

# 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)

*B.G. Greenberg Distinguished Lectures*

*Department of Biostatistics*

*University of North Carolina at Chapel Hill*

*May 12, 2016*

## The sad state of statistical testing in science

- Significance testing using  $p$ -values is by far the dominant method of testing in science.
- Its standard uncritical use is viewed by many as being the major source of the problems 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.
- My view: none of the discussion matters unless, as a profession, we can agree on an alternative to  $p$ -values.

## The $p$ -value Controversy

- Concerns with use of  $p$ -values trace back to at least Berkson (1937).
- Concerns are recurring in many scientific literatures:  
Environmental sciences: <http://www.indiana.edu/~stigtsts/>  
Social sciences: <http://acs.tamu.edu/~bbt6147/>  
Wildlife science: <http://www.npwrc.usgs.gov/perm/hypotest/>  
<http://www.cnr.colostate.edu/~anderson/null.html>
- An example - articles and books in social sciences: Rozeboom, 60; Edwards, Lindman, and Savage, 63; Morrison and Henkel, 70; Carver, 78; Meehl, 78; Shaver, 85; Oakes, 86; Kupersmid, 88; Rosnow and Rosenthal, 89; Cohen, 90, 94; Rosenthal, 91; Thompson, 93, 94, 98, 99; Volume 51 (No.4) of J. of Experimental Education; Schmidt, 96 (APA presidential address); Hunter, 97; Schmidt and Hunter, 97; Harlow, Mulaik and Steiger, 97; Levin, 98; APA task force
- Numerous works specifically focus on comparing the Fisher and N-P approaches (e.g., Lehmann, 1993 JASA: The Fisher, N-P Theories of Testing Hypotheses: One Theory or Two?)

## The Major Problem: $p$ -values are misinterpreted

- Few non-statisticians understand  $p$ -values, most erroneously thinking they are some type of error probability, Bayesian or frequentist; they are neither!
  - A survey 30 years ago:
    - \* “What would you conclude if a properly conducted, randomized clinical trial of a treatment was reported to have resulted in a beneficial response ( $p < 0.05$ )?”
      1. Having obtained the observed response, the chances are less than 5% that the therapy is not effective.
      2. The chances are less than 5% of not having obtained the observed response if the therapy is effective.
      3. The chances are less than 5% of having obtained the observed response if the therapy is not effective.
      4. None of the above.
    - \* We asked this question of 24 physicians ... Half ... answered incorrectly, and all had difficulty distinguishing the subtle differences...
    - \* The correct answer to our test question, then, is 3.”

“This isn’t right. This isn’t even wrong.” –Wolfgang Pauli, on a submitted paper

- \* **Actual correct answer:** The chances are less than 5% of having obtained the observed response *or any more extreme response* if the therapy is not effective.
- But, is it fair to count ‘possible data more extreme than the actual data’ in the evidence against the null hypothesis?  
Jeffreys (1961): “An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.”
- Matthews (1998): “The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding.”
- When testing precise hypotheses, true error probabilities (Bayesian or frequentist) are much larger than  $p$ -values.
  - Later examples.
  - *Applet* (of German Molina) available at [www.stat.duke.edu/~berger](http://www.stat.duke.edu/~berger)

## Replacing $p$ -values and fixed error probability frequentist testing with unified Bayesian/frequentist rejection odds

- 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.
- Why use of rejection odds can unify Bayesian and frequentist testing.



## Bayesian hypothesis testing and the common usage of $p$ -values are incompatible

# San Jose Mercury News

mercurynews.com WEST VALLEY 102

Friday, September 25, 2009

THE NEWSPAPER OF SILICON VALLEY 75 cents

### AIDS MILESTONE

# New path for HIV vaccine

Some in study protected from infection, but trial raises more questions

By Karen Kaplan  
and Thomas H. Maugh II  
*Los Angeles Times*

Hours after HIV researchers announced the achievement of a milestone that had eluded them for a quarter of a century, reality began

to set in: Tangible progress could take another decade.

A Thai and American team announced early Thursday in Bangkok that they had found a combination of vaccines providing modest protection against infection with the virus that causes AIDS, unleashing excitement worldwide. The idea of a vaccine to prevent infection with the human immunodeficiency virus, HIV, had long been

frustrating and fruitless.

But by Thursday afternoon, initial euphoria gave way to a more sober assessment. There is still a very long way to go before reaching the goal of producing a vaccine that reliably shields people from HIV.

Some researchers questioned whether the apparent 31 percent reduction in infections was a sta-

See **VACCINE**, Page 14



A researcher during the Thai phase III HIV Vaccine Trial, also known as RV144, tests the "prime-boost" combination of two vaccines.

ASSOCIATED PRESS

## Hypotheses and data:

- Alvac had shown no effect
- Aidsvax had shown no effect

*Question:* Would Alvac as a primer and Aidsvax as a booster work?

*The Study:* Conducted in Thailand with 16,395 individuals from the general (not high-risk) population:

- 74 HIV cases reported in the 8198 individuals receiving placebos
- 51 HIV cases reported in the 8197 individuals receiving the treatment



## The test that was performed:

- Let  $p_1$  and  $p_2$  denote the probability of HIV infection in the placebo and treatment populations, respectively.
- Test  $H_0 : p_1 = p_2$  versus  $H_1 : p_1 > p_2$
- Normal approximation okay, so

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{\sigma}_{\{\hat{p}_1 - \hat{p}_2\}}}} = \frac{.009027 - .006222}{.001359} = 2.06$$

is approximately  $N(\theta, 1)$ , where  $\theta = (p_1 - p_2)/(.001359)$ .

Test  $H_0 : \theta = 0$  versus  $H_1 : \theta > 0$ , based on  $z$ .

- Observed  $z = 2.06$ , so the  $p$ -value is 0.02.

## Bayesian analysis:

Posterior odds of  $H_1$  to  $H_0 = [\text{Prior odds of } H_1 \text{ to } H_0] \times B_{10}(z)$ ,

where

$$\begin{aligned}
 B_{10}(z) &= \text{Bayes factor of } H_1 \text{ to } H_0 = \text{'data-based odds of } H_1 \text{ to } H_0\text{'} \\
 &= \frac{\text{average likelihood of } H_1}{\text{likelihood of } H_0 \text{ for observed data}} = \frac{\int \frac{1}{\sqrt{2\pi}} e^{-(z-\theta)^2/2} \pi(\theta) d\theta}{\frac{1}{\sqrt{2\pi}} e^{-(z-0)^2/2}},
 \end{aligned}$$

For  $z = 2.06$  and  $\pi(\theta) = \text{Uniform}(0, 2.95)$ , the nonincreasing prior *most favorable* to  $H_1$ ,

$$B_{10}(2.06) = 5.63 \quad (\text{the p-value is } 0.020, \text{ commonly misinterpreted as } 50 : 1 \text{ odds})$$

(The actual subjective 'study team' prior yielded  $B_{10}^*(2.06) = 4.0$ , and the maximum Bayes factor over all possible priors is  $e^{z^2/2} = 8.35$ .)

## Bayesian hypothesis testing and fixed $\alpha$ level testing are not compatible

- Fixed  $\alpha$  level testing, i.e.
  - pre-experimentally choosing a rejection region  $\mathcal{R}$  (e.g.,  $\mathcal{R} = (1.645, \infty)$  in the vaccine example) with Type I error probability  $\alpha = Pr(\mathcal{R} | H_0)$ ,
  - and reporting the error as  $\alpha$ , *no matter where the data is in  $\mathcal{R}$* , is a valid frequentist procedure (as opposed to use of the  $p$ -value as the error probability, which is not a valid frequentist procedure).
- But it is an unconditional frequentist procedure and conditionally seems silly to a Bayesian, e.g. reporting the same  $\alpha = 0.05$ 
  - when  $z = 1.645$  (where 0.05 is a serious underestimate of the actual conditional error)
  - or  $z = 5$  (where 0.05 is a serious overestimate of the actual conditional error).

## Possible unification of frequentist and Bayesian testing

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

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

$H_0 : \theta = \theta_0$  (or  $H_0 : |\theta - \theta_0| < \epsilon$ ) versus  $H_1 : \theta \neq \theta_0$  (or  $H_1 : |\theta - \theta_0| > \epsilon$ ).

- Suppose a rejection region  $\mathcal{R}$  is specified.
- Let  $\alpha = Pr(\mathcal{R} | \theta_0)$  and  $(1 - \beta(\theta)) = Pr(\mathcal{R} | \theta)$  be the Type I error and power corresponding to the rejection region  $\mathcal{R}$ .
- Let  $\pi_0$  and  $\pi_1 = 1 - \pi_0$  be the prior probabilities of  $H_0$  and  $H_1$ .
- Let  $\pi(\theta)$  be the prior density of  $\theta$  under  $H_1$  (this could just be a point mass at a point  $\theta'$  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  $m(\mathbf{x}) = \int f(\mathbf{x} | \theta)\pi(\theta)d\theta$  is the marginal likelihood of the data  $\mathbf{x}$  under the prior  $\pi(\theta)$  (equals  $f(\mathbf{x} | \theta')$  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

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

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 for  $\alpha$  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

$$\begin{aligned} O_{post} &= \frac{\pi_1}{\pi_0} \times \frac{m(\mathbf{x})}{f(\mathbf{x} | \theta_0)} \\ &\equiv O_P \times R_{post}(\mathbf{x}), \end{aligned}$$

where  $R_{post}(\mathbf{x})$  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  $B_{10}(\mathbf{x})$ .

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

For the 21 claimed associations, these ranged between

- $O_{post} = \frac{1}{10}$  (moderate evidence of an incorrect rejection) and
- $O_{post} = 10^{68}$  (overwhelming evidence of a correct rejection).

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

**Lemma** (new): *The frequentist expectations of  $B_{10}(\mathbf{x})$  and  $B_{01}(\mathbf{x}) = 1/B_{10}(\mathbf{x})$  over the rejection region, conditional on the respective hypotheses, are*

$$E[B_{10}(\mathbf{x}) \mid H_0, \mathcal{R}] = R_{pre} = \frac{(1 - \bar{\beta})}{\alpha}, \quad \text{and} \quad E[B_{01}(\mathbf{x}) \mid H_1^*, \mathcal{R}] = [R_{pre}]^{-1},$$

where  $H_1^*$  refers to the marginal alternative model with density  $m(\mathbf{x})$ .

The first identity guarantees that, under  $H_0$ , the “average of the reported Bayes factors when rejecting” equals the actual rejection odds  $R_{pre}$ , so  $B_{10}(\mathbf{x})$  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.

## Vaccine Example:

There were widely varying opinions concerning the prior odds  $O_p = \pi_1/\pi_0$ , and the impact of this was illustrated and discussed in Gilbert et al., 2011.

But we will focus on  $O_{pre}$  and  $O_{post}(z) = B_{10}(z)$ , the odds arising from the experiment and data.

The observation is  $z = 2.06$ , where  $Z$  is approximately  $N(\theta, 1)$ , and we are testing

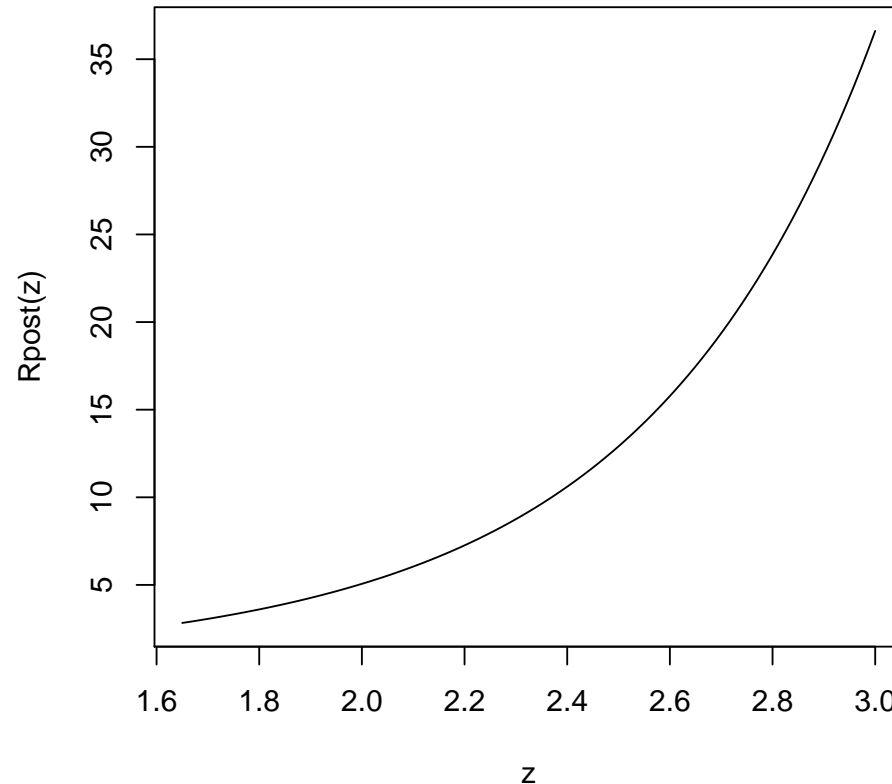
$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0.$$

The  $p$ -value is 0.02 and

$$O_{post}(z) = B_{10}(z) = \text{Bayes factor of } H_1 \text{ to } H_0 = \frac{\int \frac{1}{\sqrt{2\pi}} e^{-(z-\theta)^2/2} \pi(\theta) d\theta}{\frac{1}{\sqrt{2\pi}} e^{-(z-0)^2/2}},$$

equaling  $B_{10}(2.06) = 5.63$  for the actual data and using  $\pi(\theta) = \text{Uniform}(0, 2.95)$ , the nonincreasing prior *most favorable* to  $H_1$ .

**Odds Analysis:**  $\mathcal{R} = (1.645, \infty)$ ,  $\alpha = 0.05$  and the scientists said  $1 - \bar{\beta} = 0.45$ , so  $R_{pre} = (1 - \bar{\beta})/\alpha = 9$ . Thus, pre-experimentally, the rejection odds of a correct rejection to an incorrect rejection are nine to one.



Above is  $B_{10}(z)$  (or  $R_{post}(z)$ ), as a function of  $z \in \mathcal{R}$ . Thus, post-experimentally the odds of a correct rejection to incorrect rejection can be anywhere from 2 to  $\infty$ .

Recall  $z = 2.06$  was observed, so  $B_{10}(2.06) = 5.63$  is the actual odds.

**Another argument for  $B_{10}(\mathbf{x})$  as opposed to  $R_{pre}$ :** (The basic issue pointed out to us by Deborah Mayo)

Pre-experimentally, the larger the power, the larger is  $R_{pre}$ , so the larger the probability that a rejection from the experiment will be correct, *assuming, of course, that the power has been accurately assessed.*

But the *opposite* behavior can be true post-experimentally, with fixed data.

**Example:** Suppose we use a point mass prior at  $\theta'$  (corresponding to power at a point, and assume that this is indeed viewed as the likely alternative value).

- $R_{pre} = \frac{1-\beta(\theta')}{\alpha}$  will clearly be increasing for large  $\theta'$ .
- $B_{10}(\mathbf{x}) = \frac{f(\mathbf{x}|\theta')}{f(\mathbf{x}|\theta_0)}$  will be decreasing for large enough  $\theta'$  and fixed  $\mathbf{x}$ .

The intuition here is that the observation of (say)  $p = 0.045$  can be *less likely* under a highly powered experiment (where much smaller  $p$ -values are expected if the alternative is true) than a moderately powered experiment, and so can be less evidence in favor of the alternative for the highly powered experiment. This is correctly reflected by  $B_{10}(\mathbf{x})$ , while  $R_{pre}$  incorrectly suggests the opposite.

**A common complaint: determining Bayes factors is too hard.**  
*But  $p$ -values can be converted into a bound on Bayes factors.*

Indeed, *robust Bayesian theory* suggests a general and simple way to calibrate  $p$ -values. (Vovk, 1993, Selke, Bayarri and Berger, 2001 Am. Stat.).

- A *proper*  $p$ -value satisfies  $H_0 : p(X) \sim \text{Uniform}(0, 1)$ .
- Consider testing this versus  $H_1 : p \sim f(p)$ , where  $Y = -\log(p)$  has a decreasing failure rate (a natural non-parametric alternative).

**Theorem 1** *If  $p < e^{-1}$ ,  $B_{10} \leq \frac{1}{-ep \log(p)}$ .*

**Theorem 2** *(Examples in Edwards, Lindman and Savage, 1963). For any prior,*  

$$B_{10} \leq \sup_{\theta} f(F^{-1}(1 - p) | \theta) / f(F^{-1}(1 - p) | \theta_0),$$

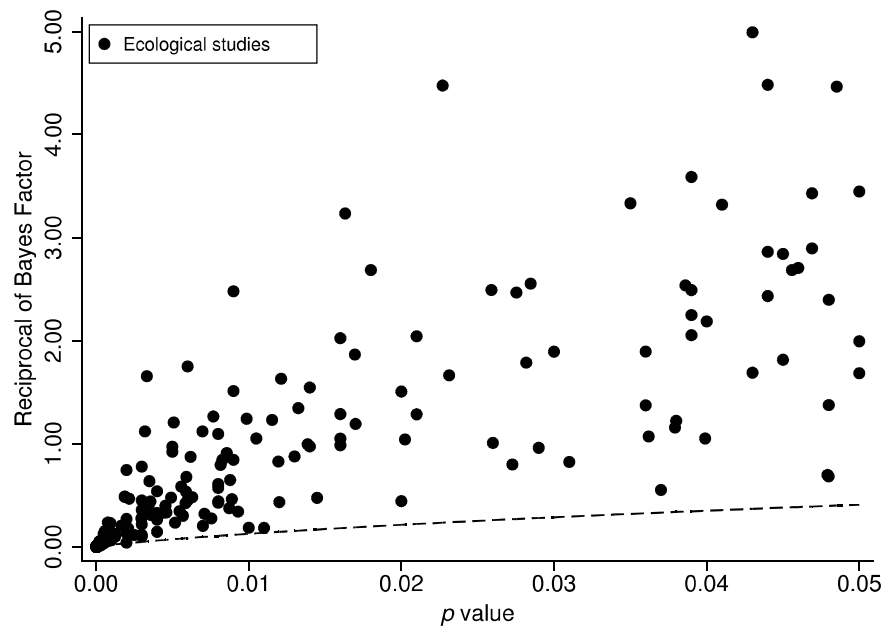
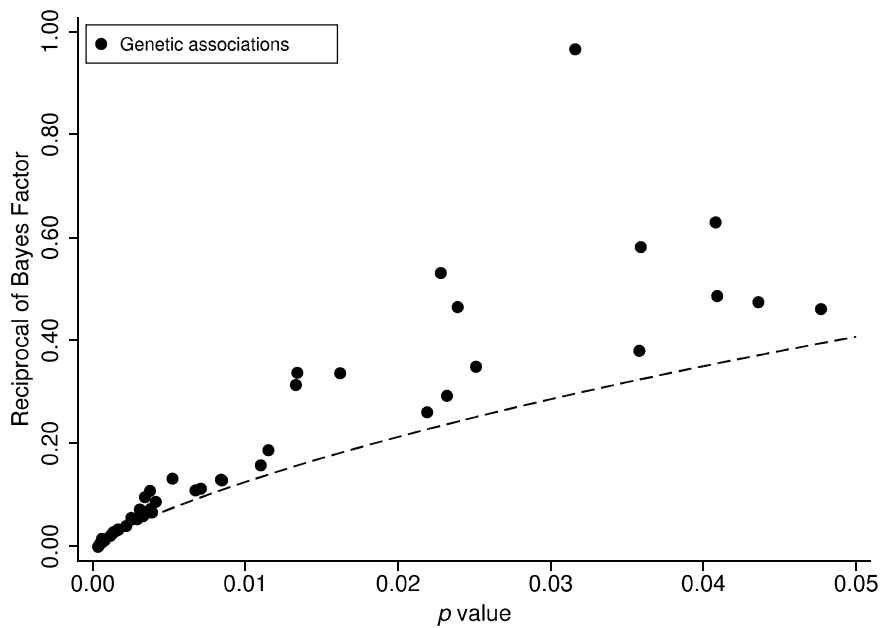
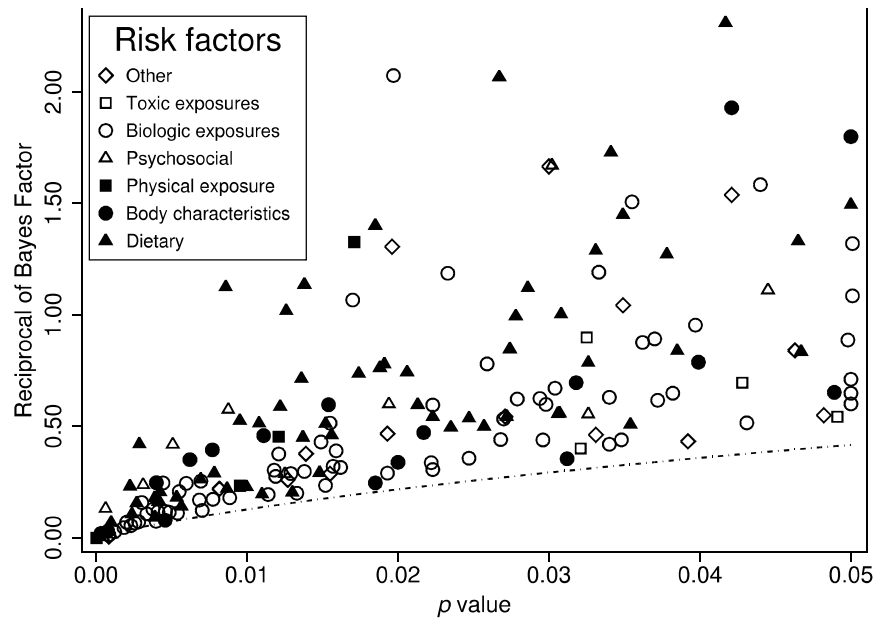
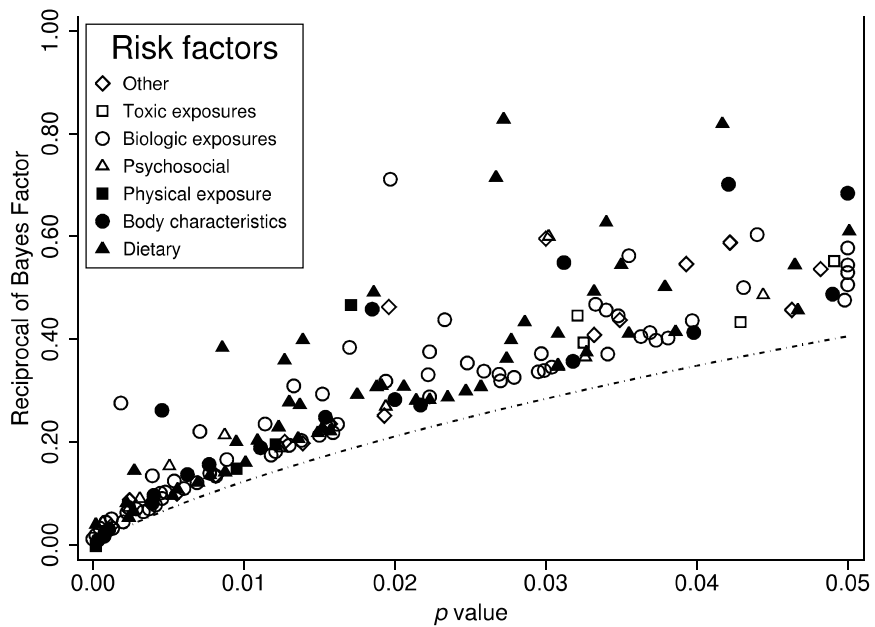
*where  $p(x) = P(T(X) > T(x) | \theta_0) \equiv 1 - F(T(x) | \theta_0)$ , with  $T$  the test statistic.*

$p$	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001	$5 \times 10^{-7}$
$\frac{1}{-ep \log(p)}$	1.60	2.44	8.13	13.9	52.9	400	3226	$2.0 \times 10^5$
$\sup_{\theta}$ , Normal	2.9	4.4	14.7	25.7	113	970	8731	$1.6 \times 10^5$



- Though very simple, there was initially concern that this bound was 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  $-ep \log(p)$  (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).



**Bayesian and frequentist possibilities for choosing the prior  $\pi(\theta)$  for the post-experimental rejection ratio or Bayes factor:**

**1. Subjective prior:** When a subjective prior is available, such as the ‘study team prior’ in the vaccine example, using it is optimal. Again, note that the resulting procedure is still as much of a frequentist procedure as is the use of the pre-experimental rejection odds  $O_{pre} = \frac{(1-\bar{\beta})}{\alpha}$ .

**2. Power considerations:** If the experiment was designed with power considerations in mind, one can use the implicit prior that was utilized to determine power. This could be a prior distribution (or weight function) if used to compute power, or a specified point (i.e., a prior giving probability one to that point) if that is what was done.

**3. Objective Bayes conventional priors:** Discussion of these can be found in Berger and Pericchi (2001). One popular such prior, that applies to our testing problem, is the *intrinsic prior* defined as follows:

- Let  $\pi^O(\theta)$  be a good estimation objective prior (often a constant), with resulting posterior distribution and marginal distribution for data  $\mathbf{x}$  given, respectively, by

$$\pi^O(\theta | \mathbf{x}) = f(\mathbf{x} | \theta)\pi^O(\theta)/m^O(\mathbf{x}), \quad m^O(\mathbf{x}) = \int f(\mathbf{x} | \theta)\pi^O(\theta) d\theta.$$

- Then the intrinsic prior (which will be proper) is

$$\pi^I(\theta) = \int \pi^O(\theta | \mathbf{x}^*)f(\mathbf{x}^* | \theta_0) d\mathbf{x}^*,$$

with  $\mathbf{x}^* = (x_1^*, \dots, x_q^*)$  being imaginary data of the smallest sample size  $q$  such that  $m^O(\mathbf{x}^*) < \infty$  (an imaginary bootstrap construction).

$\pi^I(\theta)$  is often available in closed form, but even if not, computation of the resulting Bayes factor is often a straightforward numerical exercise.

**4. Empirical Bayes prior:** This is found by maximizing the numerator of  $R_{post}(\mathbf{x}) = B_{10}(\mathbf{x})$  over some class of possible priors. Common are the class of nonincreasing priors away from  $\theta_0$  (considered in the vaccine example) or even the class of all priors as in Theorem 2.

**5.  $p$ -value bound:** Instead of picking a prior distribution to calculate  $R_{post}(\mathbf{x}) = B_{10}(\mathbf{x})$ , use the generic upper bound  $B_{10}(\mathbf{x}) < 1/[-ep \log p]$ .

These last two approaches have the problem that they are significantly biased (in the wrong way) from both Bayesian and frequentist perspectives. Indeed, if  $\bar{R}_{post}(\mathbf{x})$  is the answer obtained from either approach, then

$$R_{post}(\mathbf{x}) < \bar{R}_{post}(\mathbf{x}), \quad R_{pre} < E[\bar{R}_{post}(\mathbf{x}) \mid H_0, \mathcal{R}].$$

Thus, in either case, one is reporting *larger* rejection ratios in favor of  $H_1$  than is supported by the data. (Of course, use of either method would give much better answers than the usual misinterpretations of  $p$ -values.)

## Conclusion: saving the world from $p$ -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  $H_0$ :

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

where  $O_P$  is the prior odds and  $B_{10}(\mathbf{x})$  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. Possible choices of  $\pi(\theta)$  for computing the Bayes factor:
  - A subjective prior or ‘weight function’ chosen during a pre-experimental power computation. (Note that  $B_{10}(\mathbf{x})$  is as frequentist as  $\frac{1-\bar{\beta}}{\alpha}$ .)
  - A point mass at a value of  $\theta$  used in a pre-experimental power calculation.
  - An objective prior distribution (e.g., the *intrinsic prior* for testing).
  - The prior from a class of priors that most favors  $H_1$ .
- If determination of  $B_{10}(\mathbf{x})$  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!