P-hacking, Spurious Significance, and the Bayesian Diagnosis

A Pedagogical Treatment

Outline

Setup

Research question: Does a new teaching method improve student test scores?

- $n = 30$ students, treatment vs. control
- The true effect is zero — the method does nothing
- Researcher runs a t-test and explores the following choices

Decision	**Options explored**
Outlier rule	All / exclude $>$ 2SD / exclude $>$ 1.5SD
Covariates	None / GPA / GPA + gender
Sample	All / exclude late enrollees / exclude absentees
Directionality	Two-tailed / one-tailed
Outcome timing	Week 4 / Week 6 / Week 8

The Combinatorial Explosion

With **5 binary decisions**, there are up to $2^5 = 32$ specifications.

Under H_0, each specification has a $\sim 5\%$ false positive rate.

The probability of **at least one** significant result is:

$$1 - (0.95)^{32} \approx \mathbf{0.81}$$

Key point

There is roughly an **81% chance** of finding at least one "significant" result purely by chance, if the researcher reports only the best-looking specification.

The Bayesian Diagnosis of P-hacking

The reported p-value is valid *conditional on a single pre-specified test*. But:

- The proper frequentist correction (Bonferroni) requires

$$p < \frac{0.05}{32} \approx 0.0016$$

- From the **likelihood principle**: inference should depend only on the data observed, not on the hypothetical tests that *could* have been run
- P-hacking distorts the sampling distribution of the p-value — the unreported search path is invisible to the reader

Bayesian remedy

Bayesian model averaging (BMA) or multiverse analysis makes fragility *immediately visible*: posterior estimates that shift substantially across specifications signal the data do not support a robust conclusion.

Setup

Let $X \sim \mathcal{N}(\theta, 1)$ and consider:

$$H_0 : \theta = 0 \qquad \text{vs} \qquad H_1 : \theta = 4$$

Suppose we observe $X = 1.8$.

The p-value:

$$p = P(X \geq 1.8 \mid \theta = 0) = 1 - \Phi(1.8) \approx 0.036 < 0.05$$

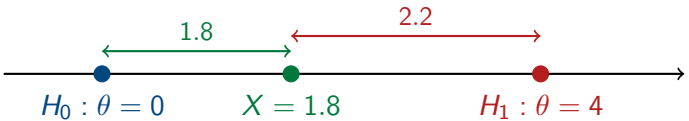
The result is "statistically significant" — a standard analysis rejects H_0.

The Likelihood Ratio Tells a Different Story

$$\frac{f(x \mid H_0)}{f(x \mid H_1)} = \frac{\phi(1.8-0)}{\phi(1.8-4)} = \frac{\phi(1.8)}{\phi(-2.2)} = \frac{\exp(-1.62)}{\exp(-2.42)} = \exp(0.80) \approx 2.23$$

Interpretation

The data are **more than twice as likely under H_0 than under H_1**. The observation $X = 1.8$ is closer to the null than to the alternative.



$X = 1.8$ sits in the tail of H_0 (hence small p-value), but in an *even more extreme* tail of H_1.

Bayesian Punchline

With equal priors $P(H_0) = P(H_1) = 0.5$:

$$P(H_1 \mid X = 1.8) = \frac{1}{1 + \mathrm{BF}_{01}} = \frac{1}{1 + 2.23} \approx \mathbf{0.31}$$

- Despite $p = 0.036$, the posterior probability of H_1 is only **31%**
- The Bayes factor $\mathrm{BF}_{10} \approx 0.45$ constitutes *evidence in favour of* H_0

Core lesson

This is a clean instance of the **Jeffreys–Lindley paradox** in spirit: *significance* and *evidential support for* H_1 are not the same thing, and can point in opposite directions. A p-value only measures distance from H_0 — it is completely blind to where H_1 is.

Setting up the Equation

We seek prior odds $\frac{P(H_1)}{P(H_0)}$ such that $P(H_1 \mid X = 1.8) = 0.90$.

By Bayes' theorem:

$$P(H_1 \mid X) = \frac{1}{1 + \mathrm{BF}_{01} \cdot \frac{P(H_0)}{P(H_1)}}$$

Setting equal to 0.9:

$$1 + \mathrm{BF} * 01 \cdot \frac{P(H_0)}{P(H_1)} = \frac{1}{0.9} \implies \mathrm{BF} * 01 \cdot \frac{P(H_0)}{P(H_1)} = \frac{1}{9}$$

Solving for the Required Prior

Since $\mathrm{BF}_{01} = 2.23$:

$$\frac{P(H_0)}{P(H_1)} = \frac{1}{9 \times 2.23} = \frac{1}{20.07}$$

Therefore the required **prior odds in favour of** H_1 are:

$$\frac{P(H_1)}{P(H_0)} \approx 20$$

which corresponds to:

$$\boxed{P(H_1) = \frac{20}{21} \approx 0.952, \qquad P(H_0) = \frac{1}{21} \approx 0.048}$$

Interpretation

What does this mean?

To obtain $P(H_1 \mid X) = 0.9$ from data that actually *favour* H_0 by a factor of 2.23, one must enter the analysis believing H_1 is **20 times more probable** than H_0 a priori.

- This is the implicit prior a researcher is effectively assuming when they treat $p < 0.05$ as strong evidence for H_1 in this setting
- It reflects a prior commitment to H_1 so strong that it overwhelms the likelihood evidence pointing the other way

Fundamental Bayesian message

A p-value **cannot be interpreted as evidence without an implicit prior**, and that prior is often unreasonably optimistic about H_1. Making the prior explicit is not a weakness of Bayesian analysis — it is its central virtue.

Summary

Scenario	p-**value**	$P(H_1 \mid X)$
Equal priors, $X = 1.8$, $H_1 : \theta = 4$	0.036	0.31
Prior odds 20:1 in favour of H_1	0.036	0.90

Three interlocking lessons:

1. **P-hacking** inflates false positives multiplicatively across specifications
2. **A significant** p-**value** can coexist with a Bayes factor favouring H_0
3. **Reaching a high posterior for** H_1 despite unfavourable data requires a heavily loaded prior — making the implicit assumptions of NHST transparent