Learning Statistics From Counterexamples

James Berger*, *Duke University*

November 5, 2024

1 Introduction

The title of this article is (essentially) the same as the famous paper Basu (2011b). Basu often opined that counterexamples were the best way to learn limitations of theories or methods and I have followed his directive in my own teaching. In fact, my PhD advisor – Larry Brown – once told me that he thought my philosophy of statistics was simply the intersection of those concepts and approaches that survived counterexamples. A number of the counterexamples that affected my philosophy of statistics are collected in this article.

2 Two counterexamples of Basu, to set the stage

Many of Basu's writings focused on survey sampling or the concept of ancillarity. Here are two of his counterexamples, the first from Basu (2011a) and the second from Basu (2011b). The description in the first example is essentially a direct quote from Basu (2011a), so as to also give a sense of the wonderful writing style of Basu.

Elephant Counterexample: "The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total

*E-mail: berger@duke.edu

weight $Y = Y_1 + \ldots + Y_{50}$ of elephants. But the circus statistician is horrified when he learns the owners purposive samplings plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $\frac{100}{99}y \approx y$ and not $50y$." [The Horvitz-Thompson estimator weights the observations by the inverse of the probability of which that unit was selected.] "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz- Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)"

Over the years, many who have commented about this example state that it is highly unrealistic, and is not the type of situation that the Horvitz-Thompson estimator was designed for. But Basu was very clear that he was not criticizing practical use of the Horvitz-Thompson estimator (although he did point out potential problems in its use when using probability proportional to size sampling — Example 4 in Basu (2011a)). Rather, he was showing that the usual logic used to justify the Horvitz-Thompson estimator — namely that it is unbiased in terms of the unit selection probabilities — is faulty logic, since using the same logic in the elephant example led to an absurdity. So he was simply pointing out that different justifications were needed for the Horvitz-Thompson estimator.

Ancillarity Counterexample: Let $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$, be iid observations whose joint distribution is bivariate normal with zero means, unit variances and correlation ρ, which is the parameter of interest. Note that $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are each a vector of independent $N(0, 1)$ random variables, so that their distributions do not depend on ρ; they are thus each an *ancillary statistic*. The ancillary method in statistics proposes conditioning on an ancillary statistic, and then analyzing the remaining data. Conditioning on $\mathbf{x}$, the y_i are independently $N(\rho x_i, 1 - \rho^2)$, so one might estimate ρ by $\sum x_i y_i / \sum x_i^2$, since it is clearly an unbiased estimate of ρ. On the other hand, one could instead condition on $\mathbf{y}$,

leading to the unbiased estimate $\sum x_i y_i / \sum y_i^2$. Thus utilizing ancillarity does not necessarily lead to a unique answer, and these estimates are likely suboptimal. (Usual estimators of ρ in this situation are based on the sample correlation $r = \sum x_i y_i / \sqrt{\sum x_i^2 \sum y_i^2}$.)

Again, Basu was not trying to say that ancillarity is a useless concept; he was actually quite supportive of its typical uses in practice (although he preferred just taking a Bayesian approach, for which one does not need the concept of ancillarity). He was, instead, simply pointing out that one cannot build a complete theory of statistical inference based on ancillarity, something that many have tried to do over the years.

Many of the examples in this paper have the same flavor as these two examples of Basu's; they often look at rather extreme situations to make a point about logical difficulties with various statistical approaches or methodologies. Some of the examples do go further, however, and call into question actual statistical practice; this distinction will be highlighted when relevant.

3 Conditioning counterexamples, the Likelihood Principle and the Stopping Rule Principle

Many believe that the biggest difference between various statistical approaches is that some (e.g. the Bayesian approach) condition on the actual observed data, while others (e.g. the traditional frequentist approach) include averages over all data. My first introduction to conditioning was the famous example of David Cox (Cox, 1958).

Example 1. A variant of the Cox example: Each day an employee arrives at work to perform measurements, and is given an unbiased instrument to make the measurements. Half of the available instruments are relatively new and have a variance of 1, while the others are older and have variance 3. The employee is assigned each type randomly but knows whether the instrument is old or new.

Conditional inference: For each measurement, report variance 1 or 3, depending on the variance of the instrument actually 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.

It seems unreasonable to perform the unconditional inference here, especially because the conditional inference is also frequentist, in that the conditional reports are just averaging over the low variance measurements or over the high variance measurements, respectively. Because the conditional inferences are also frequentist, many dismiss this example as not being

relevant to statistics. However, the virtually identical issue can arise within an experiment, as demonstrated in the following example, from Berger and Wolpert (1988).

Example 2. Pedagogical conditioning example: Two observations, X_1 and X_2, are taken, where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

Consider the following confidence set for the unknown $\theta \in (-\infty, \infty)$:

$$C(X_1, X_2) = \begin{cases} \text{the point } \{\frac{1}{2}(X_1 + X_2)\} & \text{if } X_1 \neq X_2 \\ \text{the point } \{X_1 - 1\} & \text{if } X_1 = X_2. \end{cases}$$

The frequentist coverage of this confidence set can easily be seen to be

$$P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75\,.$$

This is a silly report once the data is at hand. If $x_1 \neq x_2$, then $\frac{1}{2}(x_1 + x_2)$ is equal to θ, so the confidence set is then actually 100% accurate. If $x_1 = x_2$, we do not know whether θ equals the data's common value plus one or their common value minus one, and each of these possibilities is equally likely to have occurred. Thus intuition suggests that the confidence interval is then only 50% accurate. It is not incorrect to say that the confidence interval has 75% coverage, but it is much more scientifically useful to report 100% or 50%, depending on the data. And this conditional report can, again, be given a fully frequentist justification, as averaging over the sets of data $\{(x_1, x_2) : x_1 \neq x_2\}$ and $\{(x_1, x_2) : x_1 = x_2\}$, respectively.

The clear message from this example (and the Cox) example is that frequentists cannot ignore the need to involve conditioning on the data. This was already well known, as the need to condition on ancillary statistics. (Indeed, $T = x_1 - x_2$ in the pedagogical example is an ancillary statistic, and conditioning on T produces the correct answers.) But it is still often necessary to condition even when an ancillary statistic is not available, as we will see in many of the examples herein.

The need to condition, in general, was strongly reinforced by the Likelihood Principle (LP) and Stopping Rule Principle (SRP) and their justifications. The LP is actually the more general of the two and has the best logical justification, but we begin with the SRP because it is supported by more fun counterexamples.

To introduce the SRP, suppose an experiment E is conducted, which consists of observing data $\boldsymbol{x}$ having density $f(\boldsymbol{x} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the unknown parameters of the statistical model. Let $\boldsymbol{x}_{obs}$ denote the data actually observed.

Stopping Rule Principle (SRP): *The reasons for stopping experimentation have no bearing on the information about $\boldsymbol{\theta}$ arising from E and $\boldsymbol{x}_{obs}$.*

Serious discussion of the SRP goes back at least to Barnard (1949), who wondered why thoughts in the experimenter's head concerning why to stop an experiment should affect how we analyze the actual data that were obtained. Here are two amusing example of this, the first a variant of a well known example and the second from Berger and Berry (1988).

Example 3. Thoughts in heads 1: Two scientists are collaborating and have a joint graduate student who is conducting an experiment. They both watch her as she collects the data. The observations $x_1, x_2, \ldots$ are i.i.d. Bernoulli(θ) random variables and the scientists are testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$. After the ninth observation, they simultaneously say "That's enough," and tell the student to stop collecting data. The final data consists of 9 successes and 3 failures.

Each scientist separately analyzes the data and, when getting back together, are surprised to find that they reached different conclusions. Scientist 1 says that there is not significant evidence against H_0 at the 0.05 level, while Scientist 2 claims that there is significant evidence at the 0.05 level. How did this disagreement happen?

Scientist 1's analysis: He had planned to take just 12 observations. Thus the number of successes, x, is Binomial$(12, \theta)$, and the p-value for the observed $x = 9$ is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{12} \binom{12}{x} 0.5^x (1 - 0.5)^{(12-x)} = .0730\,. \tag{1}$$

Scientist 2's analysis: She had planned to take observations until observing 3 failures. Thus, for her, x has a Negative-binomial$(3, \theta)$ distribution, and the p-value is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{\infty} \binom{x+2}{x} 0.5^x (1 - 0.5)^3 = .0338\,. \tag{2}$$

Thus the two scientists had different *stopping rules*, but these were just thoughts in their heads; these thoughts had no effect on the experiment that was actually conducted or the data that was obtained. The SRP says that such thoughts should not matter.

Example 4. Thoughts in heads 2: A scientist comes to a statistician with 100 observations, assumed to be independent and from a N$(\theta, 1)$ distribution and desires to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The data average is $\bar{x}_n = 0.2$, so the standardized test statistic is $z = |\sqrt{n}\bar{x}_n - 0| = 2$. It is tempting to simply conclude that there is significant evidence against H_0 at the 0.05 level, but a careful classical statistician should ask the scientist "Why did you cease experimentation after 100 observations?"; in other words "what was your stopping rule?" If the scientist replies, "I decided to take an initial batch of 100 observations," there would seem to be no problem, but the words *initial batch* should give pause. Indeed, one should then (from a classical perspective) ask the followup question "What would you have

done had the first 100 observations not yielded significance?" Suppose the scientist replies: "I would then have taken another batch of 100 observations." This reply does not completely specify a stopping rule, but the scientist might agree that he was implicitly considering a procedure of the form:

- take 100 observations;
- if $\sqrt{100}\ \bar{x}_{100} > k$ then stop and reject H_0,
- but if $\sqrt{100}\ \bar{x}_{100} < k$ then take another 100 observations and reject if $\sqrt{200}\ \bar{x}_{200} > k$.

For this procedure to have level $\alpha = 0.05$, k must be chosen to be 2.18 (Pocock, 1977). Since the actual data had $\sqrt{100}\ \bar{x}_{100} = 2 < 2.18$, the scientist would not be able to conclude significance, and hence would have to take the next 100 observations. Again, the conclusion is being affected simply by thoughts in the scientist's head.

This can be carried further. Suppose the scientist duly obtains another 100 observations and brings the data back to the statistician. Suppose $\sqrt{200}\ \bar{x}_{200} = 2.2 > 2.18$ so significance has apparently been obtained. But, again, the statistician should ask the scientist what would have happened if the result had not been significant. (The statistician should have really asked this question earlier, completely nailing down the stopping rule that was in use from the beginning, but the story is more amusing this way.) Suppose the scientist says, "If my grant renewal is approved, I would then take another 100 observations but, if the grant is rejected, I would have no more funds and would have to stop the experiment." The advice of the classical statistician must then be: "We cannot make a conclusion until we find out the outcome of your grant renewal; if it is not renewed, you can claim significant evidence against H_0 while, if it is renewed, you cannot claim significance and must take another 100 observations."

More general than the SRP is the Likehood Principle (LP), which roughly states that one must always condition on only the actual data at hand.

Likelihood Principle (LP): *The information about $\boldsymbol{\theta}$, arising from just E and $\boldsymbol{x}_{obs}$, is contained in the observed likelihood function $L(\boldsymbol{\theta}) = f(\boldsymbol{x}_{obs} \mid \boldsymbol{\theta})$. Furthermore, if two observed likelihood functions are proportional (for the same $\boldsymbol{\theta}$), then they contain the same information about $\boldsymbol{\theta}$.*

Example 3 continued. For Scientist 1 the observed likelihood function was

$$\mathcal{L}_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3\,;$$

for Scientist 2 it was

$$\mathcal{L}_2(\theta) = \binom{11}{9}\theta^9(1-\theta)^3\,.$$

The first part of the LP states that one should look no further than these functions to perform inference; in particular the involvement of other possible data in (1) and (2) violates the LP. Furthermore, since $\mathcal{L}_1(\theta) \propto \mathcal{L}_2(\theta)$, the LP also says that the evidence about θ from either viewpoint is the same and, hence, that the two scientists should not have arrived at different conclusions.

The LP became prominent with the remarkable paper Birnbaum (1962), which deduced the LP as a logical consequence of the *conditionality principle* (essentially the Cox example, that one should base inference on the measuring instrument actually used) and the *sufficiency principle*, which states that a sufficient statistic for θ in E contains all information about θ that is available from the experiment. At the time of Birnbaum's paper, almost everyone agreed with the conditionality principle and the sufficiency principle, but did not agree with the LP; so it was a shock that the LP is a direct consequence of the other two principles. There are numerous qualifications relevant to the LP, and various generalizations and implications. (One such implication is the SRP, since 'stopping rules' affect $L(\theta)$ only by multiplicative constants). Many of these (and the history of the LP) are summarized in Berger and Wolpert (1988).

4 Other counterexamples relevant to classical statistics

The list of 'counterexamples' in classical statistics is enormous, but many are just examples where a particular method (such as maximum liklihood estimation) fails. Such examples are useful in reminding us of the limitations of the various methods, but they do not impact philosophical understanding; the examples in this section do have such impact.

Example 5. Shrinkage estimation: Independently, $x_i \sim \mathrm{N}(\theta_i, 1)$, $i = 1, 2, \ldots, p$, and it is desired to estimate $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ with an estimator $\boldsymbol{\delta}(\boldsymbol{x}) = (\delta_1(\boldsymbol{x}), \delta_2(\boldsymbol{x}), \ldots, \delta_p(\boldsymbol{x}))$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, under the expected loss

$$E[L(\boldsymbol{\delta}, \boldsymbol{\theta})] = E\left[\sum_{i=1}^{p} (\delta_i(\boldsymbol{x}) - \theta_i)^2\right],$$

where the expectation is over the distribution of $\boldsymbol{x}$.

Until 1961, the estimator $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x}$ was almost universally considered to be fine. It is the maximum likelihood estimator, the best unbiased estimator, the fiducial estimator, the invariant estimator and the objective Bayesian estimator. James and Stein (1961) shocked statistics by showing that the estimator

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

has smaller expected loss than $\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{x}$ if $p \geq 3$, with significantly smaller expected loss near $\boldsymbol{\theta} = \mathbf{0}$.

Besides upending standard thinking, this led to philosophical quagmires, such as the following, which has not really been resolved to this day. Suppose you are a statistician working for a company, and are given three problems to solve by different divisions in the company:

- observe $X_1 \sim N(\theta_1, 1)$ and estimate θ_1, a measure of the reliability of a production process,
- observe $X_2 \sim N(\theta_2, 1)$ and estimate θ_2, a measure of the health of employees,
- observe $X_3 \sim N(\theta_3, 1)$ and estimate θ_3, a measure of the financial return in one department.

You consult with superiors and these problems are judged to be of equal importance to the company, and sum of squares error loss is appropriate. Should you use the James-Stein estimator to analyze these problems, i.e. involve all of the x_i in the estimation of each θ_j?

While unintended, this conundrum gave considerable impetus to the hierarchical Bayes movement (see Example 12), because hiearchical Bayes is all about shrinkage for problems that are related (not completely disparate problems as above).

Example 6: Conditioning in testing: Suppose it is desired to test $H_0 : \theta = -1$ versus $H_1 : \theta = 1$, based on $x \sim \mathrm{N}(\theta, \frac{1}{4})$. The rejection region $x \geq 0$ results in a test with error probabilities (type I and type II) of 0.0228. If $x = 0$ is observed, the classical testing conclusion would be that H_0 is rejected, and that the error probability is $\alpha = 0.0228$. (Alternatively, one could state that the p-value is 0.0228.) Common sense, however, says that $x = 0$ completely fails to discriminate between the two hypotheses, since the observation is equally likely to arise from either hypothesis.

On the other hand, suppose that $x = 1$ is observed. Then, in classical frequentist testing, one can still only claim that the error in rejection is $\alpha = 0.0228$, even though $x = 1$ is four standard deviations from H_0, which seemingly implies much stronger evidence against H_0. (One could, alternatively report that the p-value is 0.000032, but we will see that this is very misleadingly small.)

Example 7. Psychokinesis and the Jeffreys-Lindley paradox: The Jeffreys-Lindley Paradox (Jeffreys (1961); Lindley (1957)) is one of the most famous counterexamples in statistics (although it is not at all clear whether it is a counterexample against frequentist or Bayesian statistics or both). I know of only one practical problem where the paradox has manifested itself, so it will be fun to begin with that example.

Jahn et al. (1987) reported an experiment involving a search for a psychokinetic effect and analyzed the data with p-values as indicated below, finding highly significant evidence of an effect. This was reanalyzed using objective Bayesian testing in Jefferys (1990) (objective in the sense that the existence and nonexistence of psychokinesis were both given prior probability of $\frac{1}{2}$) and showed strong evidence for the null hypothesis of no psychokinetic effect. This is the essence of the Jeffreys-Lindley paradox, that classical and Bayesian testing can yield dramatically different conclusions.

The experiment involved generating particles that passed through a quantum gate, with the particles emerging as either type 0 or type 1; quantum theory says that each should happen with probability 1/2. Experimental subjects were asked to try to mentally affect the outcome (presumably through some type of psychokinetic effect), either increasing the number of 0's or the number of 1's. Thus we have Bernoulli(θ) random variables, where θ is the probability of getting a 0, and are testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. For convenience in the analysis here, we will assume the Bernoulli trials are independent, although dependent analyses were also carried out in the above papers.

The Jeffreys-Lindley paradox happens as the sample size $n \to \infty$ and the sample size in this experiment was $n = 104,490,000$, qualifying as almost infinite. The number of $1's$ was $x = 52,263,471$, so that the proportion of $1's$ was $\hat{\theta} = 0.5000177$. The two-sided p-value for this data is $p = 0.0003$, leading the original authors to declare that there was highly significant evidence in favor of there being a psychokinetic effect.

The objective Bayesian analysis that was performed in Jefferys (1990) assigned prior probabilities $Pr(H_0) = Pr(H_1) = 0.5$ to the hypotheses and a uniform prior $\pi(\theta) = 1$ as the density of θ under H_1. Assigning the alternative hypothesis such a large prior probability does seem odd, and a subjective Bayesian might well assign much less weight to the alternative but, to make the Jeffreys-Lindley point, equal probabilities is best. Choice of the uniform density for θ will be discussed later.

Letting $\mathrm{Bin}(x \mid \theta)$ denote the binomial density of x, the objective posterior probability of H_0 is

$$\Pr(H_0 \mid x) = \frac{\mathrm{Bin}(\mathrm{x} \mid \frac{1}{2})\Pr(\mathrm{H_0})}{\mathrm{Bin}(\mathrm{x} \mid \frac{1}{2})\Pr(\mathrm{H_0}) + \Pr(\mathrm{H_1}) \int_0^1 \mathrm{Bin}(\mathrm{x} \mid \theta)\pi(\theta)\mathrm{d}\theta} = 0.92\,,$$

after plugging in the value of x and utilizing the objective priors. Thus the objective Bayesian analysis indicates that the evidence is strong in favor of H_0, while the classical analysis indicates that the evidence is strong in favor of H_1, the essence of the Jeffreys-Lindley paradox.

Both conclusions cannot be right and so we must look for the source of the discrepancy; indeed, there are issues with both analyses. Looking first at the Bayesian analysis, one cannot reasonably give H_1 higher prior probability than 0.5, but the choice of a uniform prior on θ is

not really reasonable; if psychokinesis leading to θ near zero or one existed, we probably would have seen it long before. More reasonable would be to choose $\pi(\theta) = \text{Uniform}(0.5-r, 0.5+r)$ for some small r, indicating that we do not expect to see a huge effect. Here are some interesting values of r and the resulting posterior probabilities of H_0:

r	0	0.00011	0.00024	0.0020	0.25	0.5
$\Pr(H_0 \mid x, r)$	0.5	0.050	0.0063	0.050	0.86	0.92

Table 1: Possible choices of the Uniform$(0.5-r, 0.5+r)$ prior for θ and the resulting posterior probabilities of the null hypothesis.

- Note that, as $r \to 0$, $\Pr(H_0 \mid x, r) \to 0.5$, so that the priors with very small r do not provide evidence against the null.
- Bayesians feel that $P(H_0 \mid x, r) \leq 0.05$ is strong evidence against H_0 and this does happen for $r \in (0.00011, 0.0020)$. However, r must be specified before seeing the data (a subjective Bayesian analysis is being done, and one must pre-specify the prior in subjective Bayes) and this is a very small target interval to hit. So it is quite unlikely that a pre-chosen r would have led to strong evidence against H_0.
- The choice $r = 0.00024$ is the choice that gives the smallest value of $P(H_0 \mid x, r)$, and this would, indeed, correspond to very strong evidence against the null. Thus 0.0063 is the smallest error that it is possible to state in rejecting the null. That the p-value is $.0063/.0003 = 21$ times smaller, reveals how much a p-value can underestimate the actual error, a message reinforced by the following example.

Example 8. A counterexample to interpreting p-values as error rates: We already saw that p-values cannot be interpreted as error rates in the previous example, where the lowest possible error rate in rejection was 21 times larger than the p-value. Here is a more general 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). Thus consider testing this hypothesis versus $H_1 : p \sim \text{Beta}(1, \text{b})$, $b > 1$. Then, with B_{01} denoting the Bayes factor of H_0 to H_1 and $f(p \,|\, H_i)$ denoting the density of p under H_i,

$$B_{01} = \frac{f(p \,|\, H_0)}{f(p \,|\, H_1)} = \frac{1}{b(1-p)^{(b-1)}} \geq -e\, p\, \log(p) \quad \text{for} \quad p < e^{-1}\,. \tag{3}$$

This follows from calculus: simply minimize the Bayes factor over $b > 1$.

Note that p will virtually always have a decreasing density under H_1 (small p-values should have larger probability under H_1); hence the choice in Vovk (1993) of the decreasing

Beta(1, b), $b > 1$, class of priors. 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, providing considerable additional support for the lower bound in (3).

An analogous bound can be given on the conditional Type I frequentist error

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

Thus, for a reported p, compute the lower bound above and interpret it as the frequentist Type 1 error probability, conditional on a minimal partition that contains the data (see Berger et al. (1994) for a full explanation).

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.

Example 9. A counterexample to non-statisticians understanding p-values: A common retort to assertions such as those in Example 8 is that everyone knows that a p-value is not an error probability, and so the difference is irrelevant. An interesting survey, relevant to this retort, was conducted 50 years ago and reported in Diamond and Forrester (1983). We quote from the article to report the survey and the results.

"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 in the Cedars-Sinai Medical Center Division of Cardiology. Half of the physicians answered incorrectly, and all had difficulty distinguishing the subtle differences between the choices. These differences, however, are crucial to any physician who wishes to understand better the clinical impact of the medical literature."

In the article, there then ensued a very good and enlightening discussion (aimed at physicians) trying to explain the differences between p-values and Bayesian and frequentist error probabilities. At the end of this discussion came the statement: "The correct answer to our test question, then, is 3." Of course, the actual correct answer to the question is: The chances are less than 5% of having obtained the observed response *or any more extreme response* if the therapy is not effective. Thus even the survey designers, who were out to show that their colleagues did not understand p-values, themselves did not understand p-values (nor, presumably, did the reviewers of the article).

5 Counterexamples relevant to subjective Bayesianism

I am a proponent of subjective Bayesian analysis when it is feasible and, especially, when it is absolutely necessary, as in Andrews et al. (1993), which reports a massive subjective elicitation performed for unknowns in a problem where no data about the unknowns was available! I am, however, primarily an advocate of objective Bayesian analysis; here are some examples indicating why this is so.

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. The next example reports an actual study of this.

Example 10. Underestimating variances involving Cepheid variable stars. In Barnes III et al. (2003), astronomical data was analyzed with the goal of determining the distance to Cepheid variable stars. As is standard in astronomy, the observations $x_1, \ldots, x_n$ were assumed to be independent and distributed as $\mathrm{N}(x \,|\, \mu, \sigma_i^2)$, with the variances σ_i^2 being specified; *i.e.*, each observation has its own known variance, arising from extensive knowledge of the astronomical measuring instruments used and all the systematic errors arising from interference by the atmosphere. In processing the raw data (photon counts) to produce the x_i and σ_i^2, many unknowns are encountered, but the astronomers feel that they know the distributions of the unknowns and can use them to compute the final σ_i^2.

A small part of Barnes III et al. (2003) was devoted to studying the accuracy of these elicitations, 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 $\pi(\tau^2) = 1/\tau^2$. Unsurprisingly, the

posterior distribution of τ^2 was centered at about 2 in one study and around 4 in another, indicating that the elicited σ_i^2 were, indeed, on the order of three times too small.

These estimated variances arose from some of the most careful subjective elicitations in science, and yet they prominently underestimated the error. What is one to think about the many casual elicitations being done in subjective Bayesian analysis? (To be fair, good subjective Bayesian training points out this common problem, and encourages elicitors to inflate their variances.)

Example 11. Hidden (bad) impacts of proper multivariate prior distributions – priors for covariance matrices: In subjective Bayesian analysis, it is common to use conjugate prior distributions, both for the ease in eliciting the parameters of the distribution and for the ease of ensuing computations. But conjugate priors can have hidden features that are detrimental. Here is one example.

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 sufficient statistic for $\boldsymbol{\Sigma}$ is $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{x}_i \, \boldsymbol{x}_i'$. By far the most commonly used subjective prior for $\boldsymbol{\Sigma}$ is the Inverse Wishart prior, for subjectively specified a and b,

$$\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-a/2} \exp\{-\tfrac{1}{2}\mathrm{tr}[b\, \boldsymbol{\Sigma}^{-1}]\}. \tag{4}$$

To see the concern with these priors, 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.

Everything on the RHS of this equation seems fine, except for the term $\prod_{i<j}(d_i - d_j)$. Indeed, this term is near zero when any eigenvalues are close; it follows that the conjugate priors try to force apart the eigenvalues of the covariance matrix.

This behavior is contrary to usual prior beliefs. Often in modelling multivariate normal data, one is deciding between choosing an exchangeable covariance structure (and hence equal eigenvalues) or a more general structure. When one is deciding whether or not to assume equal eigenvalues, it seems clearly inappropriate to use a prior distribution that gives no weight to equal eigenvalues, instead forcing them apart. Using such priors can also have a detrimental effect on inference, as was shown in Berger et al. (2020)

Objective Bayesian analysis can expose problems like this and allow for development of better subjective priors. Indeed the problem with the eigenvalues was first found in Yang and

Berger (1994), during a search for the reference prior for $\boldsymbol{\Sigma}$. The 'fix' found therein (simply divide the prior by $\prod_{i<j}(d_i - d_j)$), was then used in Berger et al. (2020) for development of better subjective priors for covariance matrices. Use of the Inverse Wishart distribution as a prior for $\boldsymbol{\Sigma}$ has also been criticized by others; see, for instance, **?**.

6 A counterexample relevant to empirical Bayes

In Bayesian hierarchical modeling, there are usually unknown *hyperparameters*. Empirical Bayes analysis estimates these hyperparameters from the data, while hierarchical Bayesian analysis assigns them a prior distribution (usually objective) and performs a full Bayesian analysis. The following basic example illustrates why we have a strong preference for objective hierarchical Bayesian analysis. Note that the same issues apply to other approaches, such as random effects modeling, best linear unbiased prediction, variance component modeling, and multilevel modeling (which are just difference names for the same thing).

Example 12. Failure of empirical Bayes in a basic normal hierarchical model. For $i = 1, \ldots, p$, suppose $x_i \sim \mathrm{N}(\cdot \mid \mu_i, 1)$ and $\mu_i \sim \mathrm{N}(\cdot \mid \xi, \tau^2)$. The marginal density of x_i given (ξ, τ^2) is found by integrating out the μ_i from the joint density of the x_i and μ_i, resulting in $x_i \sim \mathrm{N}(\cdot \mid \xi, 1 + \tau^2)$. The marginal density for the full data, $\boldsymbol{x} = (x_1, \ldots, x_p)$, is then

$$m(\boldsymbol{x} \mid \xi, \tau^2) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi(1+\tau^2)}} e^{\left[-\frac{(x_i-\xi)^2}{2(1+\tau^2)}\right]} \propto \frac{1}{(1+\tau^2)^{p/2}} \exp\left\{-\frac{p(\bar{x}-\xi)^2 + s^2}{2(1+\tau^2)}\right\}, \tag{5}$$

where $\bar{x}$ is the mean of the x_i and $s^2 = \sum(x_i - \bar{x})^2$.

Empirical Bayes analysis proceeds by estimating the hyperparamters ξ and τ^2 from this marginal liklihood, usually using either maximum likelihood or unbiased estimation. The obvious estimate of ξ is $\hat{\xi} = \bar{x}$, which is both the mle and the unbiased estimate. For τ^2, the unbiased estimate can be shown to be $\hat{\tau}^2_U = [s^2/(p-1) - 1]$, and the mle is

$$\hat{\tau}^2_{mle} = \max\left\{0, \frac{s^2}{p} - 1\right\}$$

(replace ξ by $\bar{x}$ in (5) and then maximize the resulting expression over τ^2). The unbiased estimate of τ^2 has the unfortunate property that it can be negative, which would be rather ridiculous to report. Hence we focus on use of the mle, which has become the standard empirical Bayes estimate (although see Morris (1983)).

Even use of $\hat{\tau}^2_{mle}$ is problematical. This is particularly clear if $s^2/p < 1$, in which case the mle would be $\hat{\tau}^2_{mle} = 0$. While a value of $s^2/p < 1$ is somewhat unusual here (if, for

instance, $p = 5$ and $\tau^2 = 1$, then $\Pr(s^2/5 < 1) = 0.264$), it is quite common in complicated hierarchical models to have at least one mle variance estimate equal to 0.

The problem with $\hat{\tau}^2_{mle} = 0$ is most clearly seen by looking at the marginal likelihood for τ^2 in such a situation. This marginal likelihood is given by integrating (5) over ξ, yielding

$$p(s^2 \,|\, \tau^2) \propto (\tau^2 + 1)^{-(p-1)/2} \, \exp\left\{ - \frac{s^2}{2(\tau^2 + 1)} \right\} . \tag{6}$$

Figure 1 graphs this marginal likelihood in the 'borderline' situation when $p = 5$ and $s^2 = 5$. This marginal likelihood of τ^2 is mostly decreasing away from 0, but not quickly and clearly indicates that there is considerable uncertainty as to the value of τ^2, even though the mle was $\hat{\tau}^2_{mle} = 0$. (The mle for the integrated likelihood (6) is slightly bigger than 0, reflecting the folklore that, in doing empirical Bayes analysis, it is better to use mle's from marginal likelihoods.)

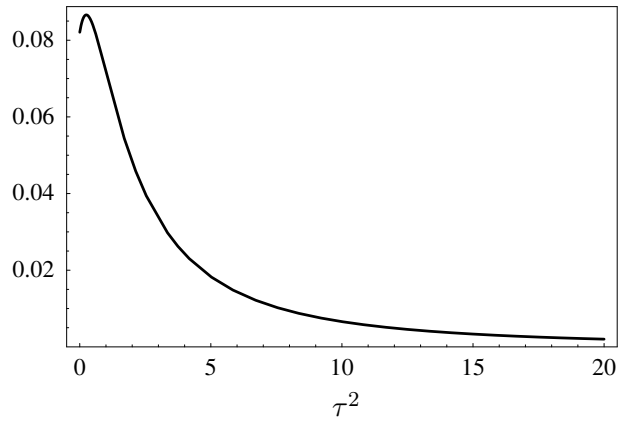

Figure 1: Marginal likelihood function of τ^2 when $p = 5$ and $s^2 = 5$ is observed.

It can be dangerous in statistical analysis to simply replace unknown parameters by their estimates, and this is particularly true in hierarchical settings. In the above example, for instance, setting τ^2 to 0 is equivalent to stating that *all the μ_i are exactly equal to each other.* This is clearly a terrible conclusion in light of the fact that there is actually great uncertainty about τ^2, as reflected in Figure 1. And, since 0 is at the boundary of the parameter space, it is also difficult to utilize likelihood or frequentist techniques to incorporate uncertainty about τ^2 into the analysis. Thus we strongly prefer full objective hierarchical Bayes analysis to empirical Bayesian analysis.

7 Counterexamples relevant to objective Bayes

Example 13. Counterexample to the view that use of improper priors is unsound. This is a counterexample to this commonly held view, but it would be more accurate to state that it is an argument for the validity of using certain (but not all) improper priors.

Reality is bounded, so one can argue that the real parameter space should, say, be some compact set $\boldsymbol{\Theta}_0$. Often, however, one only knows that the bounds are quite large, making it difficult to ascertain which $\boldsymbol{\Theta}_0$ to use. It is then tempting to pass to an unbounded space $\boldsymbol{\Theta}$ (if available) – that contains all the possible $\boldsymbol{\Theta}_0$ – and do the analysis there. This is justified if one can show that essentially the same answer is obtained using $\boldsymbol{\Theta}$, as from using any large compact $\boldsymbol{\Theta}_0$.

Thus consider a parametric model, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, an improper prior $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and an increasing compact sequence $\{\boldsymbol{\Theta}_i\}$ of subsets of the parameter space whose union is $\boldsymbol{\Theta}$ and for which the restricted priors $\pi_i(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) 1_{\boldsymbol{\Theta}_i}(\boldsymbol{\theta})$ are all proper. We seek to show that the corresponding sequence of restricted posteriors $\{\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ on $\{\boldsymbol{\Theta}_i\}$ converges to the unrestricted posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, which ensures that the restricted posteriors for large compact sets are close to the unrestricted posterior.

The method of convergence we consider is a version of Kullback-Liebler convergence. In particular we say that $\{\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ is *KL* convergent* to $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ if

$$\lim_{i \to \infty} \int \pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \log \frac{\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})} \, d\boldsymbol{x} = 0 \,.$$

Here is a lemma from Berger et al. (2009).

Lemma: $\{\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ is *KL* convergent* to $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ if $\int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) < \infty$ almost surely.

The condition in the Lemma is simply the condition that the formal posterior arising from the improper prior exists. Hence, one simply takes the improper prior, computes the posterior as one would with a proper prior and, if this posterior exists, all is well; the resulting posterior is known to be a good approximation to the restricted posterior arising from any large compact set. The Lemma should also alleviate concerns that something weird might happen when using Bayes formula with improper priors.

7.1 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

$$\pi^J(\boldsymbol{\theta}) = |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2} \,, \tag{7}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fisher information matrix. If the parameter is one-dimensional, this is a great objective prior and is viewed as the optimal prior from almost every objective Bayesian perspective.

The multiparameter case is a difference story; even Jeffreys himself did not like the outcome of using this prior for more than one parameter. For instance, if $\boldsymbol{\theta} = (\mu, \sigma)$, a normal mean and standard deviation, (7) gives $\pi^J(\mu, \sigma) = 1/\sigma^2$, whereas the standard objective prior is $\pi(\mu, \sigma) = 1/\sigma$; indeed Jeffreys preferred the latter, and it is called the 'independence Jeffreys prior.'

We know of no situation in which use of the Jeffreys-rule prior is optimal in multiparameter problems. The two counterexamples in this subsection indicate just how bad the use of the Jeffreys-rule prior can be.

Example 14. Inconsistency in the Neyman-Scott problem. Two observations are independently obtained from each of m normal distributions; the normal distributions have differing means μ_i but a common variance σ^2. Thus $\boldsymbol{x} = \{x_{ij}\}$, $i = 1, \ldots, m$, $j = 1, 2$, has density

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

Of interest is inference about the common variance σ^2. This problem was introduced by Neyman and Scott (1948), who showed that use of maximum likelihood estimation here results in an inconsistent estimate as $m \to \infty$. The counterexample has since become a test for all new methods of inference.

The Fisher information matrix can be shown to be

$$I(\mu_1, \ldots, \mu_m, \sigma^2) = \begin{pmatrix} 2/\sigma^2 & 0 & \ldots & 0 & 0 \\ 0 & 2/\sigma^2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 2/\sigma^2 & 0 \\ 0 & 0 & \ldots & 0 & m/\sigma^4 \end{pmatrix}, \tag{8}$$

so that the Jeffreys-rule prior in (7) is

$$\pi(\mu_1, \ldots, \mu_m, \sigma^2) = |I(\mu_1, \ldots, \mu_m, \sigma^2)|^{-1/2} = |m\, 2^m \sigma^{-(2m+4)}|^{-1/2} \propto \sigma^{-(m+2)}\,.$$

The posterior density for this prior can be shown to be

$$\pi^J(\mu_1, \ldots, \mu_m, \sigma^2 \,|\, \boldsymbol{x}) \propto \frac{1}{\sigma^{3m+2}} \exp\left\{ -\frac{1}{\sigma^2} \left[\frac{S^2}{2} + \textstyle\sum_{i=1}^{m} (\overline{x}_i - \mu_i)^2 \right] \right\}, \tag{9}$$

where $\overline{x}_i = (x_{i1} + x_{i2})/2$ and $S^2 = \sum_{i=1}^{m} \sum_{j=1}^{2} (x_{ij} - \overline{x}_i)^2$. Integrating out the nuisance parameters μ_i results in the marginal posterior density of σ^2 (the parameter of interest)

$$\pi^J(\sigma^2 \,|\, \boldsymbol{x}) \propto \frac{1}{\sigma^{2m+2}} \exp\left\{-\frac{S^2}{2\sigma^2}\right\},$$

which is recognizable as the InverseGamma$(m, S^2/2)$ distribution. This distribution has mean

$$\mathrm{E}[\sigma^2 \,|\, \boldsymbol{x}] = \frac{S^2}{(2m-2)}.$$

This estimator is inconsistent. The easiest way to establish this is to switch to the frequentist persective, letting σ_T^2 denote the true value of the variance and noting that the frequentist mean of S^2 is $\mathrm{E}^{\boldsymbol{x}}[S^2 \,|\, \sigma_T^2] = m\sigma_T^2$. By the law of large numbers, S^2/m thus converges to σ_T^2, as m increases. It follows that, for the Jeffreys-rule prior,

$$\lim_{m\to\infty} \mathrm{E}[\sigma^2 \,|\, \boldsymbol{x}] = \lim_{m\to\infty} \frac{S^2}{2m-2} = \frac{\sigma_T^2}{2},$$

which is only half the true value. It can also be shown that the posterior variance goes to 0 as m grows, so the posterior distribution concentrates, as is usual, but concentrates around a completely wrong value. The fact that the Jeffreys-rule prior gets worse and worse here, as the dimension grows, is a clear warning concerning its use for higher dimensional problems.

The correct objective prior to use here is the Jeffreys independence prior $\pi^{IJ}(\mu_1, \ldots, \mu_m, \sigma^2) \propto \sigma^{-2}$. This leads to a posterior density of σ^2 with mean $S^2/(m-2)$, which converges to σ_T^2 as m increases.

Example 15. Underdispersion in the Multinomial Problem. Suppose $\boldsymbol{x} = (x_1, \ldots, x_m)$ is Multinomial$(\boldsymbol{x} \,|\, n, \theta_1, \ldots, \theta_m)$ (suppressing the $(m+1)$st cell count, $x_{m+1} = n - \sum_{j=1}^{m} x_j$, and cell probability, $\theta_{m+1} = 1 - \sum_{j=1}^{m} \theta_j$, as they are determined by the other counts and probabilities) so that

$$p(\boldsymbol{x} \,|\, n, \theta_1, \ldots, \theta_m) = \frac{n!}{\prod_{j=1}^{m} x_j!(n - \Sigma x_j)!} \prod_{j=1}^{m} \theta_j^{x_j} (1 - \Sigma\theta_j)^{n - \Sigma x_j}.$$

Computation of the Fisher information matrix yields

$$I(\theta_1, \ldots, \theta_m) = \frac{n}{1 - \Sigma\theta_j} \begin{bmatrix} P\frac{1+\theta_1-\Sigma\theta_j}{\theta_1} & 1 & \ldots & 1 \\ 1 & \frac{1+\theta_2-\Sigma\theta_j}{\theta_2} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & \frac{1+\theta_m-\Sigma\theta_j}{\theta_m} \end{bmatrix}.$$

Computation of the determinant of this matrix yields

$$|I(\theta_1, \ldots, \theta_m)| = n^m \left[\left(1 - \textstyle\sum_{j=1}^{m} \theta_j\right) \prod_{j=1}^{m} \theta_j\right]^{-1}.$$

Thus, the Jeffreys-rule prior in (7) is

$$\pi^J(\theta_1, \ldots, \theta_m) \propto \left(1 - \sum_{j=1}^m \theta_j\right)^{-1/2} \prod_{j=1}^m \theta_j^{-1/2}, \tag{10}$$

which is recognizable as the (proper) Dirichlet$((\theta_1, \ldots, \theta_m) \,|\, (\frac{1}{2}, \ldots, \frac{1}{2}))$ distribution. Multiplying this by the multinomial likelihood shows that the corresponding posterior distribution is Dirichlet$((\theta_1, \ldots, \theta_m) \,|\, (x_1 + \frac{1}{2}, \ldots, x_m + \frac{1}{2}))$.

That this is a problematical posterior can be seen by considering the case where the sample size n is small relative to the number of classes $m+1$. As a specific example, suppose $n = 3$ and $m = 1000$, with $x_{240} = 2$, $x_{876} = 1$, and all the other $x_i = 0$. The posterior means resulting from using the Jeffreys-rule prior can be shown to be

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

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

The problem is that the Jeffreys-rule prior effectively added 1/2 to the 998 zero cells, making them – in concert – more important than the cells with data! That the Jeffreys-rule prior can encode much more information than is present in the data is not desirable for an objective analysis; a good objective prior needs to be much more disperse.

An alternative objective prior that is sometimes considered is the uniform prior, but this is even worse than the Jeffreys-rule prior since it adds 1 to each cell. If the total sample size n is large compared to the cell count $m+1$, either the Jeffreys-rule or uniform prior will yield more reasonable answers. But a good objective prior should be able to handle any data.

The prior that adds 0 to each cell is the improper prior $\prod_{j=1}^m \theta_j^{-1}$, but this cannot be used because it results in an improper posterior if any cell has a zero entry. The simplest (of several increasingly better) objective priors suggested in Berger et al. (2015) is the Dirichlet$((\theta_1, \ldots, \theta_m) \,|\, (\frac{1}{m}, \ldots, \frac{1}{m}))$ distribution. This only adds a total information of 1 through the prior, which is quite reasonable. The posterior means for the cells are then

$$\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = \frac{x_i + 1/m}{\sum_{j=1}^m [x_j + 1/m]} = \frac{x_i + 1/m}{n+1} = \frac{x_i + 1/1000}{4}.$$

The specific cell probabilities are then $\mathrm{E}[\theta_{240} \,|\, \boldsymbol{x}] = 0.500$ and $\mathrm{E}[\theta_{876} \,|\, \boldsymbol{x}] = 0.025$, with the cells having no observations yielding $\mathrm{E}[\theta_i \,|\, \boldsymbol{x}] = 0.00025$. These results seem much more reasonable than those that arose from the Jeffreys-rule prior. Also note that the recommended objective prior will work equally well when n is large.

7.2 Counterexamples in objective Bayesian testing

Example 16. The Bartlett counterexample to use of diffuse priors in testing. In estimation problems, objective priors can be diffuse and even improper. Problems involving testing and model uncertainty can sometimes utilize diffuse or improper priors, but more often need to utilize non-diffuse proper priors. Here is the basic version of the Bartlett (Bartlett, 1957) counterexample.

Suppose $x \sim \mathrm{N}(\cdot \,|\, \theta, 1)$, denoting the corresponding density $p(x \,|\, \theta)$. In estimation of the mean θ, it is fine to use the improper objective estimation prior $\pi(\theta) = c$, where c is any constant, since the posterior density of θ is

$$\pi(\theta \,|\, x) = \frac{p(x \,|\, \theta)\, c}{\int p(x \,|\, \theta)\, c \, d\theta} = p(x \mid \theta)\,,$$

i.e., the constants cancel.

Consider, instead, testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The Bayes factor, if $\pi(\theta) = c$ were used, is

$$B_{01}(c) = \frac{p(x \mid 0)}{\int_{-\infty}^{\infty} p(x \mid \theta)\, c \, d\theta} = \frac{p(x \mid 0)}{c}\,,$$

and, hence, depends on the arbitrary choice of c.

Even worse than use of improper priors here is use of *vague proper priors*, such as the Uniform$(-K, K)$ prior for θ, with K large, although many Bayesians erroneously view the use of vague proper priors to be better than the use of improper priors. But, for this prior, the Bayes factor becomes

$$B_{01}(K) = \frac{p(x \mid 0)}{\int_{-K}^{K} p(x \mid \theta)(2K)^{-1} d\theta} \approx \frac{2K\, p(x \mid 0)}{\int_{-\infty}^{\infty} p(x \mid \theta) d\theta} = 2K\, p(x \mid 0)\,,$$

which depends dramatically on the arbitrary choice of K. Indeed, the Bartlett paradox sends K to infinity (as is often done with vague proper priors), concluding that the Bayes factor infinitely favors the null hypothesis regardless of the data.

Example 17. A counterexample to the maximum a-posteriori model (MAP) being optimal. The Hald regression data set, that has been used by several authors (see Burnham and Anderson (1998) for references), 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. The models under consideration are all of the models defined by subsets of regressors, with the intercept being present in all models. For instance,

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

Table 3: Posterior model probabilities and corresponding excess predictive risks for the Hald regression example.

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
{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

Table 3 reports the results of a model uncertainty analysis using the encompassing AIBF approach of Berger and Pericchi (1996). The posterior probability of each model is reported, along with 'excess predictive risk', $\Delta R(M_i)$, which is the difference between the predictive risk of the model and the predictive risk of the optimal model averaged prediction, assuming squared error predictive loss.

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

$$\begin{aligned} p_1 &= \textstyle\sum_{j:x_1 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.95, \;\; p_2 = \sum_{j:x_2 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.73\,, \\ p_3 &= \textstyle\sum_{j:x_3 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.43, \;\; p_4 = \sum_{j:x_4 \in M_j} Pr(M_j \mid \boldsymbol{y}) = 0.55. \end{aligned} \tag{11}$$

Thus x_1 appears to be the most important regressor.

The maximum a-posterior model here is $\{1, 2\}$, but it is not the model with smallest excess predictive risk. That honor goes to model $\{1, 2, 4\}$, which is what is called the *median probability model*, defined as the model consisting of those regressors whose posterior inclusion probability is at least 1/2. Indeed, the median probability model is very often the optimal single predictive model; see Barbieri and Berger (2004) for conditions under which this is guaranteed to be so.

8 Counterexamples to the inadequacy of probability

There is an enormous literature claiming that ordinary probability theory is inadequate and needs to be augmented. Part of this literature is showing that people do not necessarily process probabilities well; while important to note, this does not mean we should replace

probability theory with something else. The following counterexamples are really counterexamples to counterexamples that attack probability theory; they thus defend probability theory and recommend only the most modest augmentations, to deal with issues involving precise versus imprecise probability.

Example 18. Counterexample to treating epistemic and aleatoric probabilities differently: *Aleatoric probability* is probability that arises from some random mechanism, while *epistemic probability* is probability used to describe uncertainty about some quantity that is not known, but is not random. Thus a subjective Bayesian assigning a prior distribution to an unknown – but fixed – quantity would be an example of epistemic probability.

Nearly everyone would agree with the use of probability to deal with random mechanisms, but many are uneasy with using probability to deal with epistemic uncertainty. For instance, in dealing with nuclear reactors, there are many fixed but unknown parameters that are constrained to lie in intervals; a Bayesian would usually assess a probability distribution over the intervals, whereas nuclear regulators usually do a worst-case analysis, based on seeing how the quantity of interest varies as the unknown parameters vary over the intervals.

One common misunderstanding that arises is due to use of too-simplistic probabilistic thinking. Consider the following two scenarios involving missile (or car airbag) production:

- *Scenario 1:* A production process for missiles randomly produces a faulty missile 10% of the time. So any particular missile has (aleatoric) probability of 0.1 of failing.
- *Scenario 2:* There is a 10% chance that the design of the missile is flawed, in which case all the missiles will fail. Any particular missile still has an (epistemic) probability of 0.1 of failing.

It is often argued in such situations that, while the probabilities are both 0.1, these are two very different situations that require one to think differently about the two types of probability. The problem, however, is in not using a thorough probabilistic analysis. In Scenario 1, each missile independently has probability 0.1 of failing; this also defines the Bernoulli joint probability distribution of all missile failures. In Scenario 2, however, the joint probability distribution of missile failures is that they will all fail with probability 0.1 and will all work with probability 0.9. These are very different joint probability distributions.

The two situations are also very different in terms of learning about reality. There is nothing to be learned in Scenario 1, while testing one missile in Scenario 2 will reveal whether all the missiles will work or all will fail.

Thus ordinary probability is capable of handling either aleatoric or epistemic probability. Whether or not one chooses to use aleatoric probabilities, as in the nuclear regulator situation, is a different question.

Example 19. Counterexample to the notion that Bayesian analysis requires a single prior distribution: This perception is reinforced by many of the axiomatic approaches to uncertainty, which state that a unique assessed probability distribution is needed to avoid problems like sure loss in betting. Perhaps unfortunately, this is not reality; probability distributions are themselves typically quite uncertain and there exist many efforts to deal with this uncertainty. Arguably the most useful formal approach is to attempt specification of the class of probability distributions that is compatible with beliefs or scientific understanding (such a class is often called a *credal set* – Levi (1980)), and then study the range of answers that follow from consideration of the class. This approach has many names, one of them being 'global robust Bayesian analysis' (Insua and Ruggeri (2000)). There is no space herein to discuss this in depth, but it is useful to present one example of this.

The Big Surprise: Suppose application of a statistical formalism leads to a predictive probability distribution $p(y)$ for reality $y > 0$ (perhaps mean climate temperature in 2040), but we assess that there is a 20% chance of the 'Big Surprise,' (e.g., that climate models are missing a big source of carbon sequestration that will kick in at higher temperatures). While this cannot be represented by a single probability distribution, it can be represented by the class of probability distributions

$$\mathcal{P} = \{0.2q(y) + 0.8p(y);\ q(y) \text{ being any distribution}\}\,.$$

There are many ways in which useful conclusions can be reached, even if only knowing $\mathcal{P}$. In a decision problem, for instance, one might find that a certain decision is fine for all distributions in $\mathcal{P}$. Or one can make potentially useful statements such as

$$E[y] \leq 0.8 \int yp(y)\ dy\,.$$

Example 20: Counterexample to the notion that Bayesian analysis cannot deal with imprecise probabilities. The Bayesian paradigm is typically phrased in terms of precise probabilities, e.g., the probability of rain tomorrow is 0.4 (i.e., $0.4000000000\ldots$). This is clearly not realistic. If a weather forecaster says 0.4, they perhaps mean that the probability of rain tomorrow is between 0.35 and 0.45. So one should rationally think in terms of eliciting intervals of probabilities.

The same is true for unknown input or calibration parameters. In climate models, for instance, there are many unknown parameters that are typically constrained to lie in intervals by the scientists. In high-energy physics, there are many unknown parameters, called 'systematic effects,' which are also generally constrained to lie in intervals. It is typical in both climate modeling and high-energy physics to deal with this problem by assigning uniform prior distributions to the unknowns over their intervals.

To illustrate these possibilities and provide a background for discussion, we consider the reliability example, in which a system contains m independent components, with component i having probability p_i of properly functioning over some time period of interest. Suppose that the system functions only if all components function, so the probability that the system functions is $P = \prod_{i=1}^{m} p_i$. A particularly important example of this occurs in the nuclear regulatory industry, where P is the probability of a reactor malfunction.

Suppose it is possible to restrict the p_i to intervals, namely $p_i \in (a_i, b_i)$, $i = 1, \ldots, m$, with the a_i and b_i specified. The classical analysis of this situation (used in the nuclear regulatory context for example) is to observe that, clearly, $P \in (\prod_i a_i, \prod_i b_i)$, i.e., the lower and upper endpoints of this interval are the worst-case and best-case scenarios. This conclusion is certainly a true statement but, unless m is very small or the intervals are very tight (billions of dollars are spent in the nuclear regulatory industry to make sure this is so), the range will typically be useless, a statement such as $P \in (0.4, 0.99)$.

In other disciplines (e.g., high-energy physics and climate modelling), the standard analysis is to assign uniform prior distributions to p_i in each interval. One then computes the density of P using probability theory, and reports, say, a 99% confidence set for P arising from this density. This will give a much smaller interval for P than using pointwise range analysis, e.g. $P \in (0.93, 0.98)$.

These are two extreme approaches; the first uses the crudest possible analysis, and the second specifies a single prior distribution for the unknowns. Robust Bayesian analysis can adjudicate between these extremes, by forming classes of prior probability distributions over the intervals, and then finding the range of answers corresponding to varying these priors over the classes.

In forming the class of priors, typically reasonable assumptions are that values of p_i near the midpoints of the intervals are more likely than values near the endpoints and that beliefs about the p_i are typically symmetric and unimodal in the intervals. One then considers all the probability distributions compatible with these beliefs, and finds the range of Bayesian answers over this class. Interestingly, it can be shown that the extremal 99% confidence interval over this class of prior probabilities happens to be exactly the same as that arising from use of the uniform priors over the intervals, providing strong support for the second approach above.

References

Andrews, R. W., Berger, J. O., and Smith, M. H. (1993). Bayesian estimation of fuel economy potential due to technology improvements. In *Case Studies in Bayesian Statistics*, pages

1–77. Springer.

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

Barnard, G. A. (1949). Statistical inference. *J. Roy. Statist. Soc. B*, 11:115–149.

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.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44:533–534.

Basu, D. (2011a). *An essay on the logical foundations of survey sampling, part one*. Springer.

Basu, D. (2011b). Learning statistics from counter examples: ancillary statistics. *Selected Works of Debabrata Basu*, pages 391–397.

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., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.*, 37:905–938.

Berger, J. O., Bernardo, J. M., and Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, 10:189–221 (with discussion).

Berger, J. O. and Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In *Statistical Decision Theory and Related Topics IV* **2** *(S. S. Gupta and J. O. Berger, eds). New York: Springer*, pages 29–72 (with discussion).

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. *Ann. Statist.*, 22:1787–1807.

Berger, J. O., Sun, D., and Song, C. (2020). Bayesian analysis of covariance matrix of multivariate normal distribution with a new class of priors. *Ann. Statist.*, 48:in press.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Hayward, CA: IMS, 2nd edition.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 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.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29:357–372.

Diamond, G. A. and Forrester, J. S. (1983). Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine*, 98(3):385–394.

Insua, D. R. and Ruggeri, F. (2000). *Robust Bayesian Analysis*. Springer.

Jahn, R. G., Dunne, B. J., and Nelson, R. D. (1987). Engineering anomalies research. *Journal of Scientific Exploration*, 1(1):21.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, pages 361–379.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration*, 4(2):153–169.

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

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44:187–192.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.*, 78:47–55.

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

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. 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. *Ann. Statist.*, 22:1195–1211.