Null-Hypothesis Significance Testing (NHST) critiques

EDUC 641: Week XX

TBD

UNIVERSITY OF OREGON

Goals of the unit

- Describe relationships between quantitative data that are continuous
- Visualize and substantively describe the relationship between two continuous variables
- Describe and interpret a fitted bivariate regression line
- Describe and interpret components of a fitted bivariate linear regression model
- Visualize and substantively interpret residuals resulting from a bivariate regression model
- Conduct a statistical inference test of the slope and intercept of a bivariate regression model
- Write R scripts to conduct these analyses
- Articulate modern critiques of null-hypothesis significance testing framework
- Describe strategies to improve replicability and generalizability of quantitative research

Critiques of NHST

1. Re-thinking concepts of "statistical significance"
2. A different way of thinking about probability (Bayesian)

Problems w/ "significance"

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

Aside from outright manufactured data, some problems include:

- $p<0.05$ as a condition for publication and a goal for researchers
  - $\rightarrow$ Publication bias: "successes" are published; "failures" end up in file drawers

More problems

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

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

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

- Weak theory:
  - Low-precision "differences" between groups, with little interest in quantifying the differences
  - No non-null hypotheses
  - No prior belief about likelihood of findings
  - Hypothesis After the Results are Known (HARK-ing)

NHST and the trap of $p < .05$

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 random chance, rather than the effect of income level.
- **C.** A statistically significant difference means higher income-levels yield higher life expectancies.
- **D.** The observed data would only occur 3% of the time if the null hypothesis was true.
- **E.** Previous research found that in 2005, income-level failed to produce a significant difference in average life expectancy (p = .17). Therefore, income-level is more influential on 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.

NHST and the trap of *p*-values

None of the previous interpretations were correct, yet these are commonly expressed in scientific literature!

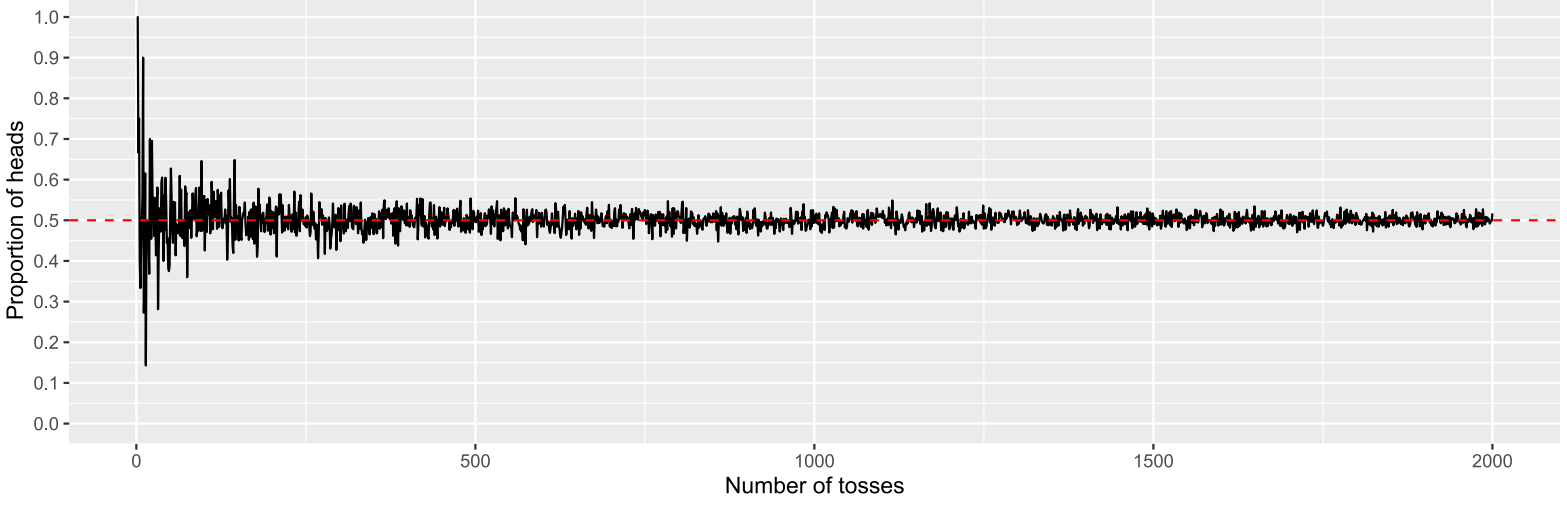
In frequentist statistics, we establish an objective decision rule: if our observed data has less than a 5 percent chance of occurring (or less than 1% or less than 0.1% or less than 10%) due to sampling idiosyncrasy, we will conclude that the observed relationship in the sample represents a true relationship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

- 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
    - In fact, the distribution of *p*-values on either side of the alpha threshold is assumed to be uniform; thus, it's not correct to describe observed vs. expected probabilities in relationship to each (e.g., a *p*-value of 0.01 is **not** three times less likely than a *p*-value of 0.03).
- Many of these concerns (and conflicts) relate to different ways of thinking about probability

(Very) brief intro to probability

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

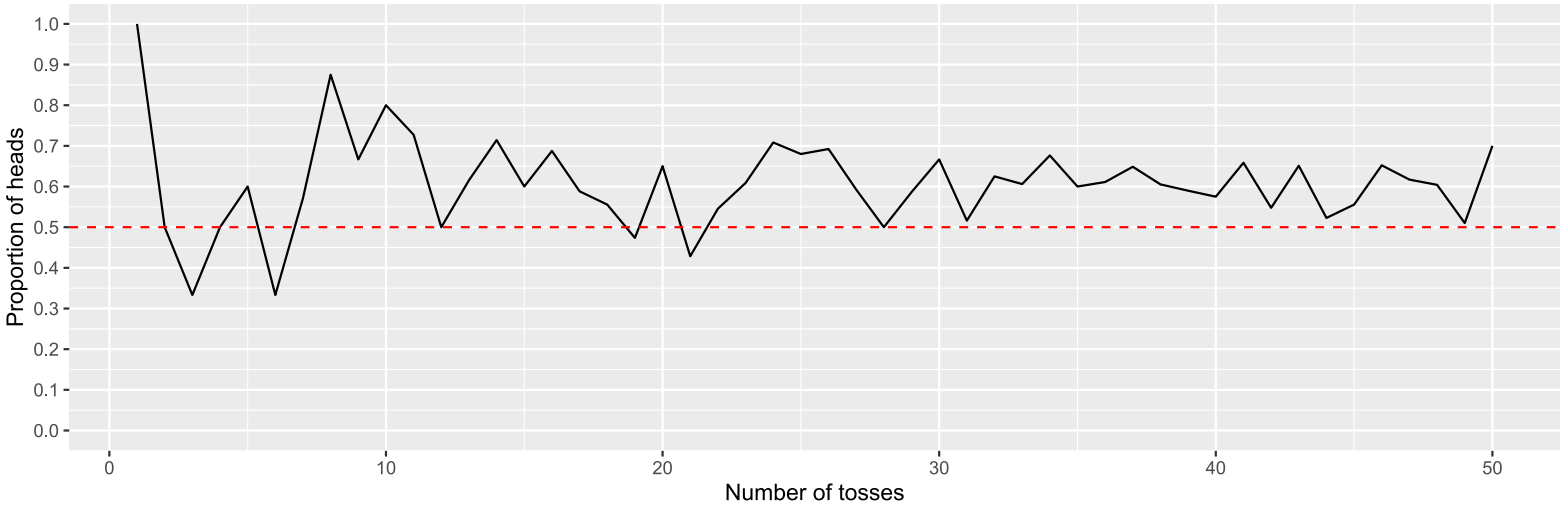
- **Frequentist:**
- With a known long-run outcome (e.g., a fair coin comes up heads half the time), we can consider the likelihood of a given short-run event
  - A coin flipped twice will come up Heads both times with a probability of 0.25 (fairly likely); but a coin flipped 20 times will come up Heads 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?

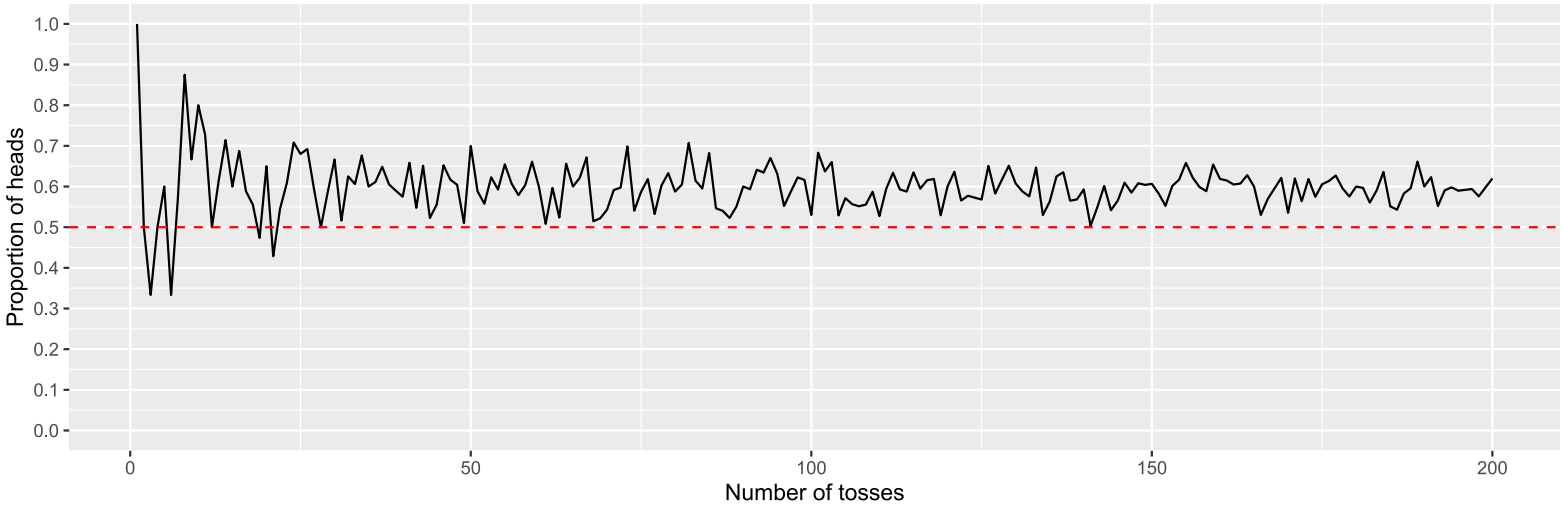
Imagine you were handed a coin and asked to determine whether it was weighted to one side. How certain would you be after 50 flips?



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?

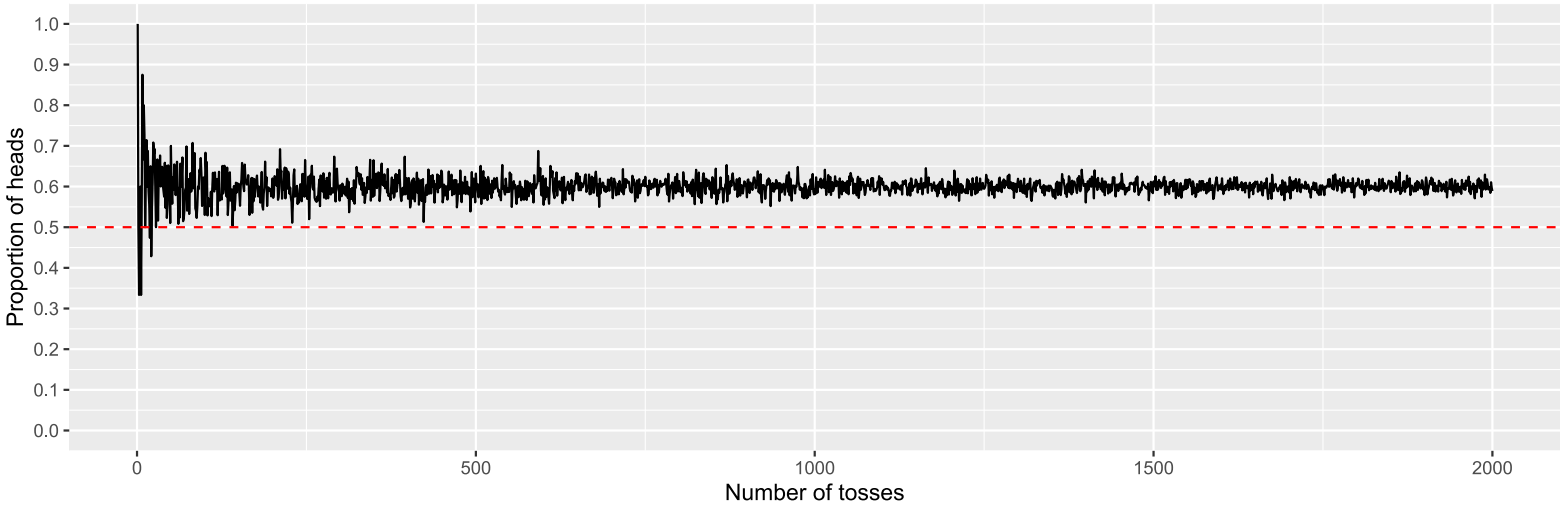
Imagine you were handed a coin and asked to determine whether it was weighted to one side. How certain would you be after 200 flips?



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):

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

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.

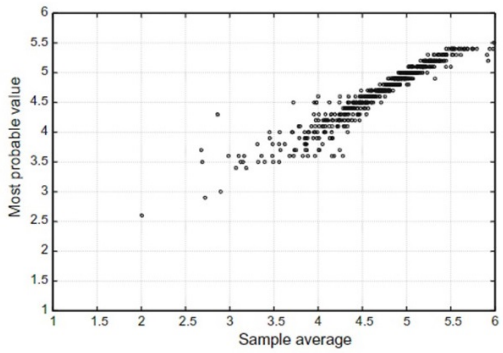


Fig. 6 Change in lecturers' evaluations as a result of estimating them as the most probable value provided by Bayesian inference instead of using sample means

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
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