Note on the Significance of 2x2 Contingency Tables and Lindley's Paradox

William H. Press

April 13, 2005

1 Introduction

There is a vast literature, both frequentist and Bayesian, devoted to the question of deciding whether a contingency table with moderate numbers of counts, say something like

	C_0	C_1
f_0	8	3
f_1	16	26
totals	24	29

(1)

shows a significant difference between a control sample C_0 and a diagnosed sample C_1 with respect to some feature or test with values f_0 and f_1.

In this note, we briefly review several frequently used methods, noting in particular the seemingly large discrepancies (previously noted by others) between the Bayesian and frequentist answers obtained. We then show that the discrepancies can be explained as an example of (so-called) Lindley's Paradox. Finally we suggest an "Ockham-removed prior", motivated within a fully Bayesian framework, that eliminates the discrepancies.

Table (1) is not chosen at random, by the way, but rather is a good example of the problem that we address. It has a p-value for the two-sided Fisher Exact Test (described below) of 0.0498; its two columns thus appear to be drawn from significantly different distributions ($p < 0.05$). A one-sided test (appropriate if the prior expectation was that f_0 would imply suppressed counts in C_1) would yield an even more significant result, by a factor of two. On the other hand, a Bayes-factor calculation, comparing a single probability model to one with independent probabilities for the two columns (with uniform priors for all probabilities) yields a Bayesian probability that the two columns are identically distributed of 0.3023; the two columns thus seem to be not significantly different. *Which is the correct answer?*

2 Standard Frequentist Approach

Agresti [1] has given an encyclopedic survey of the frequentist literature, which we will not attempt to repeat here. There is general agreement that moderate count values require so-called "exact" (as opposed to asymptotic) methods. One can quickly survey the web to determine what methods are most commonly recommended as "standard", and most commonly included in standard packages.

In brief, the standard approach for a table like (1) is:

1. Choose a test statistic that quantifies the discrepancy with the null hypothesis that there is no association between (C_0, C_1) and (f_0, f_1). Popular choices are the Wald statistic and Rao's efficient score statistic.

2. Choose a method, for example "Fisher's exact test" or "Barnard's exact test". The choice of a method is equivalent to choosing a distribution of 2×2 tables against which to compare table (1). The reason that this is not entirely straightforward is that the common probability of f_0 under the null hypothesis is not known. Different methods are, in effect, different estimations of this probability.

3. Compute the one- or two-sided p-value (as appropriate), that is, the probability of finding in the population defined by the method a value of the test as extreme as that seen.

4. Reject the null hypothesis (of no association) if $p < 0.05$ (say).

Within this paradigm, the single most popular (and therefore most standard) set of choices is probably Fisher's exact test with the Wald statistic. For a 2×2 table,

	C_0	C_1
f_0	m	n
f_1	$M - m$	$N - n$
totals	M	N

(2)

the Wald statistic is essentially the standardized difference of the observed probabilities under the null hypothesis,

$$T = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}(1-\widehat{p})(M^{-1} + N^{-1})}} \tag{3}$$

where

$$\widehat{p_1} \equiv m/M, \quad \widehat{p_2} \equiv n/N, \quad \widehat{p} \equiv (m+n)/(M+N) \tag{4}$$

(We will be careful not to confuse the p, the common probability under the null hypothesis, with the notation p that occurs in a p-value test!)

Fisher's Exact Test compares the T statistic to the distribution of T of all tables with the same marginals, both column *and row*. For the example of table (1), say, this would be all tables of the form

	C_0	C_1
f_0	m	$11-m$
f_1	$24-m$	$18+m$
totals	24	29

(5)

Since all cell counts must be positive, there are only 12 such tables in this example. The probability of each table under the null hypothesis is the hypergeometric probability, here,

$$P(m) = \binom{24}{m}\binom{29}{11-m}\Big/\binom{53}{11}, \qquad 0 \le m \le 11 \tag{6}$$

Because of the small number of distinct tables in the population, only certain discrete p-values are possible, a widely noted fact that we will revisit below. Barnard's Exact Test, proposed two decades after Fisher, was one attempt to eliminate this artifact. See [5] for a pedagogical comparison of the two methods.

Any of these, or similar, methods is open to the usual criticism of p-tests, namely that different choices of statistic can give rather different tail probabilities for a given data set. In practice, the differences are rarely large, however. Since this issue is common to all p-tests, it is always swept under the rug.

More serious is the issue in Step 2 above, namely the (implicit) estimation of the common probability P under the null hypothesis. The problem with any such choice is that, in general, it will not be the result of a consistent estimator on the population from which the table was actually drawn. Thus the p-values finally obtained are not tail probabilities of the test statistic for the actual experiment. In no sense is it precisely true that the null hypothesis will be incorrectly rejected only 1 time in 20. This is a common Bayesian objection to many frequentist procedures, namely their reliance on assumed (*non-unique*) distributions of results that might have been seen, but in fact were not.

3 Bayesian Approaches

Agresti [2] has surveyed the Bayesian literature in a recent review. Because the use of a Bayesian methodology in analyzing contingency tables is uncommon, it is harder to identify a "standard" Bayesian approach. The most obvious and

straightforward approach, namely the use of Bayes factors, is rarely used. The reason for this avoidance seems to be precisely the issue that we raise (and resolve) in this note, namely the apparent large discrepancies between Bayes factor methods and other methods (both Bayesian and tail-test), always in the demoralizing sense that the Bayes factor is less powerful in disproving the null hypothesis of no association (i.e., less able to find significant associations).

Jeffreys, in later editions of his book [4], develops the Bayes factor method for 2×2 contingency tables, similarly to the calculation below. He gives two numerical examples, one of which yields a probability (of the null hypothesis) 0.27, the other a respectable 0.0058. What Jeffreys does not mention is that Fisher's Exact Test, applied to the same data, gives probabilities of 0.057 and 0.00053, respectively. Good [3], working a series of examples, notes the discrepancies between tail area probabilities and Bayes Factors, and attempts, with very limited success, to find an empirical relation between the two. (This note can be viewed as a more principled approach to Good's program.)

Suppose H is the (null) hypothesis that the columns are identically distributed with $\mathrm{prob}(f_0) = p$, while H' is the alternative hypothesis that the columns have different probabilities $\mathrm{prob}(f_0|C_{0,1}) = p_{0,1}$. Then the Bayes factor is (see, e.g., [7])

$$\begin{aligned}\frac{\mathrm{prob}(H|D)}{\mathrm{prob}(H'|D)} &= \frac{\mathrm{prob}(D|H)}{\mathrm{prob}(D|H')} \times \frac{\mathrm{prob}(H)}{\mathrm{prob}(H')} \\ &= \frac{\int \mathrm{prob}(D,p|H)\ dp}{\int \mathrm{prob}(D,p_1,p_2|H')\ dp_1 dp_2} \times \frac{\mathrm{prob}(H)}{\mathrm{prob}(H')} \\ &= \frac{\int \mathrm{prob}(D|H,p)\ \mathrm{prob}(p|H)\ dp}{\iint \mathrm{prob}(D|H',p_1,p_2)\ \mathrm{prob}(p_1,p_2|H')\ dp_1 dp_2} \times \frac{\mathrm{prob}(H)}{\mathrm{prob}(H')}\end{aligned} \tag{7}$$

From the binomial distribution, we have

$$\begin{aligned}\mathrm{prob}(D,p|H) &= \binom{M}{m} p^m (1-p)^{M-m} \binom{N}{n} p^n (1-p)^{N-n} \\ \mathrm{prob}(D|H',p_1,p_2) &= \binom{M}{m} {p_1}^m (1-p_1)^{M-m} \binom{N}{n} {p_2}^n (1-p_2)^{N-n}\end{aligned} \tag{8}$$

For now, we take the prior ratio on the hypotheses as unity, $\mathrm{prob}(H)/\mathrm{prob}(H') = 1$. While we might well assume uniform priors on p, p_1, and p_2 in $(0,1)$, a more general choice is to use the conjugate prior

$$\begin{aligned}\mathrm{prob}(p|H) &\propto p^{-\alpha}(1-p)^{-\alpha} \\ \mathrm{prob}(p_1,p_2|H') &\propto {p_1}^{-\alpha}(1-p_1)^{-\alpha} {p_2}^{-\alpha}(1-p_2)^{-\alpha}\end{aligned} \tag{9}$$

where $0 \le \alpha < 1$. With these choices, equation (7) readily yields

$$F \equiv \frac{\mathrm{prob}(H|D)}{\mathrm{prob}(H'|D)} = \frac{B(m+n+1-2\alpha, M+N-m-n+1-2\alpha)}{B(m+1-\alpha, M-m+1-\alpha) B(n+1-\alpha, N-n+1-\alpha)} \tag{10}$$

where B is the beta function.

In equation (10), one may interpret $1-\alpha$ as a constant number of counts added by the prior to each of the observed counts m, n, $M-m$, and $N-n$. (This is a typical outcome of using conjugate priors.) We will generally take $\alpha = 1/2$, but our results are not sensitive to this choice.

We can now readily demonstrate what is the problem that Jeffreys ignored and Good puzzled over. We define a population of 2×2 contingency tables by the prescription:

- Choose M and N uniformly i.i.d. between 5 and 100.
- Choose m uniformly in $0 \ldots M$, and n uniformly in $0 \ldots N$.

For tables drawn randomly from this population we compute the two-sided Fisher's Exact Test (with the Wald statistic) p-value, and also the Bayesian probability of the null hypothesis from equation (10), that is, $F/(1+F)$. The result is shown in Figure 1. Evident is a strong tendency for the Bayes probability to lie at values > 0.1, making rejection of the null hypothesis impossible, even in cases where Fisher's test rejects the null hypothesis as strongly as 0.005.

While some more fervent Bayesians have rationalized this result as, somehow, a good thing – an intrinsic conservatism of the Bayes factor – most have instead substituted different Bayesian methods (as reviewed in [2]) without this flaw. A simple example is to use a Bayesian quantity like $\mathrm{prob}(p_1 > p_2)$ as a tail statistic (Good's so-called "Bayes/non-Bayes compromise"). Similar arguments to those leading to equation (10) give

$$\mathrm{prob}(p_1 > p_2) = \frac{\iint\limits_{p_1 > p_2} dp_1 dp_2 \; {p_1}^{m-\alpha}(1-p_1)^{M-m-\alpha}{p_2}^{n-\alpha}(1-p_2)^{N-n-\alpha}}{B(m+1-\alpha, M-m+1-\alpha)B(n+1-\alpha, N-n+1-\alpha)} \tag{11}$$

where the integrals must be done numerically for each set of $\{m, n, M, N, \alpha\}$. To get a two-sided tail probability to compare to two-tailed Fisher, we take the smaller of $\mathrm{prob}(p_1 > p_2)$ and $\mathrm{prob}(p_2 > p_1)$ and multiply it by 2.

Figure 2 shows the result. Much of the vertical dispersion can be understood as due to the discreteness of the Fisher Exact Test's p-values. Simply changing $\leq$ to $<$ in the definition of the Fisher test moves many points that lie above the diagonal to locations below the diagonal. With this caveat, it is fair to conclude from the Figure that the two tail tests are measuring essentially the same property of the contingency tables.

The linearity and small dispersion of Figure 2 further suggests that there is nothing "wrong" with the Bayesian probabilities p_1 and p_2, and that the discrepancy shown in Figure 1 must lie in either the common probability p (which turns out not to be the case), or in the way that the Bayes factor compares p with p_1 and p_2 (which turns out to be precisely the case).

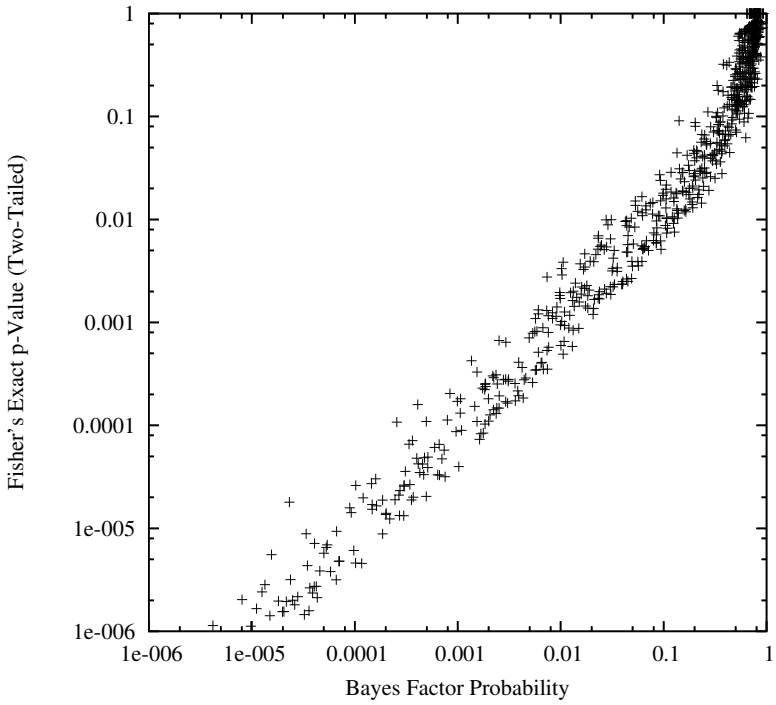

Figure 1: The Bayes factor probability of the null hypothesis $F/(1+F)$ is plotted against Fisher's Exact Test p-value (two-tailed) for a sample of 2×2 contingency tables. The Bayes factor often fails to reject the null hypothesis, even when it is strongly rejected by the p-value test.

4 Lindley's Paradox

Lindley's Paradox (see [6] for a review of the literature) is a name given to exactly the situation that we have just seen: Analysis based on Bayes factor odds ratios can award a high probability to a sharp null hypothesis, even when that hypothesis is easily rejected by a tail test.

The canonical example of Lindley's paradox is that of measuring a single normal variable y with known (small) σ, but unknown mean μ. The null hypothesis H is that μ has a certain value μ_0. The alternative hypothesis H' is

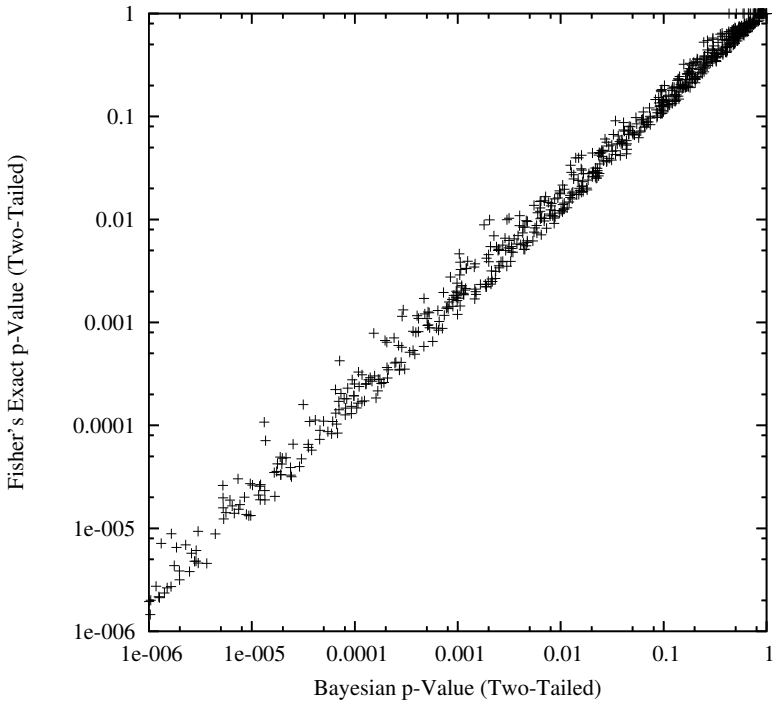

Figure 2: The Bayes tail probability p-values are plotted against Fisher's Exact Test p-values (two-tailed) for the same sample of 2×2 contingency tables. Some of the variance seen results from the discreteness of Fisher p-values.

that it has some other value. We then write (cf. equation 7),

$$\begin{aligned}
\frac{\text{prob}(H|D)}{\text{prob}(H'|D)} &= \frac{\text{prob}(D|H)}{\text{prob}(D|H')} \times \frac{\text{prob}(H)}{\text{prob}(H')} \\
&= \frac{\text{prob}(D|H)}{\int \text{prob}(D|H',\mu)\ \text{prob}(\mu|H')\ d\mu} \times \frac{\text{prob}(H)}{\text{prob}(H')} \\
&= \frac{(2\pi\sigma^2)^{-1/2} \exp(-[y-\mu_0]^2/[2\sigma^2])}{\int (2\pi\sigma^2)^{-1/2} \exp(-[y-\mu]^2/[2\sigma^2]) \text{prob}(\mu|H')\ d\mu} \times \frac{\text{prob}(H)}{\text{prob}(H')} \\
&\approx \frac{(2\pi\sigma^2)^{-1/2} \exp(-[y-\mu_0]^2/[2\sigma^2])}{\text{prob}(y|H')} \times \frac{\text{prob}(H)}{\text{prob}(H')}
\end{aligned} \tag{12}$$

where the approximation is that, for small σ, the Gaussian in the denominator approximates a Dirac delta function.

Attention now focusses on the surviving prior on y in the denominator. The broader we make our prior on y, the smaller $\text{prob}(y|H')$ becomes, and the more the null hypothesis is favored. The so-called paradox is that as we try to take the limit of complete ignorance of y – seemingly allowing it to have an

arbitrarily large range – we simultaneously make it impossible to disprove the null hypothesis that it has the value y_0.

Lindley's paradox is also a good example of "it's not a bug, it's a feature!". If we take

$$\mathrm{prob}(y|H') = \frac{1}{y_{\max} - y_{\min}}, \qquad y_{\min} < y < y_{\max} \tag{13}$$

(and zero elsewhere), then many Bayesians (e.g., [7]) would rewrite the last line of equation (12) as

$$\frac{\mathrm{prob}(H|D)}{\mathrm{prob}(H'|D)} \approx \exp(-[y-\mu_0]^2/[2\sigma^2]) \times \frac{y_{\max} - y_{\min}}{(2\pi)^{-1/2}\sigma} \times \frac{\mathrm{prob}(H)}{\mathrm{prob}(H')} \tag{14}$$

They would then identify the second factor on the right as a so-called "Ockham factor", by which any *arbitrary* new parameter added to a theory ought to be penalized, so as to avoid the overfitting of data. Indeed, the automatic emergence of Ockham factors in Bayesian calculations is taken as a strength, not a weakness, of the formalism. Ockham factors play, in a Bayesian context, the role that Bonferroni corrections play in a frequentist context: both serve to discount the significance of inferences made from multiple hypotheses.

Notice that the third factor in equation (14) is also a prior, namely the prior odds ratio between H and H'. This is often taken as unity, meaning that there is no reason to prefer H over H' *a priori*. *But is unity actually the correct "neutral" prior?*

5 Use of Ockham-Removed Priors to Resolve the Paradox

The perspective of this note is that there is *less* here than meets the eye; that Lindley's paradox results simply from a confusion between two conceptually different uses of added parameters; and that the Bayesian framework already provides the means for disentangling this confusion.

When we add a parameter μ to a model in order to fit the data better, we hope for a narrow posterior probability for its values, which we will likely summarize as a value and uncertainty. We are disappointed if the posterior is broad and uninformative. It is of no particular consequence if the posterior "notches out" (i.e., eliminates) any particular value μ_0 of the new parameter.

On the other hand, when we add a parameter to a model specifically as as a foil for a null hypothesis value μ_0, then the situation is completely reversed: We are not bothered if the posterior on μ is broad, and we are not particularly interested in its value if it is narrow. Rather, we hope for a clear "notch" on μ_0, such that that particular value can be rejected.

The Bayes factor formalism provides the means for distinguishing between these two different situations, by allowing us to choose different interpretations for the third factor, the overall prior odds ratio, in equation (14). In the case of the first situation, the interpretation given above is appropriate: Unity prior

odds ratio means neutrality on whether to add a fitting parameter. Indeed, a slight reworking of equations (12) and (14) would yield the *Bayes Information Criterion* (BIC) as an indicator of whether an additional model parameter is favored. The "automatic" Ockham factor is entirely appropriate in this situation.

In the case of the second situation, however, there is no reason for us to be slaves to the previous meaning of the overall prior. Rather, knowing that our interest is in disproving a null hypothesis, we are free to choose a prior that corrects for (i.e., "undoes") the Ockham factor. In other words, the truly neutral prior for this situation is the inverse of the Ockham factor, favoring the alternative hypothesis. The distinction is between *parameter fitting* (with a possibly variable number of parameters), on the one hand, and *significance testing* on the other.

The population of 2×2 contingency tables defined above provides a nice test of our claims. The only complication is that there is a parameter p in the null hypothesis, and two parameters $p_{0,1}$ in the alternative hypothesis. Thus our neutral prior for significance testing will be the ratio of the two (estimated) Ockham's factors.

For the null hypothesis, the Ockham factor K_0 is estimated as the range of p (that is, unity) divided by an estimate of the uncertainty in p,

$$K_0 \approx \frac{1}{\sqrt{\hat{p}(1-\hat{p})/(M+N)}} \tag{15}$$

where

$$\hat{p} \equiv (m+n)/(M+N) \tag{16}$$

For the alternative hypothesis, the Ockham factor K_1 is estimated as the area of the unit square, divided by the product of the uncertainties of p_0 and p_1.

$$K_1 \approx \frac{1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\hat{p}_1(1-\hat{p}_1)/(MN)}} \tag{17}$$

where

$$\hat{p}_0 \equiv m/M, \qquad \hat{p}_1 \equiv n/N \tag{18}$$

The neutral prior is thus taken as

$$\frac{\text{prob}(H)}{\text{prob}(H')} = \frac{K_0}{K_1} \tag{19}$$

Figures 3 and 4 show the comparison between the Bayes factor probability with Ockham-removed prior and either the two-tailed Fisher Exact Test (Figure 3) or, from equation (11), the Bayesian tail probability (two-tailed). The latter figure is most enlightening, since it does not have the discreteness artifacts of Fisher's test.

With the Ockham-removed prior, there is a very tight agreement between the tail probability and the Bayes factor probability, extending from small p-values all the way up to almost unity. The curvature near $p = 1$ is readily

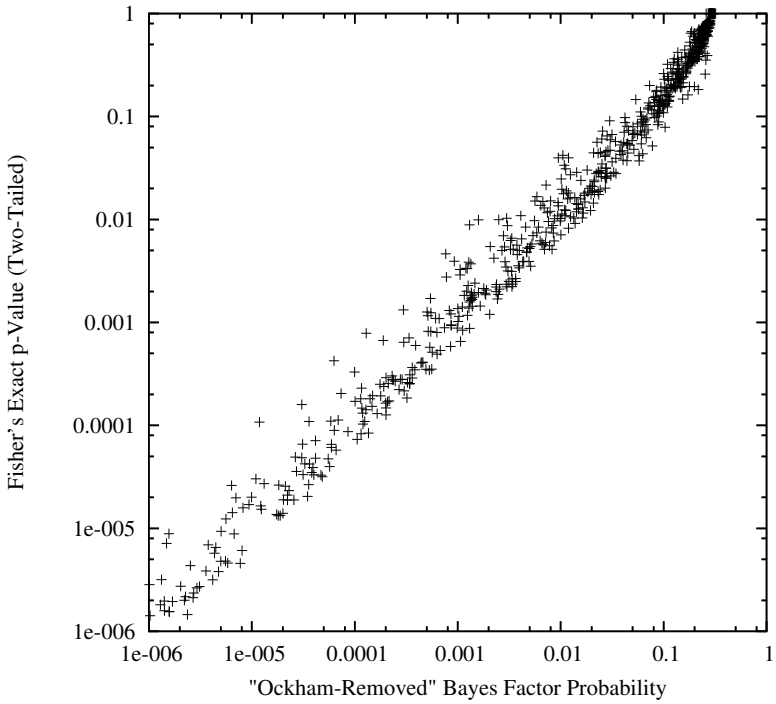

Figure 3: Use of an Ockham-removed prior (equation 19) brings the Bayes factor method into good agreement with Fisher's Exact p-value. The discrepancy near $p = 1$ is because any two-tailed test has the value 1 when its tails "cross", while the Bayes factor method never assigns probability 1 to the null hypothesis.

explained as an artifact of tail tests: As the tail probability increases, it eventually crosses 50% (so that the two-tailed probability reaches unity), at which point the definition of the tails is reversed. Thus there will be a substantial population very near unity. For the Bayes factor probablity, on the other hand, unity probability for the null hypothesis is a limiting case that is never reached.

6 Conclusions

When applied to contingency tables, Bayes factor methods have long been known to support the null hypothesis (of no association), even when tail tests strongly indicate otherwise. This tendency is an example of Lindley's Paradox, and is due to the so-called Ockham factor that naturally arises in Bayes factor methods.

Ockham factors are appropriate in parameter-fitting applications, as safeguards against overfitting. In such applications, they are closely related to the Bayes Information Criterion (BIC) for deciding whether to add a new parame-

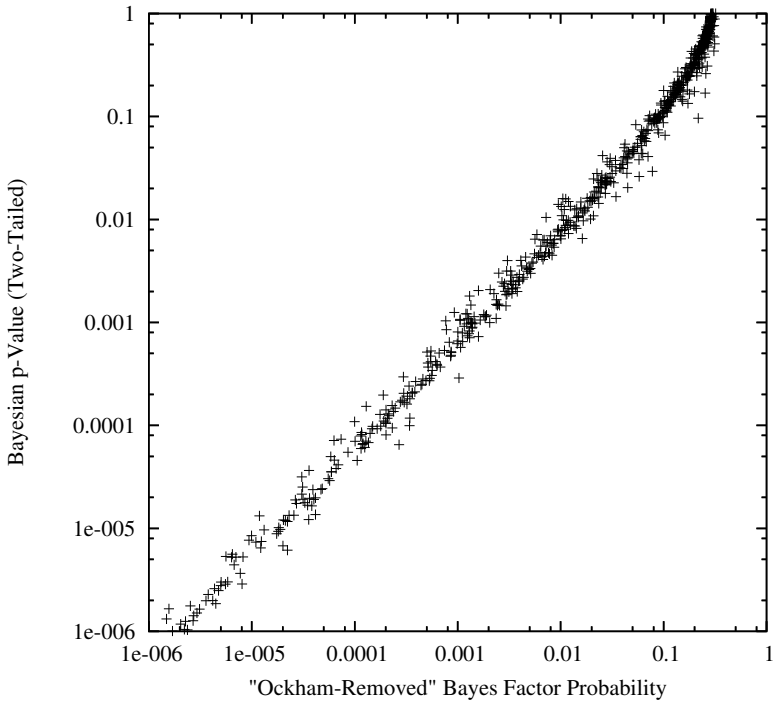

Figure 4: Same as Figure 3, but using the Bayesian tail probability (equation 11) instead of the Fisher test, thus eliminating the variance due to the latter's discreteness. As in Figure 3, the curvature at the upper left is an artifact of the tail test when its tails cross.

ter.

However, Ockham factors are not appropriate when an added parameter is simply a nuisance "foil" against which the significance of a null hypothesis is to be tested. In such a case, the "neutral" prior odds ratio is not unity, but is rather the inverse of the Ockham factor.

If we use an "Ockham-removed" prior, then tail tests and the the Bayes factor method give very nearly identical results. We should not expect *exactly* identical results, even on average: On the frequentist side, the choice of a different tail statistic will give different results. On the Bayesian side, our estimation of the Ockham factor is only approximate, and is open to discussion at the level of factors close to unity.

7 Acknowledgments

I thank Rick Picard for pointing me to the literature on Lindley's Paradox, and Dave Higdon for useful discussions.

References

[1] Agresti, A. (1992) "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, vol. 7, pp. 131-177.

[2] Agresti, A. (2004?) "Bayesian Inference for Categorical Data Analysis: A Survey", at www.stat.ufl.edu/∼aa/cda/bayes.pdf.

[3] Good, I.J. (1967) "A Bayesian Significance Test for Multinomial Distributions," *J. Roy. Stat. Soc., Ser. B*, vol. 29, pp. 399-431.

[4] Jeffreys, H. 1961, *Theory of Probability*, 3rd ed. (Cambridge: Cambridge University Press).

[5] Mehta, C.R. and Senchaudhuri, P. (2003) "Conditional versus Unconditional Exact Tests for Comparing Two Binomials", at www.cytel.com/Papers/twobinomials.pdf

[6] Shafer, G. 1982, "Lindley's Paradox", *J. Am. Stat. Assoc.*, vol. 77, pp. 325-334; also several commentaries *loc. cit.*

[7] Sivia, D.S. (1996) *Data Analysis: A Bayesian Tutorial* (Oxford: Clarendon Press).