The Use of Rejection Odds and Rejection Ratios in Testing Hypotheses

Jim Berger

Duke University

with M.J. Bayarri, Daniel J. Benjamin (University of Southern California, and Thomas M. Sellke (Purdue University)

ASA Symposium on Statistical Inference

Bethesda, MD

October 11, 2017

Outline

- Discussion of the current problems with using p-values for hypothesis testing.
- Review of why Bayesian hypothesis testing and the common usage of p-values are incompatible.
- Review of why Bayesian hypothesis testing and fixed error probability frequentist testing are incompatible.
- A mathematical bound that can be used to quickly convert p-values into Bayes factors.
- Why use of rejection odds provides a frequentist justification for Bayes factors.

The sad state of statistical testing in science

- Significance testing of a null hypothesis, H_0, using p-values is by far the dominant method of testing in science.
- It is a major cause of the lack of reproducibility of science.
- Everyone is talking about it:
  - articles in all the major science journals;
  - changes in editorial policy (the journal *Basic and Applied Social Psychology* banned p-values);
  - the recent ASA position statement about p-values and discussion;
  - article that just appeared in *Nature Human Behaviour* with 72 authors recommending 'significance' should be at $p = 0.005$, rather than $p = 0.05$.
- The problem is that p-values are typically misinterpreted.
  - $p = 0.05$ is often interpreted as 1 to 20 odds against H_0, when it is really no more than 1 to 2.5 odds.
  - $p = 0.005$ is often interpreted as 1 to 200 odds against H_0, when it is really no more than 1 to 14 odds.

Incompatibility of Bayesian testing with p-values and fixed error probability frequentist testing

To test: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ based on data $\boldsymbol{x} \sim f(\boldsymbol{x} \mid \theta)$.

- p-value, for test statistic $T(\boldsymbol{x})$, is $P_0(T(\boldsymbol{X}) > T(\boldsymbol{x}_{obs}))$.
- Bayes factor (or odds) of H_1 to H_0, for prior $\pi(\theta)$ under H_1,

$$B_{10}(\boldsymbol{x}) = \frac{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)d\theta}{f(\boldsymbol{x} \mid 0)} ,$$

  is much smaller than $1/p$ (e.g., $B_{10} = 2.5$ when $1/p = 1/[0.05] = 20$).

- Fixed α-level frequentist testing chooses a rejection region $\mathcal{R}$ and computes Type I error probability $\alpha = P_0(\mathcal{R})$, reporting error α *no matter where the data is in* $\mathcal{R}$.
  - This seems wrong to a Bayesian, e.g. reporting the same $\alpha = 0.05$
    - when $p = 0.05$ (where 0.05 is a serious underestimate of the error)
    - or $p = 0.00001$ (where 0.05 is a serious overestimate of the error).

A common complaint: determining Bayes factors is too hard.

But p-values can be converted into bounds on Bayes factors.

Indeed, *robust Bayesian theory* suggests general and simple ways to calibrate p-values. (Vovk, 1993, Sellke, Bayarri and Berger, 2001).

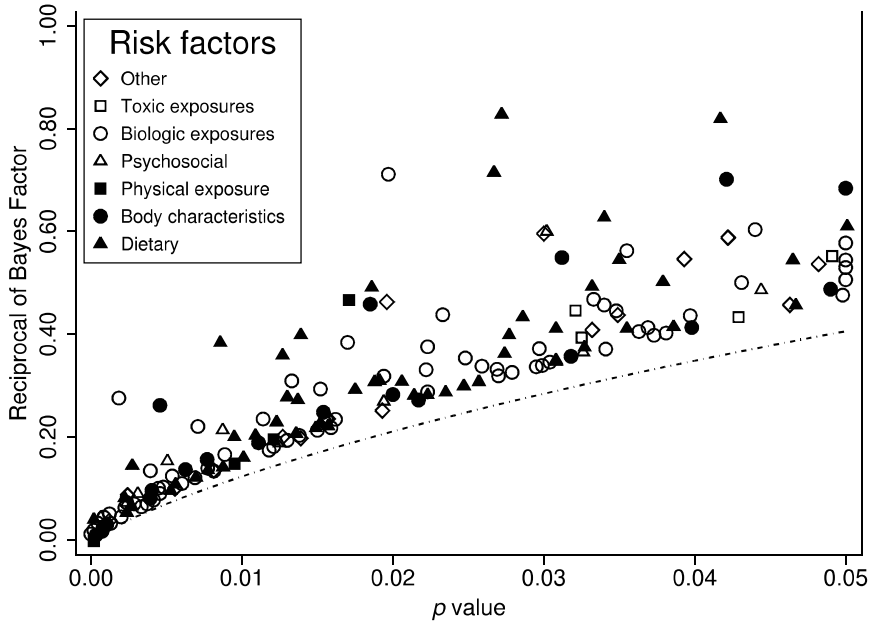
Theorem 1 *A proper p-value satisfies $H_0 : p(X) \sim Uniform(0,1)$, so consider testing this versus $H_1 : p \sim g(p)$, where $Y = -\log(p)$ has a non-increasing failure rate (a natural non-parametric condition on g). Then*

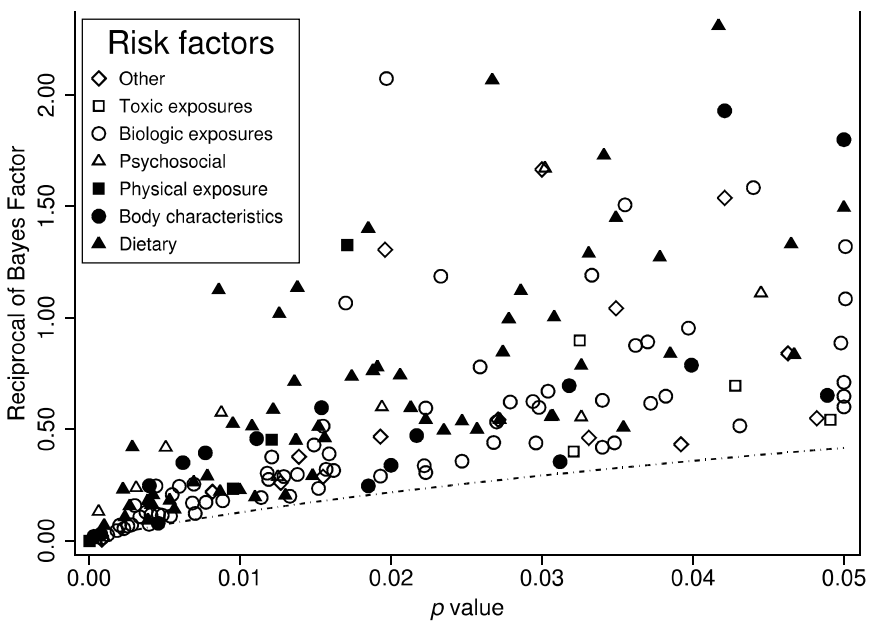
$$B_{10} \leq \frac{1}{-e\, p\, \log(p)} \quad \textit{for} \quad p < e^{-1}\,.$$

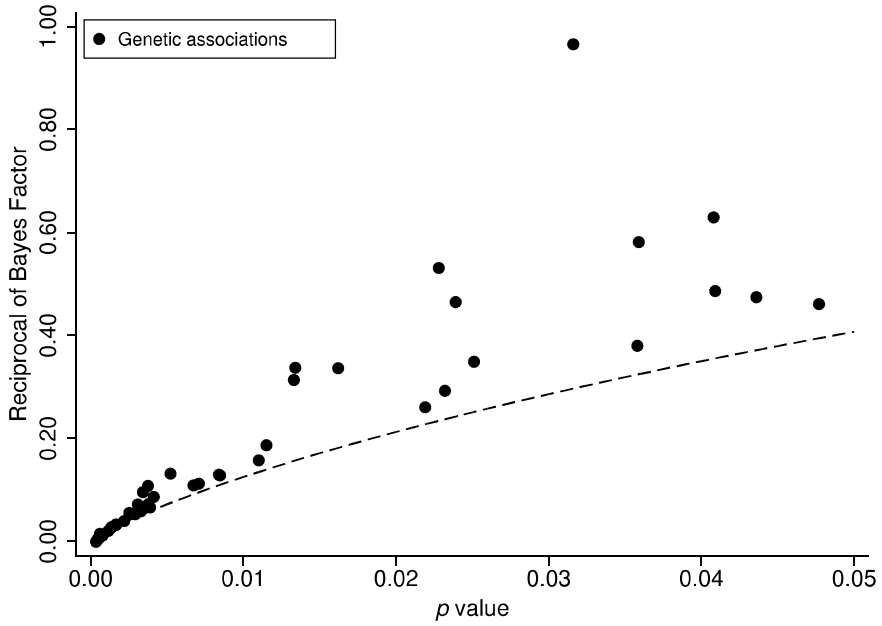
p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001	5×10^{-7}
$\frac{1}{-ep\log(p)}$	1.60	2.44	8.13	13.9	52.9	400	3226	2.0×10^5

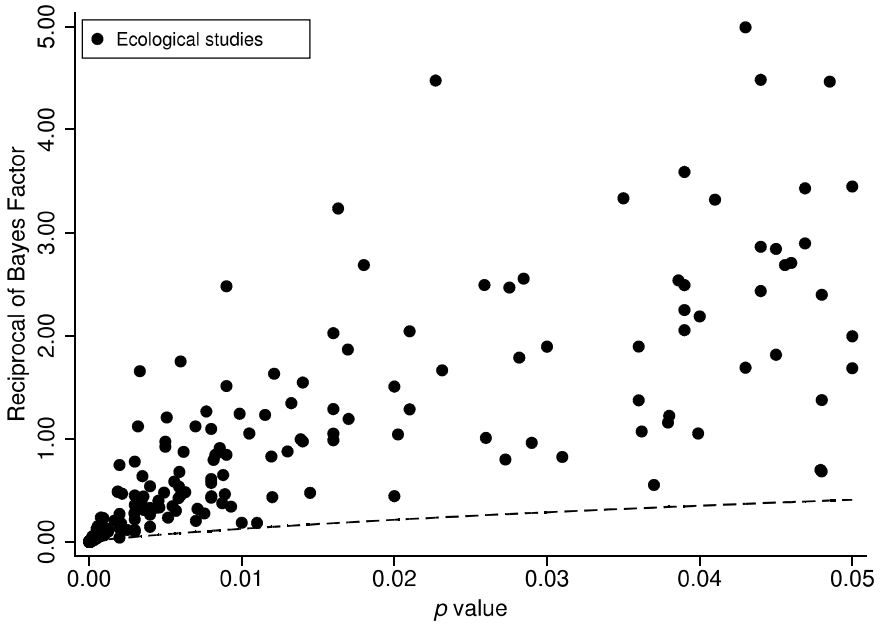
- Although very simple, there was initially concern that the $\frac{1}{-ep\log(p)}$ bound is too large, since it is known that Bayes factors can depend strongly on the sample size n, and the bounds do not.
- But the following studies indicate that this might not typically be a problem. These studies
  - look at large collections of published studies where $0 < p < 0.05$;
  - compute a Bayes factor, $B_{01} = 1/B_{10}$, for each study;
  - graph the Bayes factors versus the corresponding p-values.
- The lower boundary in all figures is essentially the lower bound $-e\,p\log(p)$ (the corresponding bound for $B_{01} = 1/B_{10}$ and given by the dashed lines in the figures), indicating that it is often an accurate bound.

The first two graphs are for 272 'significant' epidemiological studies with two different choices of the prior; the third for 50 'significant' meta-analyses (these three from J.P. Ioannides, Am J Epidemiology, 2008); and the last is for 314 ecological studies (reported in Elgersma and Green, 2011).


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.50
1.00
1.50
2.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Genetic associations
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Ecological studies
Reciprocal of Bayes Factor
0.00
1.00
2.00
3.00
4.00
5.00
0.00
0.01
0.02
0.03
0.04
0.05
p value

Bayes factors have a frequentist justification, through the notion of rejection odds

Setup (for now a mix of frequentist and Bayes):

We observe data $\boldsymbol{x}$ from the density $f(\boldsymbol{x} \mid \theta)$ and wish to test

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0 \quad .$$

- Suppose a rejection region $\mathcal{R}$ is specified.
- Let $\alpha = Pr(\mathcal{R} \mid 0)$ and $(1 - \beta(\theta)) = Pr(\mathcal{R} \mid \theta)$ be the Type I error and power corresponding to the rejection region $\mathcal{R}$.
- Let π_0 and $\pi_1 = 1 - \pi_0$ be the prior probabilities of H_0 and H_1.
- Let $\pi(\theta)$ be the prior density of θ under H_1 (this could just be a point mass at a point θ' for which power is to be evaluated).
  - Then $(1 - \bar{\beta}) = \int (1 - \beta(\theta))\pi(\theta)d\theta$ is the average power wrt the prior $\pi(\theta)$ (equals $[1 - \beta(\theta')]$ if power at a point is used).
  - And $m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\pi(\theta)d\theta$ is the marginal likelihood of the data $\boldsymbol{x}$ for the prior $\pi(\theta)$ under H_1 (equals $f(\boldsymbol{x} \mid \theta')$ for a point mass prior).

Pre-experimental analysis (not new):

The pre-experimental probability of incorrectly rejecting H_0 is then $\pi_0\alpha$, while the pre-experimental probability of correctly rejecting H_0 is $\pi_1(1-\bar{\beta})$.

Definition: The *pre-experimental odds of correct to incorrect rejection of* H_0 are

$$\begin{aligned} O_{pre} &= \frac{\pi_1}{\pi_0} \times \frac{(1-\bar{\beta})}{\alpha} \\ &\equiv O_P \times R_{pre} \\ &\equiv [\text{prior odds of } H_1 \text{ to } H_0] \times [\textit{rejection odds} \text{ of } H_1 \text{ to } H_0]\,. \end{aligned}$$

Reporting of the rejection odds, R_{pre}, recognizes the crucial role of power in understanding the strength of evidence in rejecting, and does so in a simple way (reducing the evidence to a single number).

average power	0.05	0.25	0.50	0.75	1.0	0.01	0.25	0.50	0.75	1.0
type I error	0.05	0.05	0.05	0.05	0.05	0.01	0.01	0.01	0.01	0.01
R_{pre}	1	5	10	15	20	1	25	50	75	100

Example: Genome-wide Association Studies (GWAS)

- Early genomic epidemiological studies almost universally failed to replicate (estimates of the replication rate are as low as 1%), because they were doing extreme multiple testing at non-extreme p-values.
- A very influential paper in Nature (2007) by the Wellcome Trust Case Control Consortium proposed the cutoff $p < 5 \times 10^{-7}$.
  - Found 21 genome/disease associations; 20 have been replicated.
- The frequentist Bayesian argument for the cutoff:
  - They wanted an experiment with O_{pre}, the pre-experimental odds of a true to false positive, equal to $10:1$.
  - They assessed O_P, the prior odds of a true to false positive, to be $\frac{1}{100{,}000}$. (This is their implementation of Bayesian control for multiple testing; O_P could, instead, have been estimated from the data.)
  - Typical GWAS studies had power $(1-\bar{\beta}) = 0.5$.
  - Solving $[\frac{10}{1} = \frac{1}{100{,}000} \times \frac{0.5}{\alpha}]$ gave $\alpha = 5 \times 10^{-7}$.

Post-experimental odds analysis (not new):

Once the data is at hand a Bayesian would focus on the posterior odds of H_1 to H_0 given by

$$\begin{aligned} O_{post} &= \frac{\pi_1}{\pi_0} \times \frac{m(\boldsymbol{x})}{f(\boldsymbol{x} \mid 0)} \\ &\equiv O_P \times R_{post}(\boldsymbol{x}) \,, \end{aligned}$$

where $R_{post}(\boldsymbol{x})$ is the data-dependent odds of a true to false rejection, more commonly called the Bayes factor of H_1 to H_0 and denoted $B_{10}(\boldsymbol{x})$.

GWAS example: Parts of the Nature article argued that it is best to just compute the Bayes factors, $B_{10}(\boldsymbol{x})$, and the posterior odds O_{post}.

For the 21 claimed associations, these ranged between

- $O_{post} = 10^{68}$ (overwhelming evidence of a correct rejection) and
- $O_{post} = \frac{1}{10}$ (evidence of an *incorrect* rejection; note that this is the one claimed association in the article that has *not* been replicated).

Reporting these these seems much more reasonable than always saying $O_{pre} = \frac{10}{1}$, but the article did not base decisions on them, presumably because they are not frequentist measures. Is that true?

Lemma (new): *The frequentist expectation of $B_{10}(\boldsymbol{x})$, over the rejection region and under H_0, is*

$$E[B_{10}(\boldsymbol{x}) \mid H_0, \mathcal{R}] = R_{pre} = \frac{(1-\bar{\beta})}{\alpha}\,.$$

This guarantees that, under H_0, the "average of the reported Bayes factors when rejecting" equals the actual rejection odds R_{pre}, so $B_{10}(\boldsymbol{x})$ *is as valid a frequentist report as is* R_{pre}.

How can a valid frequentist procedure depend on a prior distribution?

- Any power assessment requires at least specification of a point at which to assess power, and that can be used as the prior if nothing else is available.
- Thus, if one is willing to consider power, then R_{post} is much better than R_{pre}, since it has the same frequentist justification and is fully data dependent.

Conclusion: Saving the world from misuse of p-values

- We need to agree that the direct use of p-values as confirmatory evidence should stop (the ASA statement more or less says this); the historical evidence is clear that p-values cannot be properly interpreted by most users.
- The ideal replacement for p-values would be the posterior odds of H_1 to H_0:

$$O_{post} = O_P \times B_{10}(\boldsymbol{x}) \quad (\text{superior to } O_{pre} = O_P \times \tfrac{1-\bar{\beta}}{\alpha}) \,,$$

  where O_P is the prior odds and $B_{10}(\boldsymbol{x})$ is the Bayes factor of H_1 to H_0.
- The Bayes factor can be the only report if use of prior odds is problematical, and this report *has as much frequentist justification as does reporting the pre-experimental rejection odds* $(1-\bar{\beta})/\alpha$.
- If determination of $B_{10}(\boldsymbol{x})$ is not feasible, report the upper bound on the Bayes factor, $1/[-ep\log p]$; this is much less likely to be misinterpreted than p.

Thanks!