An introduction to the theory and practice of Bayesian hypothesis testing: A workshop using JASP

Thomas J. Faulkenberry

Tarleton State University

Plan for today

Our guiding questions:

- WHY should I change how I do statistics?
- WHAT should I do instead?
- HOW can I do it?

All materials can be found at:

- `http://github.com/tomfaulkenberry/bayesWorkshop`

An exercise

Suppose you have a treatment that you suspect may alter performance on a certain task. You find that the experimental group **significantly** outperforms the control group, $t(18) = 2.7$, $p = 0.01$.

How many of the following are True statements?

- The probability that the null hypothesis is true is 1%
- The probability that the research hypothesis is true is 99%
- If you reject the null, the probability that you are making the wrong decision is 1%
- If you repeated the experiment over and over, then you would get a significant result 99% of the time.

An exercise

Suppose you have a treatment that you suspect may alter performance on a certain task. You find that the experimental group **significantly** outperforms the control group, $t(18) = 2.7$, $p = 0.01$.

They're all false!

What is a p-value?

The p-**value** is the probability of obtaining a test statistic at least as extreme as the one that was observed, given that the null hypothesis is true...i.e., $P(D \mid \mathcal{H}_0)$.

Core logic:

- suppose the opposite of what we want (i.e., that there is no effect)
- show that this *under this null hypothesis*, our data is very unlikely (our p-value is very small)
- conclude that there must be an effect
- Fisher: as $p \to 0$, evidence for $\mathcal{H}_1 \to \infty$

Criticisms

Nelder (1999)

> *"The most important task before us in developing statistical science is to demolish the p-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology. (p. 261)"*

Nelder, J. A. (1999). From statistics to statistical science. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*, 257-269. 10.1111/1467-9884.00187

Criticisms

Lindley (1999)

> *"My personal view is that p-values should be relegated to the scrap heap and not considered by those who wish to think and act coherently. (p. 75)"*

Lindley, D.V. (1999). Comment on Bayarri and Berger. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics* (Vol. 6, p. 75). Oxford: Clarendon.

Criticisms

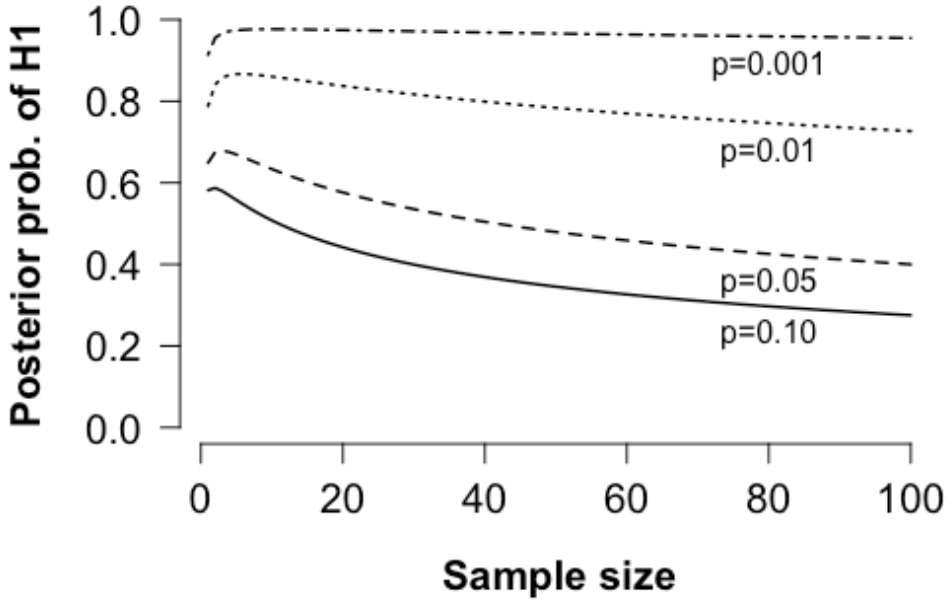
Townsend (2008)

> *"[R]ecent years have seen a lively debate about the value of null hypothesis testing, in addition to various means of improving the strategy and avoiding its threatening sloughs of despond. (p. 271)"*

Townsend, J. T. (2008). Mathematical Psychology: Prospects For The 21st Century: A Guest Editorial. *Journal of Mathematical Psychology, 52*, 269–280. 10.1016/j.jmp.2008.05.001

Criticisms

Berger & Sellke (1987) - p-values **overstate** evidence



Note: for a given p-value, that p-value becomes **less evidential** for $\mathcal{H}_1$ as sample size increases

Criticisms

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with *p*-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of *p*-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today

Criticisms

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless. Made

Bayesian inference

Bayes theorem:

$$\underbrace{p(\mathcal{H} \mid \text{data})}_{\substack{\text{Posterior beliefs} \\ \text{about hypothesis}}} = \underbrace{p(\mathcal{H})}_{\substack{\text{Prior beliefs} \\ \text{about hypothesis}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H})}{p(\text{data})}}_{\text{predictive updating factor}}$$

Bayesian inference

Natural action in science is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})} =$$

Bayesian inference

Natural action in science is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})} = \frac{p(\mathcal{H}_1) \cdot \frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data})}}{p(\mathcal{H}_0) \cdot \frac{p(\text{data}|\mathcal{H}_0)}{p(\text{data})}}$$

Bayesian inference

Natural action in science is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\begin{aligned}\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})} &= \frac{p(\mathcal{H}_1) \cdot \frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data})}}{p(\mathcal{H}_0) \cdot \frac{p(\text{data}|\mathcal{H}_0)}{p(\text{data})}} \\ &= \frac{p(\mathcal{H}_1) \cdot p(\text{data} \mid \mathcal{H}_1)}{p(\mathcal{H}_0) \cdot p(\text{data} \mid \mathcal{H}_0)}\end{aligned}$$

Bayesian inference

Natural action in our discipline is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{predictive updating factor}}$$

Bayesian inference

The predictive updating factor

$$B_{10} = \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}$$

tells us how much better $\mathcal{H}_1$ predicts our observed data than $\mathcal{H}_0$. This ratio is called the **Bayes factor**

Bayes factors

Example: suppose $B_{10} = 5$.

Interpretation: the observed data are 5 times more likely under the alternative hypothesis $\mathcal{H}_1$ than the null hypothesis $\mathcal{H}_0$.

This is taken as **positive evidence** for the alternative $\mathcal{H}_1$

Bayes factors

Jeffreys (1961) proposed the following thresholds for evidence:

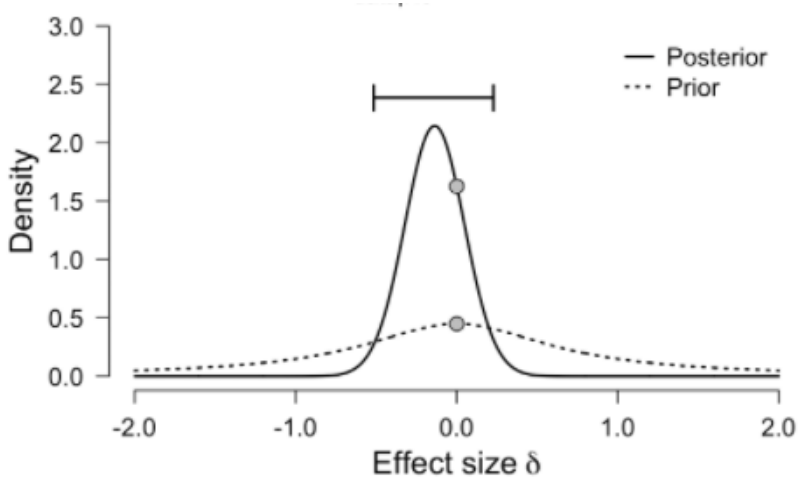
Bayes factor	Evidence
1-3	anecdotal
3-10	moderate
10-30	strong
30-100	very strong
$>$100	extreme

www.jasp-stats.org

Savage-Dickey Density Ratio

When $\mathcal{H}_0$ is a "point null hypothesis", B_{01} can be computed by comparing the density of the point null $\delta = 0$ in the posterior distribution compared to the density of $\delta = 0$ in the prior[12]

$$B_{01} = \frac{p(\delta = 0 \mid \text{data}, \mathcal{H}_1)}{p(\delta = 0 \mid \mathcal{H}_1)}.$$



[1]Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology, 60*, 158-189

[2]Faulkenberry, T. J. (2019). A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors. *Communications for Statistical Applications and Methods, 26*(2), 1-22

Contact info

- Thomas J. Faulkenberry
- Department of Psychological Sciences
- Tarleton State University
- faulkenberry@tarleton.edu
- Twitter: @tomfaulkenberry