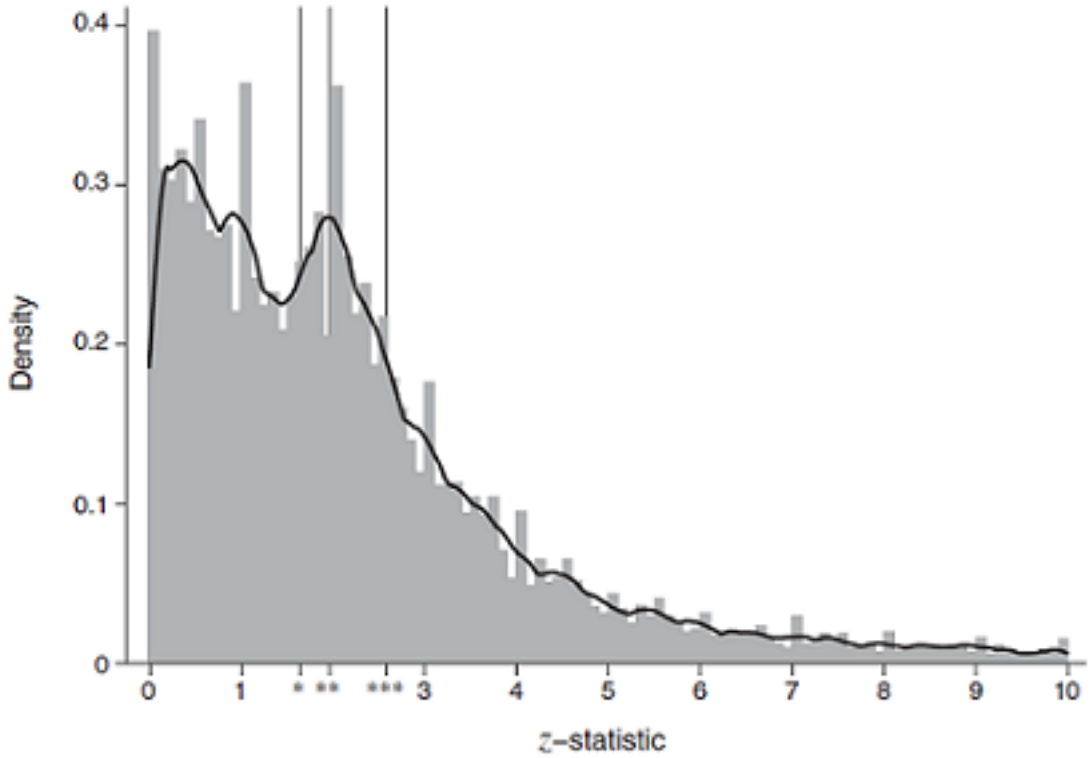
At the core of this problem is the need to balance (avoid) two types of inference errors

	Reject H_0	**Fail to reject**
H_0 = true	Type I error	Correct decision
H_0 = false	Correct decision	Type II error

The rate of false positives (**Type I errors**) is equal to your α-threshold. You choose this rate.

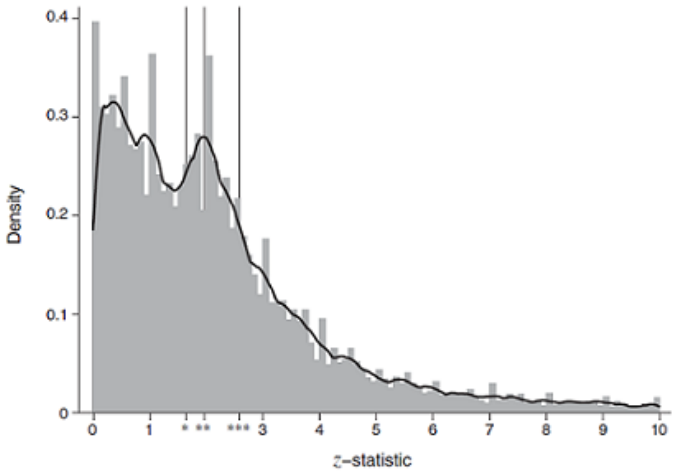
NHST and the trap of $p < 0.05$

If you set the goal at 0.05 (or 0.01 or 0.001), that's what you'll get:

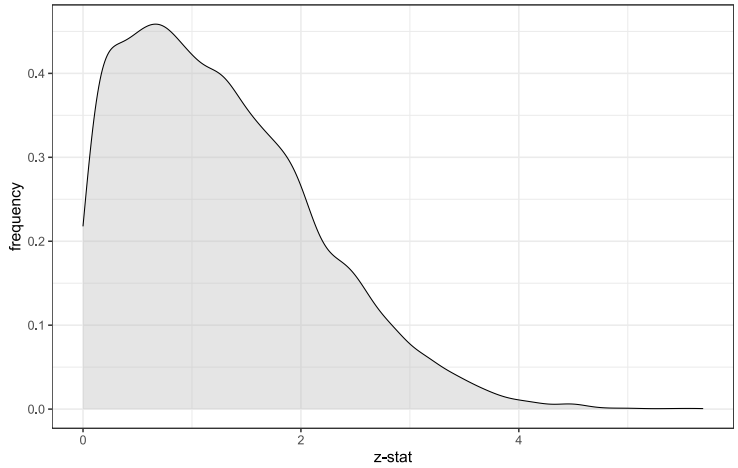


Brodeur, A., Cook, N. & Heyes, A. (2020). Methods matter: p-Hacking and publication bias in causal analysis in economics. *American Economic Review, 110*(11), 3634-3660.

What should that distribution be?

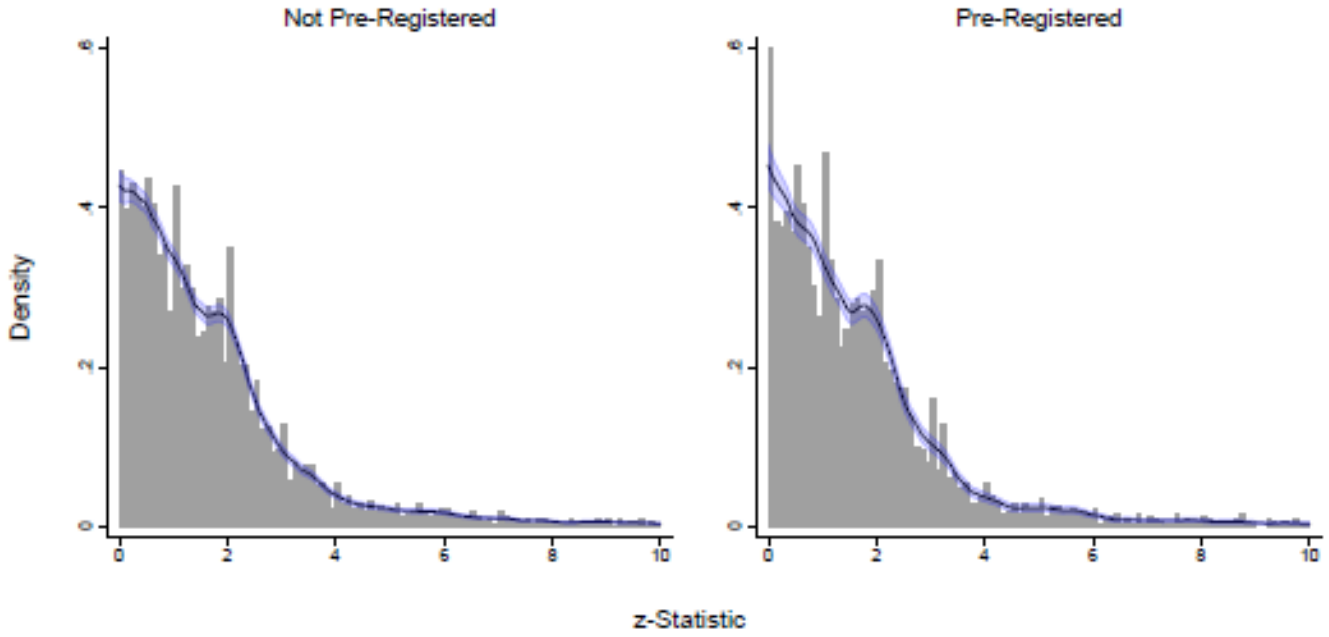


What we observe



H_0 false, 10,000 simulations,
$N = 30, \mu = 0.2, \sigma = 1$

Open Science alone won't save us



Notes: This figure displays the distribution of test statistics for $z \in [0, 10]$ from randomized control trials from 2018–2021 by pre-registration status. We define a pre-registered RCT as a study that was registered before its trial end date listed in a registry. Studies that were registered after the trial end date are counted as non-pre-registered. The tests are from studies published in 15 leading economics journals. Bins are 0.1 wide. We have also superimposed an Epanechnikov kernel. We do not weight articles.

Brodeur, A., Cook, N. Hartley, J. Heyes, A. (2022). Do pre-registration and pre-analysis plans reduce p-hacking and publication bias?

Dangers of multiple hypothesis testing

If your goal is to find a "statistically significant" result, you will detect such a relationship 1 out of 20 times (on average).

Imagine rolling a die. What is the probability you roll a 1? $1/6 = 0.167$

Now, roll it twice, what is the probability **at least** one of your rolls is a 1?
$1 - (5/6 * 5/6) = 0.306$

If you conduct enough tests, you'll detect a relationship eventually.

One fix

Instead of using $\alpha = 0.05$ for each individual test, use $\alpha = 0.05$ for the **family of tests** when we examine multiple contrasts to test a single hypothesis.

Bonferroni method

Take a given α-threshold and "split it" across the entire family of tests. Assuming $\alpha = 0.05$:

- For 2 tests, conduct each at 0.025 level;
- For 3 tests, conduct each at 0.0167 level; etc. ...

Use this new threshold to identify the critical t-statistic given the number of degrees of freedom.

Other approaches exist! Bonferroni is an extremely conservative one--beware!

As tests increase, so do critical t-values:

# tests	# new α	t-statistic (df = ∞)
1	0.0500	1.96
2	0.0250	2.24
3	0.0167	2.39
4	0.0125	2.50
5	0.0100	2.58
6	0.0083	2.64
10	0.0050	2.81
20	0.0025	3.02
50	0.0010	3.29
100	0.0005	3.48

Type II error

Equally important as the rate of false positives is the number of false negatives (**Type II errors**) you're going to get. **Why do we care about this? Why not just set** $\alpha = 0.00000001$**?**

Failure to detect an effect when there is one there is just as important (but more often ignored) of a concern. Under-powered studies (studies that suffer from a high-likelihood of Type II error) have just as much potential to mis-inform us about the relative value of a particular intervention, practice or policy.

Type II error depends on the magnitude of the effect, the number of observations, the total variation in your variables, and the amount of the variation in the outcome you can explain.

Other data and analysis challenges

Small, subjective researcher choices *can* have meaningful effects on results of study.

Take the "Many Analysts"[1] project in which 29 teams used the same dataset to address the same question: "Are soccer referees more likely to give red cards to dark-skin-toned players than to light-skin-toned ones?"

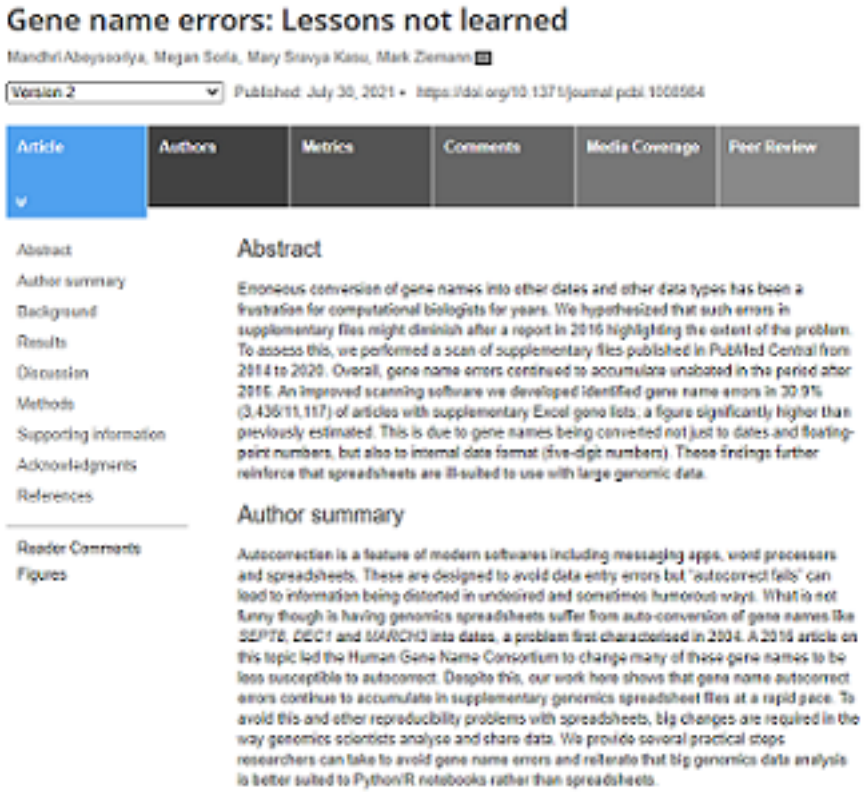
- Estimated effect sizes ranged from 0.89 to 2.93 (in odds-ratio units)
- 20 teams found a "significant" effect, 9 did not

Critically, these were all studies done with the **exact same data** and the **exact same research question**!

[1] Silberzahn, R. et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356.

Other data and analysis challenges

Never **EVER** use Excel to read in your data



Gene name errors: Lessons not learned

Mandhri Abeysooriya, Megan Soria, Mary Sravya Kasu, Mark Ziemann

Version 2 Published: July 30, 2021 • https://doi.org/10.1371/journal.pcbi.1008984

Article | Authors | Metrics | Comments | Media Coverage | Peer Review

Abstract
Author summary
Background
Results
Discussion
Methods
Supporting information
Acknowledgments
References
Reader Comments
Figures

Abstract

Erroneous conversion of gene names into other dates and other data types has been a frustration for computational biologists for years. We hypothesized that such errors in supplementary files might diminish after a report in 2016 highlighting the extent of the problem. To assess this, we performed a scan of supplementary files published in PubMed Central from 2014 to 2020. Overall, gene name errors continued to accumulate unabated in the period after 2016. An improved scanning software we developed identified gene name errors in 30.9% (3,436/11,117) of articles with supplementary Excel gene lists; a figure significantly higher than previously estimated. This is due to gene names being converted not just to dates and floating-point numbers, but also to internal date format (five-digit numbers). These findings further reinforce that spreadsheets are ill-suited to use with large genomic data.

Author summary

Autocorrection is a feature of modern softwares including messaging apps, word processors and spreadsheets. These are designed to avoid data entry errors but "autocorrect fails" can lead to information being distorted in undesired and sometimes humorous ways. What is not funny though is having genomics spreadsheets suffer from auto-conversion of gene names like *SEPT8*, *DEC1* and *MARCH3* into dates, a problem first characterised in 2004. A 2016 article on this topic led the Human Gene Name Consortium to change many of these gene names to be less susceptible to autocorrect. Despite this, our work here shows that gene name autocorrect errors continue to accumulate in supplementary genomics spreadsheet files at a rapid pace. To avoid this and other reproducibility problems with spreadsheets, big changes are required in the way genomics scientists analyse and share data. We provide several practical steps researchers can take to avoid gene name errors and reiterate that big genomics data analysis is better suited to Python/R notebooks rather than spreadsheets.

More importantly, without being able to observe researcher decisions in their code, we can't assess the reasonableness of choices and the potential for replication. **We need to learn how to code and not just point-and-click for Open Science.**

The convoluted logic of NHST

Many of the preceding issues result from the fundamentally convoluted rationale of the null-hypothesis significance testing approach. An example:

Assume that we find that average national life expectancy is significantly greater in high-income countries compared to low-income countries. With an α of .05, the (hypothetical) statistical test returns p = .03, meaning we can reject the null hypothesis.

Which of the following statements is correct, given p = .03:

- **A.** There is a 3% probability that high-income countries do not have a higher average life expectancy (the null hypothesis).
- **B.** There is a 3% probability that the results are due to sampling idiosyncrasy, rather than a true relationship in the population.
- **C.** A statistically significant difference means higher income-levels yield higher life expectancies.
- **D.** The observed data would occur 3% of the time if the null hypothesis were true.
- **E.** Previous research found that in 2005, income-level was not related to average life expectancy (p = .17). Therefore, income-level is more stronyl related to life expectancy in 2015 than in 2005.
- **F.** Another research group finds a statistically significant difference between countries with socialized medicine compared to those without (p = .001). Their smaller p value means that socialized medicine is more effective at increasing life expectancy than income level.

The convoluted logic of NHST

Only D on the previous slide is correct, yet all of these are commonly expressed in the scientific literature!

In frequentist statistics, we establish an **objective decision rule**: if our observed data has less than a 5 percent chance of occurring (or less than 1% or less than 0.1% or less than 10%) due to sampling idiosyncrasy, we will conclude that the observed relationship in the sample represents a true relationship in the population.

- Once we set a threshold (an α-theshold), we are making a binary decision about significance and non-significance
- Relationships are *NOT* "more" or "less" significant[1]
  - In fact, the distribution of *p*-values under the null hypothesis is a uniform one; thus, it's not correct to describe observed vs. expected probabilities in relationship to each other (e.g., a *p*-value of 0.01 is **not** three times less likely than a *p*-value of 0.03).
- This has become such an issue that in 2016, the American Statistical Association released a statement touching on many of these issues, and a subset of committee members explicitly recommended to not use the term "statistically significant"

- Many of these concerns relate to different ways of thinking about probability...

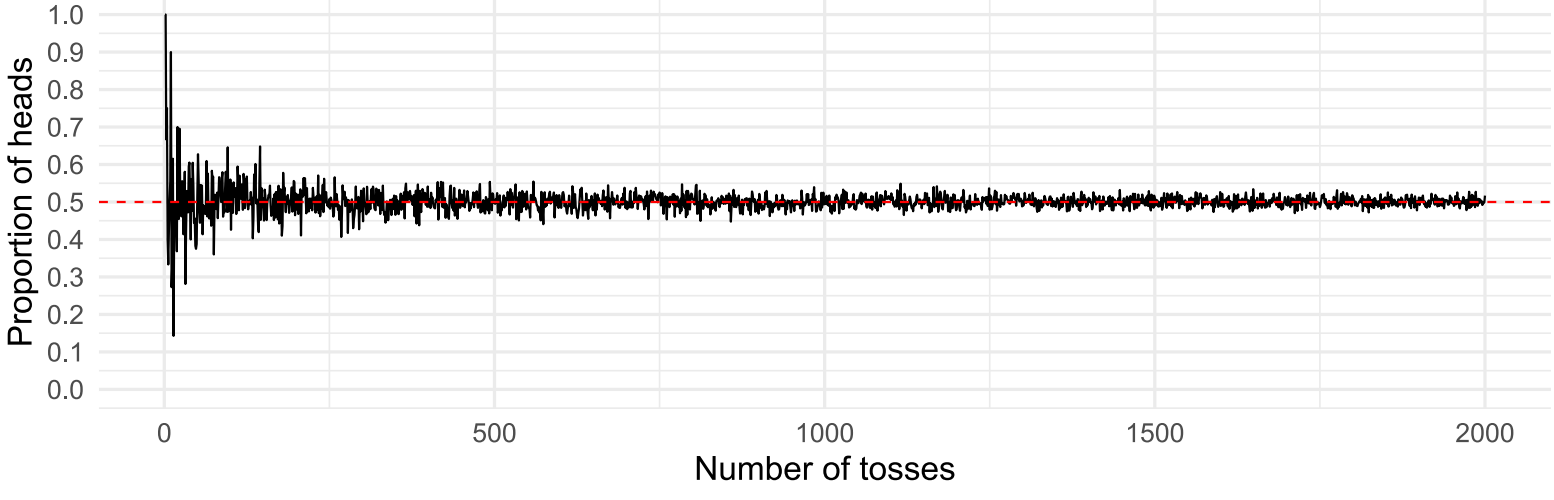
[1] This is the classical/standard way of thinking about inference. More recently, scholars have begun to question this way of thinking and consider this sort of comparison appropriate.

Bayesian probability and inference

(Very) brief intro to probability

Two common ways to think about probability: the long-run view (frequentist) and priors (Bayesian)

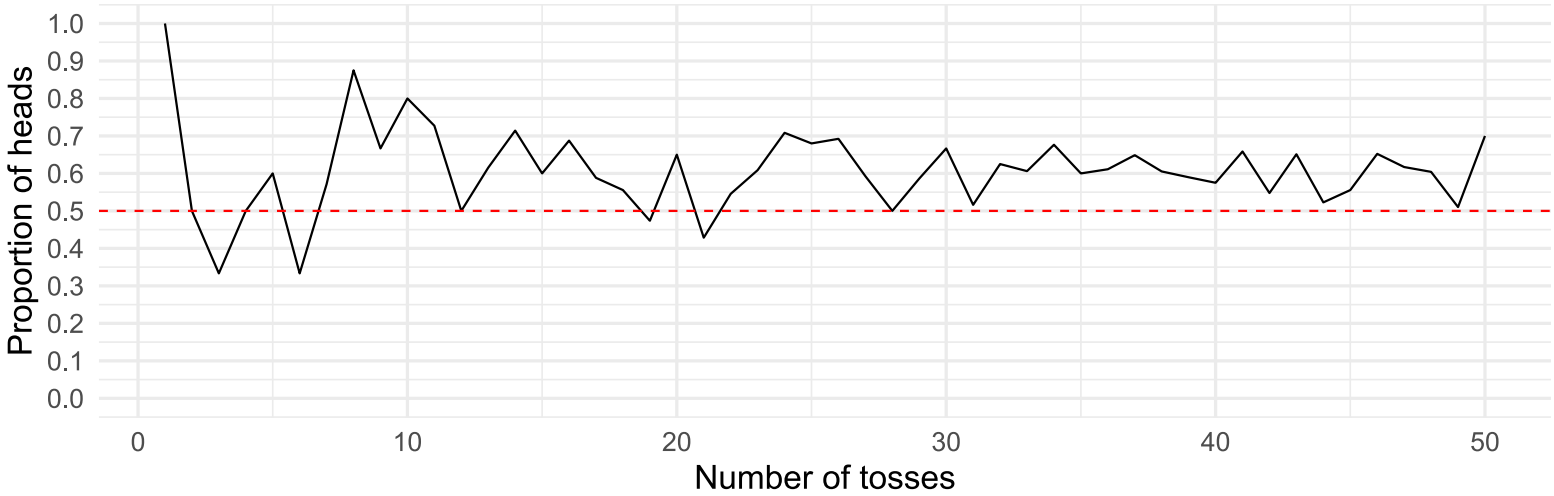
- **Frequentist:**
- With a known long-run outcome (e.g., a fair coin comes up heads half the time), we can consider the likelihood of a given short-run event
  - A coin flipped twice will come up Heads both times with a probability of 0.25 (fairly likely); but a coin flipped 20 times will come up Heads every time with a probability of 0.000000095 (very unlikely)
  - In the short-run, you can expect some variability, but in the long run, it will converge to the known distribution



Probability and testing H_0

We are asking how likely is it that we observe event X happening with Y frequency, if the expected probability of X happening is Z?

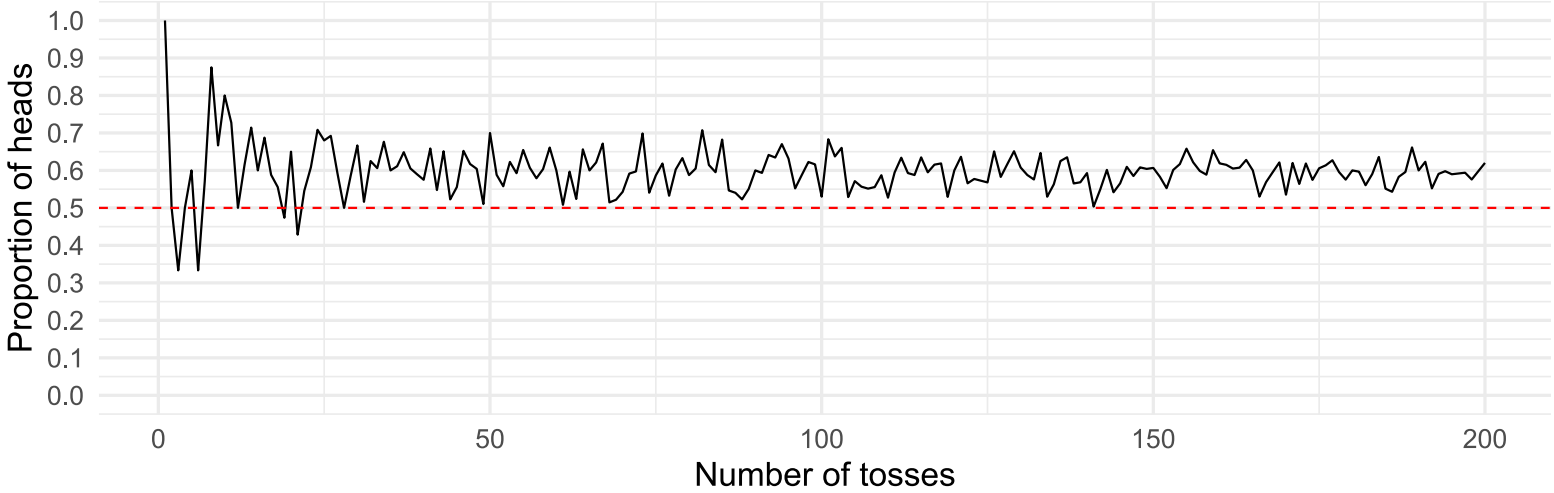
Imagine you were handed a coin and asked to determine whether it was weighted to one side. How certain would you be after 50 flips?



Probability and testing H_0

We are asking how likely is it that we observe event X happening with Y frequency, if the expected probability of X happening is Z?

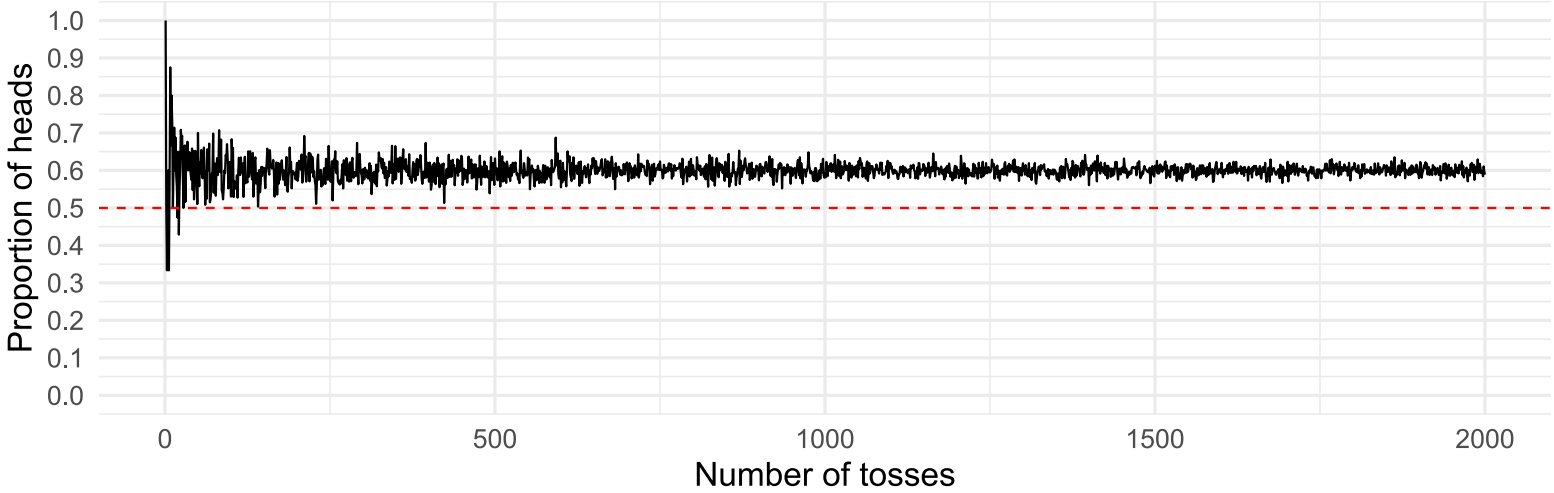
Imagine you were handed a coin and asked to determine whether it was weighted to one side. How certain would you be after 200 flips?



Probability and testing H_0

We are asking how likely is it that we observe event X happening with Y frequency, if the expected probability of X happening is Z?

Imagine you were handed a coin and asked to determine whether it was weighted to one side. How certain would you be after 2000 flips?



Problems with probability

In frequentist statistics, we assume that the null hypothesis has as good a chance as any other hypothesis at being true, and we then test how likely we are to observe the data as we see them when the null is actually true...**but is this the right framework???**

How believable are the following findings?

- People who wear glasses are more likely to be empathetic $(p = 0.023)$
- Medical doctors have fewer close relationships than those in other professions $(p = 0.007)$
- People who engage in power posing experience increased testosterone $(p = 0.045)$ (*a real paper*)

Perhaps it is wiser to start with a *prior* belief about the probability of an event?

Bayesian statistics

In contrast to frequentist approaches, Bayesian probability takes into account one's prior beliefs in determining the probability of an event

- Key insight: adjust current beliefs based on previous knowledge/beliefs
  - Adjust prior belief towards evidence in observed data
- Prior beliefs could come from theory, research or personal belief

Bayes Theorem (don't need to know this):

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

I found this to be a helpful explainer video.

Bayesian Example

I have a theory that Oregonians are (generally) outdoors-y. I meet an outdoors-y person. What is the probability I have met an Oregonian?

$P(Oregonian) = 0.02$ (to make it simple 1 of 50 U.S. states)
$P(Outdoorsy|Oregonian) = 0.65$
$P(Outdoorsy) = 0.20$

$$P(Oregonian|Outdoorsy) = \frac{(0.65)(0.02)}{0.20} = 0.065$$

Not surprisingly, I've updated my beliefs and now think there is a greater-than-base-rate likelihood (over 3x) that this person is an Oregonian. But it also tells me not to be too excited and claim the probability is 0.65 that I've met an Oregonian just because they are outdoors-y.

- Inferential statisticians (and applied researchers like yourselves) make statements like, "the chance we would observe differences between the treatment and control as large as the ones we did when there was actually no effect is less than 5%"
- Bayesian statisticians (and applied researchers like yourselves) make statements like, "the chance that the treatment is more effective than the control is 92%"

An applied example

University instructors are appraised based (in part) on student evaluations. Fraile & Bausch-Morrell (2010) collected data on these evaluations over two years.

They use the sample estimates from the first year of data, to estimate instructor evaluation ratings in the second year -- *given their prior evaluation ratings*.

The range of the Bayesian estimates for the Year 2 student evaluation ratings is narrower than the sample estimates (shrinks from score range of 2 - 6 to 2.5 - 5.5). The precision with which they make their estimates is much tighter. More to come on confidence intervals (or in Bayesian terms, credible intervals) in EDUC 643.

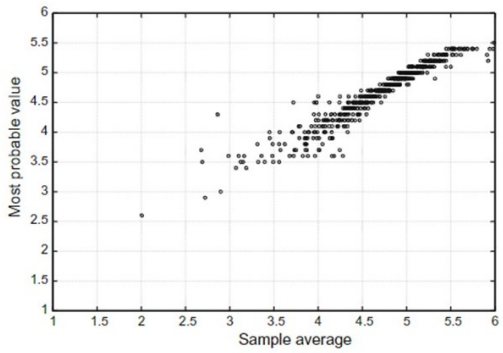


Fig. 6 Change in lecturers' evaluations as a result of estimating them as the most probable value provided by Bayesian inference instead of using sample means

With our data

```
bayes ← stan_glm(life_expectancy ~ schooling, data = who,
                 prior=normal(3, 1), prior_intercept = normal(54, 4))
```

```
#>
#> SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
#> Chain 1:
#> Chain 1: Gradient evaluation took 0 seconds
#> Chain 1: 1000 transitions using 10 leapfrog steps per transition would tak
#> Chain 1: Adjust your expectations accordingly!
#> Chain 1:
#> Chain 1:
#> Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
#> Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
#> Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
#> Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
#> Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
#> Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
#> Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
#> Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
#> Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
#> Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
#> Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
```

With our data

```
print(bayes)
```

```
#> stan_glm
#>  family:       gaussian [identity]
#>  formula:      life_expectancy ~ schooling
#>  observations: 173
#>  predictors:   2
#> ------
#>             Median MAD_SD
#> (Intercept) 42.5    1.6
#> schooling    2.2    0.1
#>
#> Auxiliary parameter(s):
#>       Median MAD_SD
#> sigma 4.6    0.3
#>
#> ------
#> * For help interpreting the printed output see ?print.stanreg
#> * For info on the priors used see ?prior_summary.stanreg
```

Why not Bayes?

1. Selecting a prior can be **subjective**
2. Difficulty in finding informative prior
3. Disputes over value of uninformative priors
4. Similarity in results across approaches
5. Stasis

Much more complexity to be explored in another course

A path forward...

So where does all this leave us? Some recommendations for moving forward:

1. Move beyond a single metric to evaluate practice and theory
    - Multiple studies w/ focus on generalizability
    - Replication studies
    - Systematic reviews/meta-analysis
    - Focus on *confidence intervals* and *magnitude* of effects (**more in EDUC 643**)
2. Open Science
    - Pre-registration (REES, OSF, etc.)
    - Registered reports
    - Public materials and data
    - Use of scripts!!!
    - Statistical tests (GRIM, SPRITE, etc.)
3. Changes in academic incentives
4. Recognition of the impact that subjective research decisions have on (some) quantitative empirical results

Synthesis and wrap-up

Goals of the unit

- Articulate modern critiques of null-hypothesis significance testing framework
- Describe strategies to improve replicability and generalizability of quantitative research

To Dos

Final project

Due Dec. 7, 11:59pm

Re-submission

Assignments <13.5, resubmissions due on Canvas 12/2 at 5pm

End-of-term SES

Thanks!