Counterexamples that helped to shape the objective Bayesian philosophy

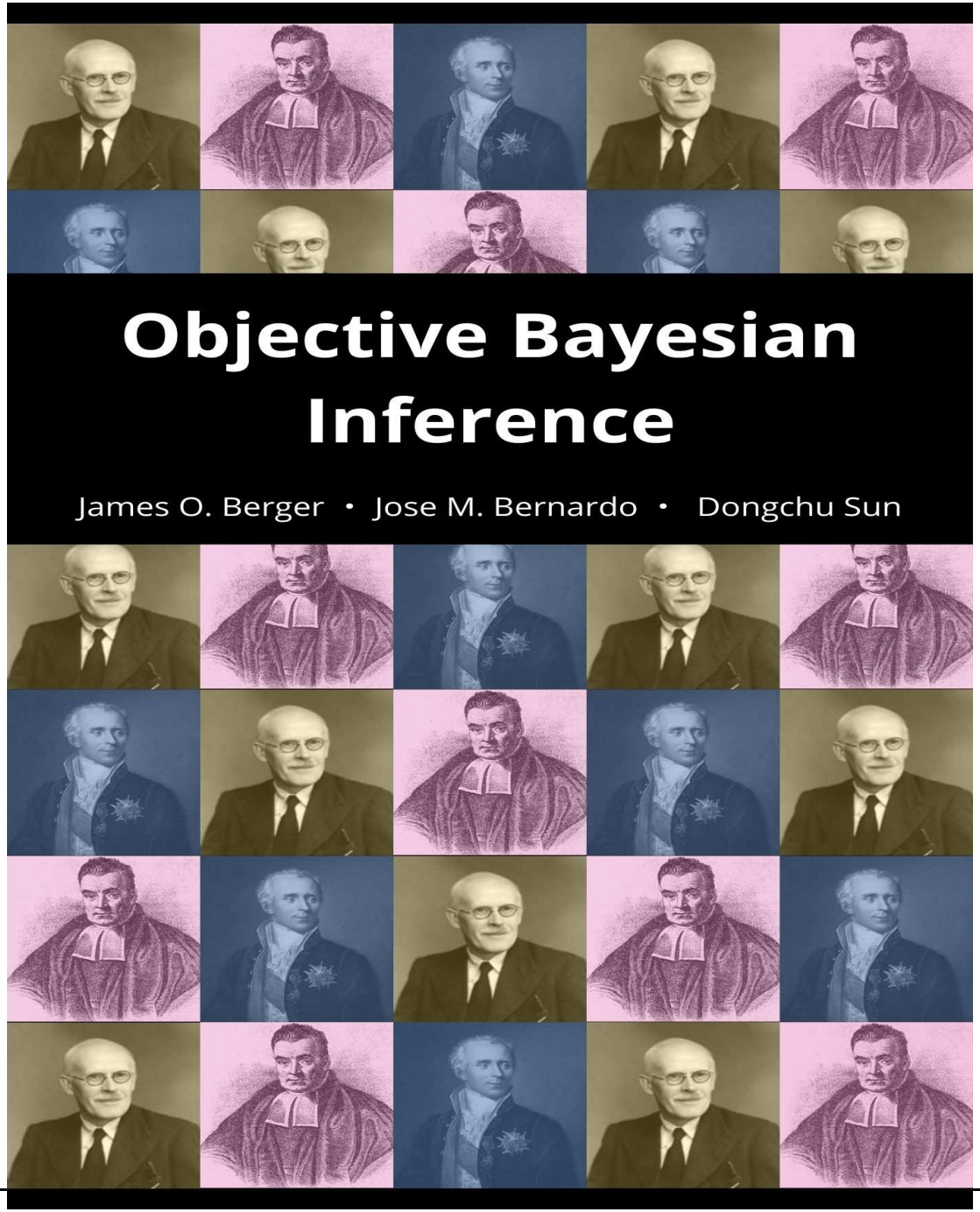
Jim Berger

Duke University and Texas A&M University

OBayes 2025

April 9, 2025

- *Learning statistics from counter examples: ancillary statistics* was a famous article by Debabrata Basu (Basu, 2011).
- For a recent volume in Sankyha honoring Basu, I recently wrote an article with the same title (Berger, 2024).
- The counterexamples in this talk are taken from the 22 counterexamples in that paper, or from *Objective Bayesian Inference*.


Objective Bayesian Inference
James O. Berger • Jose M. Bernardo • Dongchu Sun

Counterexample to Inverse Probability

Laplace invented inverse probability (a name given by Augustus de Morgan in 1838) and rediscovered Bayes theorem.

- Inverse probability proceeded by
  - choosing or developing a probability model $f(x \mid \theta)$ for the data x, given unknown parameters θ;
  - choosing the prior $\pi(\theta) = 1$;
  - obtaining the posterior
$$\pi(\theta \mid x) = \frac{f(x \mid \theta) \times 1}{\int f(x \mid \theta) \times 1 \; d\theta} \; ;$$
  - finding the median of this posterior distribution (as well as other features).
- This was *the standard method of statistical analysis* until about 1930 (i.e., for over 150 years), and is still in use today.

The main counterexample to inverse probability: it is not invariant to parameterization.

Example: In analysis using the normal distribution, the parameterizations used for the scale parameter in the 19th century were σ, σ^2, $1/\sigma^2$ and $\log \sigma$. Using a constant prior for each parameterization results, say, in the posterior distribution of the normal mean being a t-distribution, but with differing degrees of freedom.

Proposed fix during the 1930's by Harold Jeffreys: If the data model density is $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ the *Jeffreys-rule prior* for the unknown $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is

$$|I(\boldsymbol{\theta})|^{1/2} d\theta_1 \ldots d\theta_k$$

where $I(\boldsymbol{\theta})$ is the $k \times k$ matrix Fisher's information matrix with (i, j) entry

$$I(\boldsymbol{\theta})_{ij} = \mathrm{E}_{\boldsymbol{X} \mid \boldsymbol{\theta}} \left[- \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\boldsymbol{X} \mid \boldsymbol{\theta}) \right].$$

- This is invariant to parameterization!
- But yielded the wrong degrees of freedom for the normal mean problem.
- Welch and Peers (1963) showed that, for one dimensional θ, the Jeffreys-rule prior essentially gives optimal frequentist answers.

The greatest counterexample in statistics: the *Likelihood Principle*

- Two core components of frequentist theory in 1961 were
  - The *sufficiency principle.*
  - The *conditionality principle*; as in the following David Cox example.

Example: An employee is randomly given either a measurement instrument with variance 1 (new) or one with variance 3 (old) to perform assays.

- *Conditional inference:* For each measurement, report variance 1 or 3, depending on the instrument being used.
- *Unconditional inference:* The overall variance of the assays is $\frac{1}{2} \times 1 + \frac{1}{2} \times 3 = 2$, so report a variance of 2 regardless of the instrument actually being used.

The conditionality principle says to do the conditional inference.

Focus on the likelihood function $\mathcal{L}(\boldsymbol{\theta}) = f(\boldsymbol{x} \mid \boldsymbol{\theta})$, for the observed data $\boldsymbol{x}$.

Likelihood Principle (LP):

- All the information about $\boldsymbol{\theta}$ obtainable from an experiment is contained in $\mathcal{L}(\boldsymbol{\theta})$. Thus frequentist averaging over $\boldsymbol{x}$ would be precluded!
- Two likelihood functions $\mathcal{L}_1(\boldsymbol{\theta})$ and $\mathcal{L}_2(\boldsymbol{\theta})$ (from the same or different experiments but about the same $\boldsymbol{\theta}$) contain the same information about $\boldsymbol{\theta}$ if they are proportional to one another.

Virtually all frequentists viewed the LP as being wrong, but Birnbaum (1962) proved that the LP is a logical consequence of the sufficiency principle and conditionality principle!

The LP does not say how to use $\mathcal{L}(\boldsymbol{\theta})$, but OBayes provides the most natural use.

Frequentist counterexamples that strongly impacted OBayes

Optimal frequentist procedures must be Bayesian.

- Consider a *decision rule* $\boldsymbol{\delta}(\boldsymbol{x})$, with $L(\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta})$ being the loss if the decision rule is used and $\boldsymbol{\theta}$ is the parameter.
- The quality of $\boldsymbol{\delta}$ is measured by the *risk function*

$$R(\boldsymbol{\delta}, \boldsymbol{\theta}) = E[L(\boldsymbol{\delta}(\boldsymbol{X}), \boldsymbol{\theta})]\,,$$

  where the expectation is with respect to $\boldsymbol{X}$ given $\boldsymbol{\theta}$.
- $\boldsymbol{\delta}$ is admissible [inadmissible] if it cannot [can] be improved in risk, improvement meaning there is a $\boldsymbol{\delta}^*(\boldsymbol{x})$ such that $R(\boldsymbol{\delta}^*, \boldsymbol{\theta}) \leq R(\boldsymbol{\delta}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ with strict inequality for some $\boldsymbol{\theta}$.
- **Theorem** (Wald, Stein, Farrell): Any admissible decision rule must be *generalized Bayes*, i.e. Bayes with respect to a proper or improper prior.

Stein shrinkage estimation

- Independently, $X_i \sim N(\theta_i, 1)$, $i = 1, 2, \ldots, p$; define $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$.
- It is desired to estimate $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ with an estimator $\boldsymbol{\delta}(\boldsymbol{x}) = (\delta_1(\boldsymbol{x}), \delta_2(\boldsymbol{x}), \ldots, \delta_p(\boldsymbol{x}))$, under the loss $L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{p} (\delta_i(\boldsymbol{x}) - \theta_i)^2$.
- The maximum likelihood estimate, unbiased estimate, fiducial estimate and inverse probability Bayes estimate is $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x}$.
- Blyth (1951) showed this was admissible for $p = 1$.
- Stein (1959) showed this was admissible for $p = 2$.
- James and Stein (1960) showed this was inadmissible for $p \geq 3$, with improvement obtained by the *shrinkage estimator*

$$\boldsymbol{\delta}^{JS}(\boldsymbol{x}) = \left(1 - \frac{(p-2)}{|\boldsymbol{x}|^2}\right) \boldsymbol{x}\,.$$

- While unintended, this gave considerable impetus to the hierarchical Bayes movement, because hierarchical Bayes was all about shrinkage.

Counterexamples to use of the multivariate Jeffreys-rule prior:
The most commonly used prior in objective Bayesian analysis is the Jeffreys-rule prior (Jeffreys, 1961), given by $\pi^J(\boldsymbol{\theta}) = |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2}$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information matrix.

- This is great if the parameter is one-dimensional.
- It is bad in higher dimensions; here are two examples.

Example. Inconsistency in the Neyman-Scott problem.
$\boldsymbol{x} = \{x_{ij}\}$, $i = 1, \ldots, m$, $j = 1, 2$, has density

$$p(\boldsymbol{x} \,|\, \mu_1, \ldots, \mu_m, \sigma^2\} = \prod_{i=1}^{m} \prod_{j=1}^{2} \mathrm{N}(x_{ij} \,|\, \mu_i, \sigma^2) \,.$$

Neyman and Scott (1948) showed that the maximum likelihood estimator of σ^2 is inconsistent as $m \to \infty$. So is the posterior distribution of σ^2 when using the Jeffreys-rule prior $\pi(\mu_1, \ldots, \mu_m, \sigma^2) \propto \sigma^{-(m+2)}$; indeed, the posterior converges to a point mass at half the true value of σ^2.

Example. Underdispersion in the Multinomial Problem (Berger, Bernardo and Sun (2015)). Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is Multinomial$(\boldsymbol{x} \,|\, n, \theta_1, \ldots, \theta_m)$. The Jeffreys-rule prior is

$$\pi^J(\theta_1, \ldots, \theta_m) \propto \left(1 - \sum\nolimits_{j=1}^{m} \theta_j\right)^{-1/2} \prod_{j=1}^{m} \theta_j^{-1/2}\,, \tag{1}$$

which is the Dirichlet$((\theta_1, \ldots, \theta_m) \,|\, (\frac{1}{2}, \ldots, \frac{1}{2}))$ distribution. The corresponding posterior distribution is Dirichlet$((\theta_1, \ldots, \theta_m) \,|\, (x_1 + \frac{1}{2}, \ldots, x_m + \frac{1}{2}))$. This is problematical:

- Suppose $n = 3$, $m = 1000$, $x_{240} = 2$, $x_{876} = 1$, and the other $x_i = 0$.
- The posterior means can be shown to be

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/2}{\sum_{j=1}^{m} [x_j + 1/2]} = \frac{x_i + 1/2}{n + m/2} = \frac{x_i + 1/2}{503}\,.$$

- Thus $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] = 2.5/503 = 0.005$, $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] = 1.5/503 = 0.003$, and $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = 0.5/503 = 0.001$ for the cells with no observations.
  - Particularly troubling is that cell 240 has two of the three observations, but posterior probability 0.005.

Counterexamples suggesting elements of reference prior theory

- In nonregular problems the Fisher information will typically not exist, so there is no Jeffreys-rule prior; reference prior theory utilizes a much more general notion of asymptotic missing information that applies in almost complete generality.
- The problems with the multivariate Jeffreys-rule prior are resolved in reference prior theory by
  - recognizing that objective priors should depend on the parameter of interest – σ^2 in the Neyman-Scott example and individual cell probabilities in the multinomial problem –
  - sequentially deriving the reference prior one parameter at a time, a generalization of the independence Jeffreys prior for the normal problem.

Counterexamples relevant to OBayes hypothesis testing

- The Bartlett (1957) paradox that, in testing a point null hypothesis versus a compound alternative, the use of increasingly vague proper priors on the alternative causes the posterior probability of the null hypothesis to go to 1, regardless of the data. (Of course, this was implicit in Jeffreys development of Bayesian hypothesis testing.)
- In model uncertainty, assigning all models equal prior probability often fails to lead to multiplicity control of false positives (Scott and Berger (2010)).
- In model uncertainty, the *maximum posterior probability (MAP)* model is often not the best single model. Often better (Barbieri and Berger (2004)) is the *median probability model* defined by
  - Calculating the *posterior inclusion probability* of the features used to define the models.
  - Choosing the model to be that which includes only the features whose posterior inclusion probability exceeds 0.5.

Example. The Hald regression data set, that has been used by several authors (see Burnham and Anderson (1998)), has $n = 13$ observations $\boldsymbol{y}$ that are regressed on four possible regressors: x_1, x_2, x_3, x_4, the full model being

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon, \;\; \epsilon \sim \mathrm{N}(0, \sigma^2)\,,$$

with σ^2 unknown. Consider the models defined by subsets of regressors, with the intercept being present in all models. Thus

$$\text{Model } \{1,3,4\} \quad \text{denotes the model} \quad y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon\,.$$

Table 1 reports the results of a model uncertainty analysis using the encompassing AIBF approach of Berger and Pericchi (1996).

Table 1: Posterior model probabilities and excess predictive risks (the difference between predictive risk of the model and the predictive risk of the optimal model averaged prediction, assuming square error predictive loss).

Model	$Pr(M_i \mid \boldsymbol{y})$	$\Delta R(M_i)$
null	0.000003	2652.44
{1}	0.000012	1207.04
{2}	0.000026	854.85
{3}	0.000002	1864.41
{4}	0.000058	838.20
{1,2}	0.275484	8.19
{1,3}	0.000006	1174.14
{1,4}	0.107798	29.73

Model	$Pr(M_i \mid \mathbf{y})$	$\Delta R(M_i)$
{2,3}	0.000229	353.72
{2,4}	0.000018	821.15
{3,4}	0.003785	118.59
{1,2,3}	0.170990	1.21
{1,2,4}	0.190720	0.18
{1,3,4}	0.159959	1.71
{2,3,4}	0.041323	20.42
{1,2,3,4}	0.049587	0.47

The posterior inclusion probabilities here (the overall probability that a variable is in a model) are

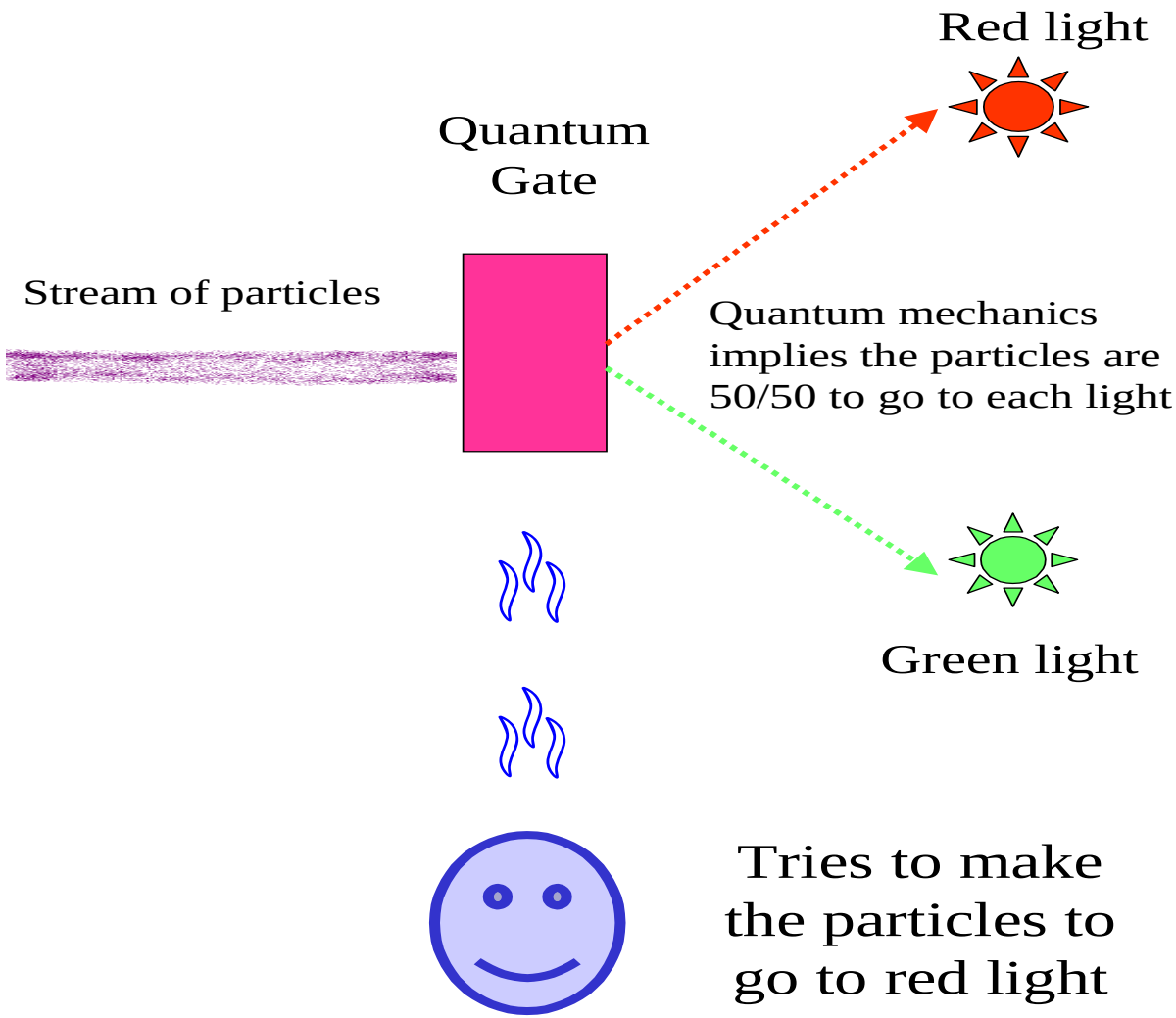
$$p_1 = \sum_{j:x_1 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.95, \;\; p_2 = \sum_{j:x_2 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.73\,,$$
$$p_3 = \sum_{j:x_3 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.43, \;\; p_4 = \sum_{j:x_4 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.55.$$

Thus the median probability model, the model consisting of those variables whose posterior inclusion probability exceeds 0.5, is $\{1, 2, 4\}$.

The Jeffreys (1939)-Lindley (1957) 'Paradox': In testing with very large sample sizes, a frequentist can think that there is overwhelming evidence against the null hypothesis, while a Bayesian thinks there is overwhelming evidence in favor of the null hypothesis.

Psychokinesis Example: Do people have the ability to perform *psychokinesis*, affecting objects with thoughts?

The experiment: Schmidt, Jahn and Radin (1987) used electronic and quantum-mechanical random event generators with visual feedback; the subject with alleged psychokinetic ability tries to "influence" the generator.


Red light
Quantum
Gate
Stream of particles
Quantum mechanics
implies the particles are
50/50 to go to each light
Green light
Tries to make
the particles to
go to red light

Data and model:

- Each "particle" is a Bernoulli trial (red = 1, green = 0)
  θ = probability of "1"
  $n = 104,490,000$ trials
  X = # "successes" (# of 1's), $\quad X \sim \text{Binomial}(n, \theta)$
  $x = 52,263,470$ is the actual observation

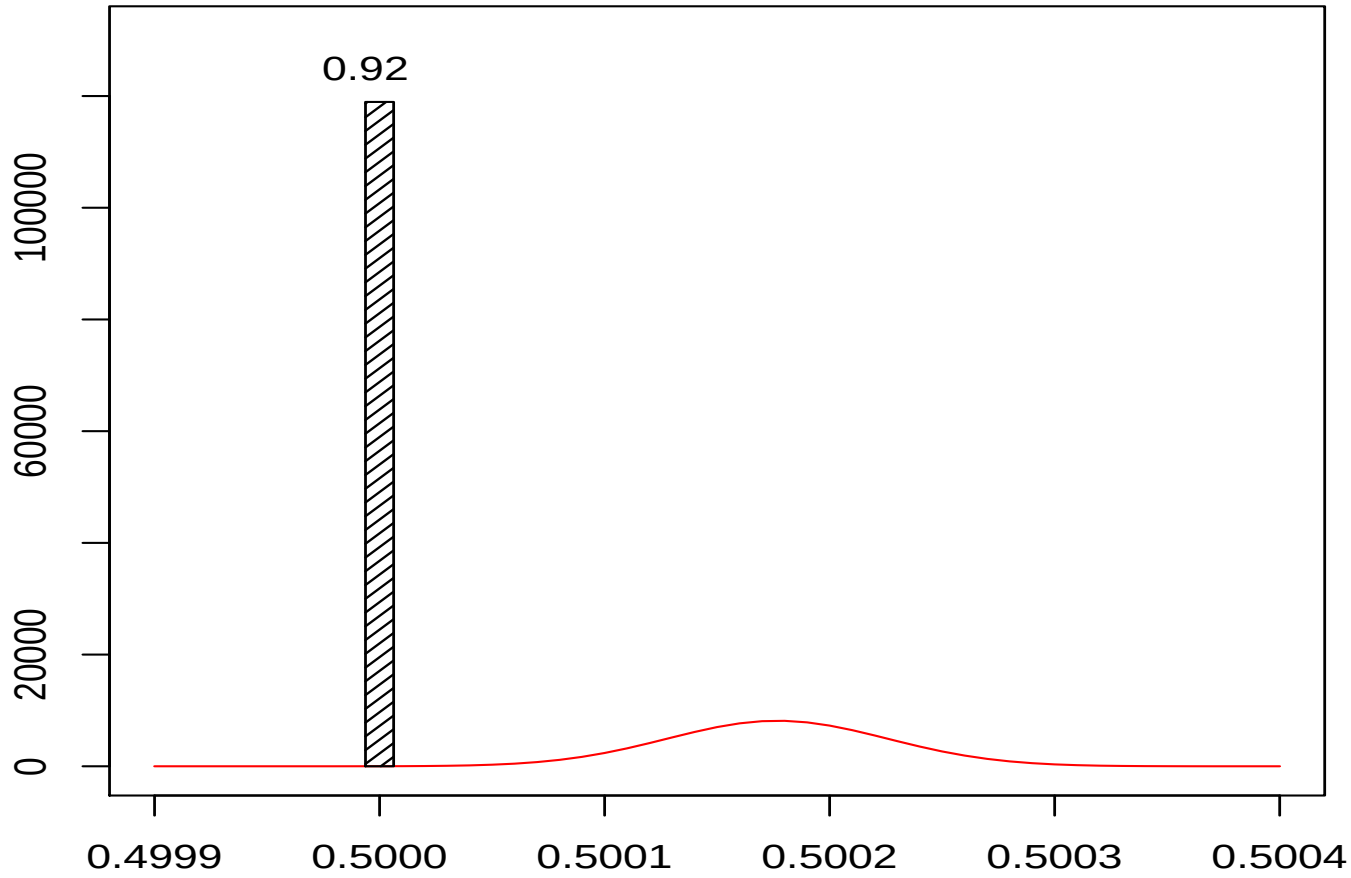
  To test $H_0 : \theta = \frac{1}{2}$ (subject has no influence)

  versus $H_1 : \theta \neq \frac{1}{2}$ (subject has influence)

- P-value $= P_{\theta=\frac{1}{2}}(|X - \frac{n}{2}| \geq |x - \frac{n}{2}|) \approx .0003.$
  From a frequentist perspective, this would seem to ve very strong evidence in favor of psychokinesis.

Bayesian Analysis: (Jefferys, 1990)

- The objective Bayesian prior distribution would be
  $Pr(H_0) = Pr(H_1) = \frac{1}{2}$ and $\pi(\theta) = 1 \quad$ (on $0 < \theta < 1$).
- Computation yields
  - $Pr(H_0 \mid x = 52,263,470) \approx 0.92$ (recall, p-value $\approx .0003$).
  - Posterior density on $H_1 : \theta \neq \frac{1}{2}$ is the
    $Be(\theta \mid 52,263,471\ ,\ 52,226,531)$ density.


0.92
0
20000
60000
100000
0.4999
0.5000
0.5001
0.5002
0.5003
0.5004

- Choice of the Uniform prior on H_1 is highly questionable.
- A robust Bayesian analysis would, say, consider $\pi_r(\theta) = U(\theta \mid \frac{1}{2} - r, \frac{1}{2} + r)$; here r could be interpreted as the the largest change in success probability that you would expect, given that psychokinesis exists.
- One could then study the posterior probability as a function of r.

A counterexample to interpreting p-values as error rates: Data X has a specified distribution under the null hypothesis, and $p(\cdot)$ is a proper p-value under the null. Here is a result from Vovk (1993).

Theorem. A *proper* p-value, $p(\cdot)$, satisfies $H_0 : p(X) \sim \text{Uniform}(0,1)$ (the definition of a proper p-value). Test this hypothesis versus $H_1 : p \sim \text{Beta}(1, \text{b})$, $b > 1$. Letting B_{01} denote the Bayes factor of H_0 to H_1,

$$B_{01} = \frac{1}{b(1-p)^{(b-1)}} \geq -e\, p\, \log(p) \quad \text{for} \quad p < e^{-1}\,. \tag{2}$$

- The $\text{Beta}(1, \text{b})$, $b > 1$, are decreasing in p, which is natural.
- This class can be generalized to the class of all priors such that $Y = -\log(p)$ has a non-increasing failure rate (Sellke et al., 2001), a natural non-parametric condition that covers most cases of interest.

An analogous bound can be given on the conditional Type I frequentist error (see Berger et al. (1994) for definition):

$$\alpha(p) \geq (1 + [-e\, p \log(p)]^{-1})^{-1}.$$

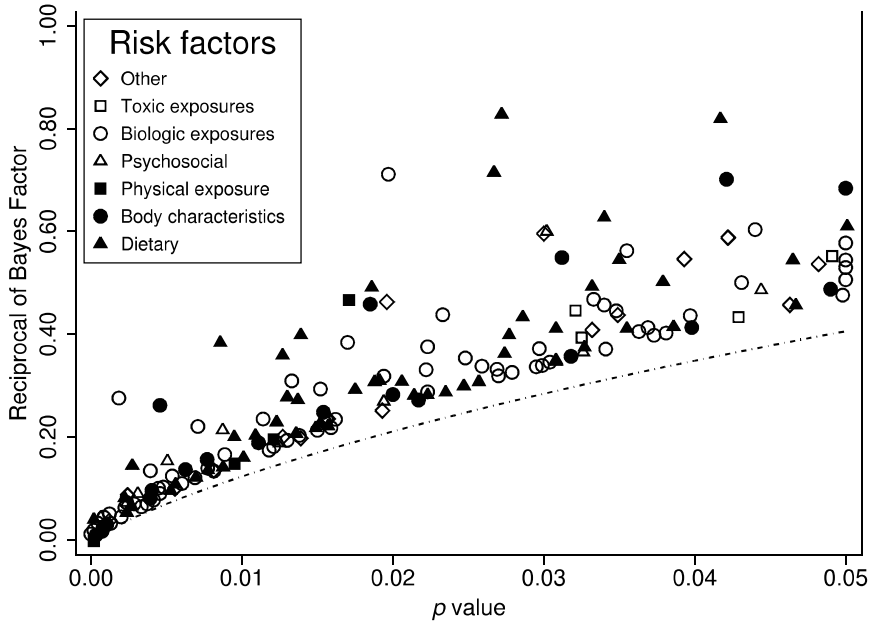
p	.2	.1	.05	.01	.005	.001	.0001	.00001
$-ep\log(p)$	.879	.629	.409	.123	.072	.0189	.0025	.00031
$\alpha(p)$	.465	.385	.289	.111	.067	.0184	.0025	.00031

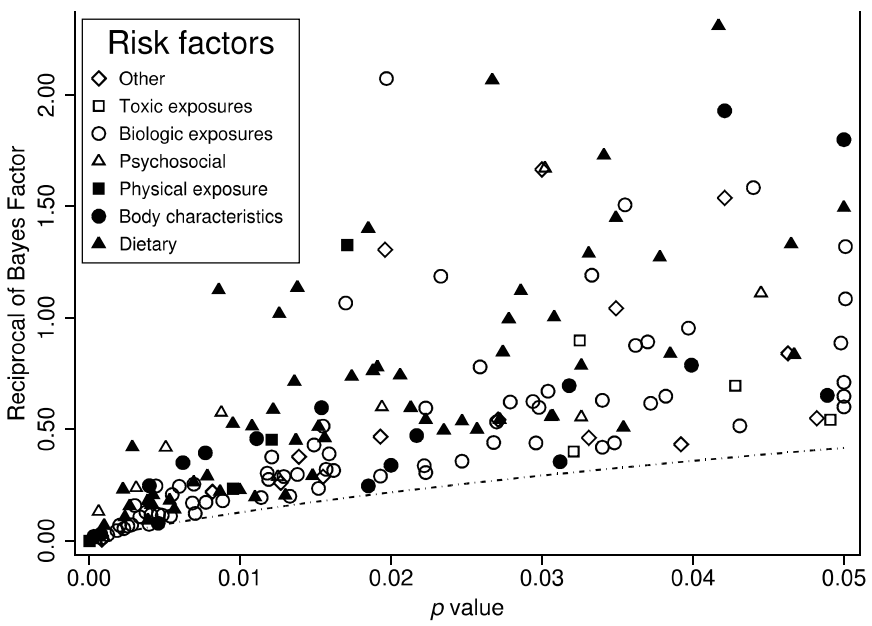
Table 2: p-values and the associated lowest possible Bayes factors and conditional frequentist error probabilities.

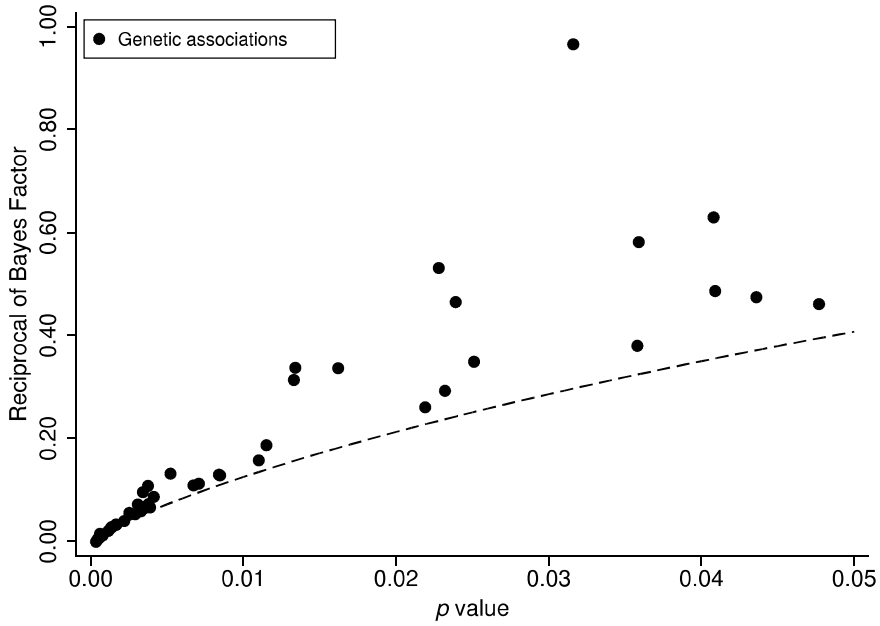
So p-values are much too small (often orders of magnitude too small) to have any interpretation as error probabilities.

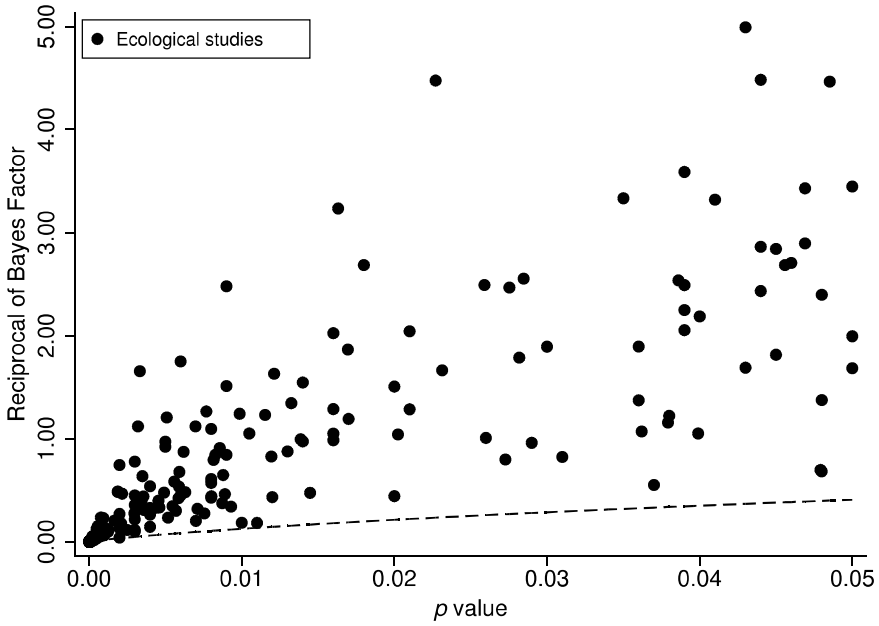
- Although very simple, there was initially concern that the $-ep\log(p)$ bound is too small, since it is known that Bayes factors can depend strongly on the sample size n, and the bound does not.
- But the following studies indicate that this might not typically be a problem. These studies
  - look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

Counterexamples to subjective Bayes

Overly precise elicitations: The folklore in Bayesian statistics is that, if someone is asked to give their prior estimate of an unknown quantity and assess the likely error in their estimate (say by stating the variance of their estimate), they will underestimate the error by at least a factor of 3.

Example. Underestimating variances involving Cepheid variable stars.

- Observations $x_1, \ldots, x_n$ were independentlhy $\mathrm{N}(x \,|\, \mu, \sigma_i^2)$, with the σ_i^2 being specified and claimed to be very accurate.
- A small part of Barnes III et al. (2003) studied this claim, by modeling the observations x_i as, instead, being $\mathrm{N}(x_i \,|\, \mu, \tau^2\sigma_i^2)$ random variables, with τ^2 unknown and assigned the objective prior $\pi(\tau^2) = 1/\tau^2$.
- The posterior distribution of τ^2 was centered at about 2 in one study and around 4 in another. These estimated variances arose from some of the most careful subjective elicitations in science, and yet they prominently underestimated the error.

Example. Hidden (bad) impacts of conjugate prior distributions for covariance matrices:

- Consider i.i.d. multivariate normal data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where each k-dimensional column vector $\boldsymbol{x}_i \sim \mathrm{N}_k(\boldsymbol{x} \,|\, \boldsymbol{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ unknown.
- The most commonly used subjective prior for $\boldsymbol{\Sigma}$ is the Inverse Wishart prior, for subjectively specified a and b, $\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-a/2} \exp\{-\frac{1}{2}\mathrm{tr}[b\, \boldsymbol{\Sigma}^{-1}]\}$.
- Consider the spectral decomposition $\boldsymbol{\Sigma} = \boldsymbol{O}\boldsymbol{D}\boldsymbol{O}'$, with $\boldsymbol{O}$ being an orthogonal matrix and $\boldsymbol{D}$ being a diagonal matrix with diagonal entries $d_1 > d_2 > \cdots > d_k$.
- Changing variables to $\boldsymbol{O}$ and $\boldsymbol{D}$ yields (see Yang and Berger (1994))

$$\pi(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma} \propto |\boldsymbol{D}|^{-a/2} \exp\{-\tfrac{1}{2}\mathrm{tr}[b\, \boldsymbol{D}^{-1}]\} \prod_{i<j} (d_i - d_j) \cdot I_{[d_1 > \cdots > d_k]}\, d\boldsymbol{D}\, d\boldsymbol{O},$$

  where $I_{[d_1 > \cdots > d_k]}$ denotes the indicator function on the given set.
- The term $\prod_{i<j}(d_i - d_j)$ is near zero when any eigenvalues are close, so the inverse Wishart prior forces apart the eigenvalues of the covariance matrix, contrary to typical judgement.

Thanks

References

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32:870–897.

Barnes III, T. G., Jefferys, W. H., Berger, J. O., Mueller, P. J., Orr, K., and Rodriguez, R. (2003). A Bayesian analysis of the Cepheid distance scale. *Astrophysical J.*, 592:539.

Berger, J. and Pericchi, L. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Annals of Statistics*, 22:1787–1807.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–326, with discussion.

Burnham, K. P. and Anderson, D. (1998). *Model Selection and Inference – A Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edition.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, 16:1–32.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38:2587–2619.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55:62–71.

Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(2):317–341.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, 22:1195–1211.