Bayes Factor Null Hypothesis Tests are still Null Hypothesis Tests

Matt N. Williams
School of Psychology
Massey University, New Zealand

Presentation at Modern Modeling Methods Conference,
Storrs CT, 24-25 May 2016

MASSEY UNIVERSITY

Context part I: Default statistical methods

- “Statistics: The science of defaults” (Gelman, 2014)
- Many presentations in this conference focus on complex methods, conducted by skilled analysts or statisticians
- This is wonderful! But it’s also important to look at the simpler methods used by everyday researchers
- Crucial topic in psychology given the recent replication crisis - mainstream research practices seem to often produce incorrect or overconfident conclusions (even to very simple research questions).

Context part II: Dissatisfaction with NHST

In psychology, as in many fields, there is both growing unease with the ubiquitous use of null hypothesis significance tests (NHST)

Problems are well established, but include:

- NHST tells us P(Data|Model) when we want to know P(Model|Data)
- Usually takes the form of testing a null (or "nil") hypothesis that a parameter (effect or relationship) is *exactly* zero - may be implausible?
- Asymmetry of result: Significant *p* value taken to establish truth of alternate hypothesis, but a non-significant value often taken to imply only uncertainty ("insufficient evidence to reject null")
    - May contribute to selective reporting of results

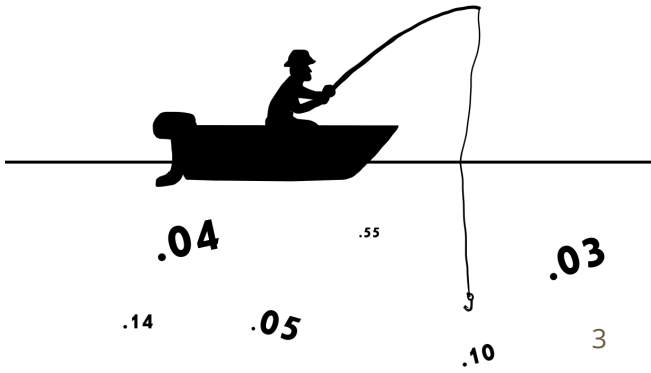


Image by Ed Witt http://kantorwont.com/statistics-comics/

A solution? Bayes factor null hypothesis tests

- Bayes factor developed by Jeffreys in 1925; a statistic for comparing the probability of a set of observed data under two models
- More recently applied to the problem of null hypothesis testing
- Grew to fame in psychology thanks in part to a paper by Wagenmakers et al. (2011) re-analysing the infamous Bem (2011) study using Bayes factors
- Implemented in several online apps and in the easy-to-use SPSS alternative JASP (JASP Team, 2016).

$$BF = \frac{P(D|M_1)}{P(D|M_2)} \; e.g., \; \frac{P(D|H_1)}{P(D|H_0)}$$

Bayes factor null hypothesis tests

- In their application as alternatives to NHST, Bayes factor tests generally take a specific form:
- Null hypothesis H0: Parameter is exactly zero (i.e., H0 specifies a prior in which all prior mass is on a single point)
- Alternate hypothesis H1: Parameter is not exactly zero (i.e., H1 specifies a **prior** that is spread over a range of values)
- The difficulty is in choosing what prior to use for the alternate hypothesis
  - Sidenote: I find the concept of a prior *within* a hypothesis somewhat confusing; I find it helps to think of the priors for the alternate hypothesis in a Bayes Factor test as a "conditional prior": It is the prior probability we would place on different parameter values, if we knew H1 was true.

Bayes factor null hypothesis tests: Specific forms

- Recent innovations in the area have largely taken the form of developing suggested priors (for the alternate hypothesis, and for nuisance parameters) to produce Bayes factor alternatives to several common null hypothesis significance tests, for example:
  - *t* test (Rouder et al., 2009)
  - correlation (Wetzels & Wagenmakers, 2012)
  - ANOVA (Rouder et al., 2012)

Apparent intent is that these tests can be used as a generic default replacement for common significance tests (e.g., Wetzels et al., 2011 - re-analysed 855 recently reported *t* tests in psychology using Bayes factors)

Bayesian *t* test

Bayesian *t* test (Rouder et al., 2009) is a good example of a Bayes factor null hypothesis test:

- Parameterised with respect to standardised effect size $\delta = (\mu_1 - \mu_2)/\sigma$
- H0: $\delta = 0$
- H1: $\delta \sim$ Cauchy (0, 1)
  - Technically accomplished by setting prior as $N(0, \sigma^2_p)$ and hyperprior $\sigma^2_p \sim$ inverse $\chi^2(1)$
- In either case, Jeffrey's prior on variance $P(\sigma^2) = 1/\sigma^2$

On not stomping flowers...

"In a desert of incoherent frequentist testing there blooms a Bayesian flower. You may not think it is a perfect flower. Its color may not appeal to you, and it may even have a thorn. But it is a flower, in the middle of a desert. Instead of critiquing the color of the flower, or the prickliness of its thorn, you might consider planting your own flower — with a different color, and perhaps without the thorn. Then everybody can benefit." -EJ Wagenmakers, datacolada.org/35

Image in public domain, via Wikimedia Commons

On not stomping flowers...

- Bayes factor null hypothesis testing definitely has advantages over frequentist NHST:
  - Considers likelihood under both H0 and H1 (not just H0)
  - Can provide evidence *for* (and not just against) H0
  - More intuitive to interpret
- Great to see methodologists building accessible tools like JASP for Bayesian analysis
- Psychologists and other social scientists *do* need methods for testing hypotheses - not just for estimating effects and relationships
- My concern primarily just with the idea of using a Bayes factor test *of a point null hypothesis* as a default approach.

The problem: Lack of posterior

- The Bayes factor itself tells us only about the probability of the data under each hypothesis- *not* how certain we can be that a particular hypothesis is correct
  - I.e., like frequentist analyses, it focuses only on P(D | H) under each hypothesis and not P (H | D)
  - No posterior probability distribution - because it doesn't take into account the probability that each hypothesis is correct.

"Researchers should report the Bayes factor, and readers can update their own priors accordingly... Sophisticated researchers may add guidance and value to their analysis by suggesting prior odds, or ranges of prior odds."
-(Rouder et al., 2012, p. 359).

- Ok, but when we are talking about everyday (non-statistician) researchers and readers, is it realistic to expect them to do this?

Interpreting Bayes factors

- I suspect users of these tests *will* interpret Bayes factors as posterior statements about which model is more plausible.
- This interpretation is supported by popular qualitative guidelines for interpreting Bayes factors (e.g. below from Wetzels et al., 2011, as based on Jeffreys, 1961).
- These seem to imply to users that the Bayes factor itself is the "endpoint" for drawing conclusions.

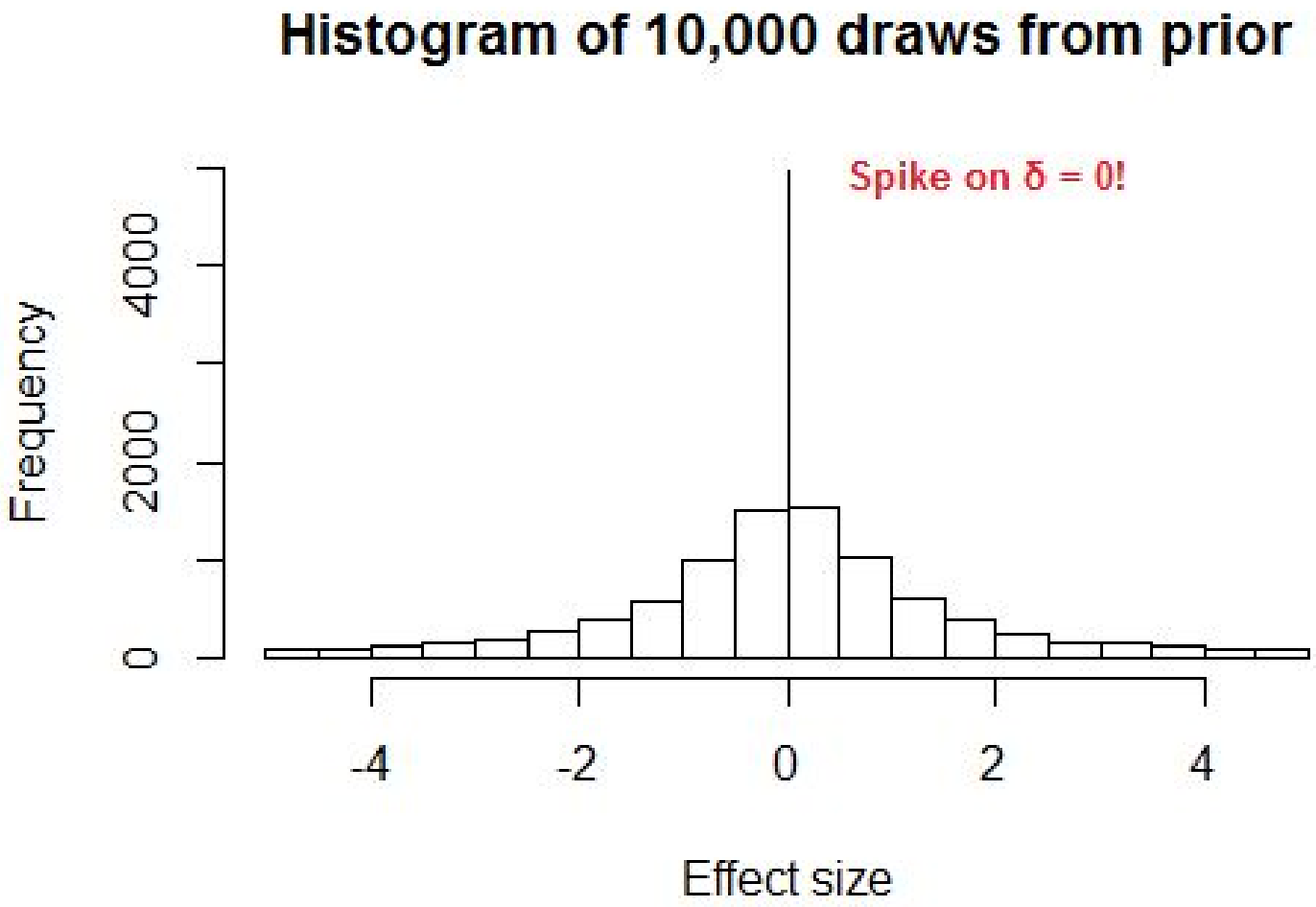
Bayes factor	
>100	Decisive evidence for H_A
30–100	Very strong evidence for H_A
10–30	Strong evidence for H_A
3–10	Substantial evidence for H_A
1–3	Anecdotal evidence for H_A
1	No evidence
1/3–1	Anecdotal evidence for H_0
1/10–1/3	Substantial evidence for H_0
1/30–1/10	Strong evidence for H_0
1/100–1/30	Very strong evidence for H_0
<1/100	Decisive evidence for H_0

Obtaining a posterior

So... Under what conditions **can** a Bayes factor test be interpreted as a statistical model producing the posterior odds that H1 is correct?

A Bayes factor can be interpreted as the posterior odds if our prior is that P(H0) = P(H1) = 0.5

The implied prior (for Bayesian *t* test)



Note. Effect sizes outside [-5, 5] trimmed out.

Spike and slab?

Referred to this prior as a “spike and slab” in my abstract, but maybe better described as “tower on hill!”

Image: In public domain, via Wikimedia Commons

Is this a sensible default?

- If we interpret a Bayes Factor as a direct statement about the posterior odds for the two hypotheses, the implied prior is *highly* informative
- The prior suggests that one particular value of the effect size ($\delta = 0$) is **vastly** more likely than any other value
  - I.e., still a very strong focus on a null hypothesis, which is held to be especially plausible
- Is such an informative prior suitable as a general default? Is it likely to accurately represent our prior knowledge, or what we really expect in psychology?

Why fight the null strawman?

- Existing Bayes factor null hypothesis tests don't take into account an important fact: *Most* of the time, in social science research the hypothesis of interest to the researcher will be that an effect falls in a particular *direction*
  - with no reason whatsoever to expect that the effect is *exactly* zero in size
  - In such a situation, the null hypothesis is just a **strawman** to be rejected.

If we want to compare a directional research hypothesis against a plausible competitor, why not compare it to an hypothesis that **the effect is in the other direction**?

So what are the alternatives?

My suggestion:

1. **Test *directional* hypotheses** (no point null, unless there is a specific reason to consider an exactly zero effect to be especially plausible)

2. **Use an *informative* prior variance for the effect size**

Most effects in psychology and other social sciences are small - if we fail to take this information into account (e.g., by using non-informative priors), our conclusions about directionality will be overconfident.

Priors - some rough ideas

- So we need procedures that use priors that (even if roughly) take into account the fact that we know most effects are fairly small.
- Where can we get these from?
- The prior distributions specified for alternate hypotheses in the various Bayes factor null hypothesis tests might provide a good starting point with respect to prior shape (e.g., Cauchy on effect sizes)
- For prior variance, we could use *empirical* information about the size of effects in various fields.
- E.g. in psychology we could use Richard et al.'s (2003) meta-meta-analysis of 25k social psychological studies and 8m people: Found mean $|r|$ of 0.21, which equates to a Cohen's *d* of ~0.43.
    - So for mean differences/binary predictors, could set a Cauchy (0, 0.43) on δ?

Computation

Can achieve posterior probability statements about directional hypotheses simply using Bayesian estimation with informative priors

- Focus on proportion of draws from the posterior with estimated effect size > 0
- I.e., posterior probability that effect is positive

But could also using Bayes factor tests - just with the proviso that the default comparison is *directional*

- *Not* a comparison of a point null against alternate H1
- Then Bayes factor becomes posterior odds, if our prior is that an effect in either direction is equally likely.)

References

Bem DJ (2011) Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. J Pers Soc Psychol 100:407–425.

Gelman A (2014) How do we choose our default methods?, in: Lin X, Genest C, Banks DL, Molenberghs G, Scott DW, Wang J-L (Eds.), Past, Present, and Future of Statistical Science pp. 293–301.

JASP Team (2016). JASP (Version 0.7.5.5)

Jeffreys H (1935) Some tests of significance, treated by the theory of probability. Math Proc Cambridge Philos Soc 31:203–222.

Richard FD, Bond CF Jr, Stokes-Zoota JJ (2003). One hundred years of social psychology quantitatively described. Rev Gen Psychol 7:331-363

Rouder JN, Morey RD, Speckman PL, Province JM (2012) Default Bayes factors for ANOVA designs. J Math Psychol 56:356–374.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 16:225.

Wagenmakers EJ, Wetzels R, Borsboom D, Van der Maas H (2011) Why psychologists must change the way they analyze their data: The case of psi. J Pers Soc Psychol 100:426–432.

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers E-J (2011) Statistical evidence in experimental psychology an empirical comparison using 855 t tests. Perspect Psychol Sci 6:291–298.

Wetzels R, Wagenmakers E-J (2012) A default Bayesian hypothesis test for correlations and partial correlations. Psychon Bull Rev 19:1057–1064.

Comments or questions?

Ask now, or:

m.n.williams@massey.ac.nz

twitter.com/matthewmatix