Math. Scientist **36,** 117–125 (2011)
Printed in England

UNDERSTANDING AND USING FISHER'S *p*. PART 2: A REFERENCE BAYESIAN HYPOTHESIS TEST

K. R. W. BREWER * AND
GENEVIEVE HAYES,** *Australian National University*

Abstract

In Part 2 of this article, a reference Bayesian hypothesis test is constructed that corresponds exactly to the single-parameter ABIC_1 of Part 1. An important role is played here by a hitherto rather neglected (and initially purely empirical) law of numbers (see Benford (1938)) that had first been propounded in 1881 (see Newcomb (1881)). This hypothesis test is then extended to small samples, where another important role is played by the p-statistic; this time in setting an upper bound to the false discovery rate, regardless of the number of degrees of freedom involved.

Keywords: Augmented Bayesian information criterion, Bayesian hypothesis test, Benford's law, complete ignorance

2010 Mathematics Subject Classification: Primary 62A01
Secondary 62C10

1. Towards a reference Bayesian hypothesis test

The history of attempts to construct a reference Bayesian hypothesis test has been long and somewhat painful. An important turning point came in 1995 when Lavine and Wolpert recognised that attempts to do so using Bayes factors were fundamentally unsound (see Lavine and Wolpert (1995) which is part of O'Hagan (1995)). A few years later Brewer proposed the use of the reference posterior odds in place of Bayes factors and of Lebesgue-type measures in the place of probabilities Brewer (1999), (2002), but it was not until a further important result presented in Part 1 of this article (Brewer and Hayes (2011) in *Math. Scientist* **36**) became available, namely that an effective Bayesian information criterion would necessarily be a function only of the T-statistic, that it became relatively easy to construct a useful reference Bayesian hypothesis test.

The Lebesgue-type measures mentioned above behave exactly like probabilities, except that they are not required either to sum or to integrate to unity. Using these, the role previously played by the unstable Bayes factor can be taken over by the stable reference Bayesian posterior odds Brewer (1999), (2002).

The hypothesis test about to be developed in this part of the article is approximately Bayesian whenever the observation is significant or close to significance, and asymptotically Bayesian as $|T|$ increases. It is also a conservative test, always understating the extent to which the observation is significant, and one that reflects exactly the behaviour of the single-parameter augmented Bayesian information criterion or ABIC_1, defined in Equation (4.3) of Part 1 (Brewer

Received 21 September 2010; revision received 21 December 2010.
* Postal address: School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Canberra, ACT 0200, Australia. Email address: ken.brewer@anu.edu.au
** Email address: ghayes17@gmail.com

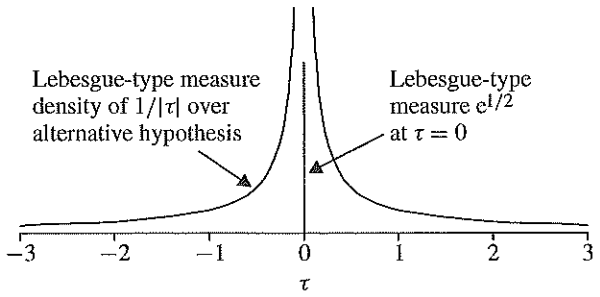

FIGURE 1: The rectangularly hyperbolic shape of complete ignorance for a reference prior.

and Hayes (2011)) as

$$\mathrm{ABIC}_1 = T^2 - \ln T^2 - 1. \tag{1.1}$$

This ABIC_1 formula also determines the augmented Bayesian false discovery rate (ABFDR), defined in Part 1 as

$$\mathrm{ABFDR} = \frac{1}{1 + \exp(\mathrm{ABIC}_1/2)}. \tag{1.2}$$

The approximately and asymptotically Bayesian hypothesis test that has the properties stipulated above can be specified as follows. The null hypothesis H_0 is located precisely and uniquely at $\tau = 0$, where τ is the number of sample standard deviations between that H_0 location and the true size of the relevant effect. The prior Lebesgue-type measure for H_0 at $\tau = 0$ is chosen to have the finite measure $\exp(\frac{1}{2})$. The alternative hypothesis H_1 exists over the entire real line, with the trivial exception of $\tau = 0$. Its measure density over this real line is $|1/\tau|$, which corresponds to the two rectangular hyperbolae in Figure 1. The relevant approximation, however, consists in regarding the whole of the H_1 measure as concentrated at $\tau = T$ and as experiencing the diminution that corresponds to that location.

The manner in which this test has been made to correspond exactly to the ABFDR of (1.2) is indicated by Table 1. Row (1) of this table displays the values of $\exp(-T^2/2)$,

TABLE 1: Constructing the ABFDRs for the integer values of $|T|$ from 0 to 4. (*Except where $|T| = 0$, which is the H_0 location. Note also that the entire posterior measure of H_1 is treated as located at the observed value of T.)

			Observed values of $\lvert T\rvert$				
	Row	Formula	0	1	2	3	4
H_0 likelihood at $\tau = T$	(1)	$\exp(-T^2/2)$	1.000 000	0.606 531	0.135 335	0.011 109	0.000 335
H_0 prior measure at $\tau = 0$	(2)	$\exp(\frac{1}{2})$	1.648 721	1.648 721	1.648 721	1.648 721	1.648 721
H_0 posterior measure at $\tau = T$	(3)	(1) × (2)	1.648 721	1.000 000	0.223 130	0.018 316	0.000 553
H_1 posterior measure at $\tau = T$	(4)	$1/\lvert T\rvert$*	0	1.000 000	0.500 000	0.333 333	0.250 000
Sum of those last two measures	(5)	(3) + (4)	1.648 721	2.000 000	0.723 130	0.351 649	0.250 553
ABFDR	(6)	(3) ÷ (5)*	0	0.500 000	0.308 561	0.052 086	0.002 207

which is the reciprocal of the likelihood ratio L, and reflects the 'bell-shaped curve' of the normal distribution. Row (2) holds only the constant $\exp(\frac{1}{2})$, the Lebesgue-type measure at the H_0 location $\tau = 0$. The product of these two rows, the $\exp((1 - T^2)/2)$ in row (3), reflects the posterior measure for H_0 at $\tau = T$. Row (4) reflects the rectangularly hyperbolic manner in which the corresponding H_1 measure density diminishes, namely as $1/|T|$ over the entire range of H_1.

To obtain the $ABIC_1$'s ABFDR value in row (6), however, it is necessary to regard the *whole* of H_1's measure in row (4) as concentrated at the location $\tau = T$, and as experiencing the diminution $1/|T|$ corresponding to that location. Further, since H_1's total prior measure is infinite, while the H_0 measure at $\tau = 0$ is finite, the prior probability attributable to H_1 is 'almost one', and that attributable to H_0 is infinitesimally small. However, once even a single observation is made, the posterior measure attributable to H_1 becomes finite, and also if at any stage T happens to be observed to be equal to one, the two hypotheses contain equal measures, so their probabilities become equal, each at 0.5, as also happens with the $ABIC_1$.

The correspondence between this hypothesis test and the ABFDR of Part 1 is therefore close, but in order to justify the manner in which the original shape of the finite prior measure density over H_1 was chosen, we also need to explain the nature and relevance of Benford's (initially empirical) law of numbers. Since that law, which can be seen to imply such a shape precisely, is less well known than it deserves, we next digress to introduce it and to exhibit some of its implications.

2. Benford's law of numbers

In the late 19th century the Canadian–American astronomer Newcomb noticed that the early pages in his printed tables of logarithms were more heavily used than the later pages (Newcomb (1881)). On closer inspection he found empirically that the proportions of numbers having the leading digit m could be described by the formula $\log_{10}(m + 1) - \log_{10} m$, and found a rough proof. Those proportions are displayed as percentages in Table 2.

The percentages shown in Table 2 are, however, somewhat counterintuitive. Sceptical readers might like to take their daily newspaper and record all the leading digits they can find in it (excluding those of the page numbers, which are not random, because they follow a predictable sequence). They would then see how well or otherwise those random digits were following the pattern indicated in Table 2. Such a natural scepticism may, however, explain why Newcomb's

TABLE 2: Percentages of leading digits found in random observations.

Leading digit	Percentages of occurrences
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%
Total	100.0%

TABLE 3: Relative frequencies of leading digits found in random observations.

Leading digit	Relative frequencies of leading digits
Small values	← 58 etc.
0.8	51
0.9	46
1	30.1
2	17.6
3	12.5
4	9.7
5	7.9
6	6.7
7	5.8
8	5.1
9	4.6
10	3.01
20	1.76
Large values	1.25 etc. →

findings were virtually ignored for half a century, until his American compatriot Benford carried out a much more exhaustive investigation of his own (Benford (1938)).

These days, auditors use Benford's law to detect fraudulent uses of data, and it is just possible that they might have been over-reluctant to advertise its use too openly! However, in the current situation, where misunderstandings of the significance of Fisher's p-statistic demand our attention, the validity and relevance of Benford's law are too important for it to remain in obscurity. Table 2 indicates clearly that, in naturally random populations, small numbers predominate over large numbers. This fact has had a bearing on our choice of the measures used to describe prior ignorance, a choice to which we now return.

While the percentages in Table 2 relate only to leading digits combined over all 'denary cycles' (such as the ranges between 1 and 10, 10 and 100, or 0.01 and 0.1) it is possible to define a complete ignorance prior by treating those percentages as applying within each denary cycle individually. It is also possible to visualize this possibility by extending Table 2 vertically – see Table 3.

The complete ignorance prior corresponding to what we can now describe as 'Benford's law of numbers (extended)', or BLNE, thus consists of the two rectangular hyperbolae in Figure 1, one on either side of the null hypothesis location, $\tau = 0$. Further, the manner in which the original H_1 prior, $1/|\tau|$, corresponds to the extended version of Benford's law of numbers (Benford (1938)) can therefore be stated as follows: that the penalty term $-\ln T^2$ (or equivalently $-2\ln|T|$) in (1.1) corresponds to the prior measure density of the alternative hypothesis (H_1) falling off on either side of $\tau = 0$ as $|\tau|^{-1}$ (the integral of any x^{-1} being $\ln x$ plus an arbitrary constant, here set equal to zero).

It might be useful here to recall the particular instance in which we earlier used the $ABIC_1$ equivalent to this test. In our second example of Part 1 of this article (Brewer and Hayes (2011), Section 3), the number of heads thrown exceeded the expected number by 8000. Moreover, the standard deviation of that observed or realized difference was 2000, so $|T| = \frac{8000}{2000} = 4$. Substituting 4 for the $|T|$ of (1.1) we obtained $R = \exp(ABIC_1/2) = \exp((T^2-1)/2))/|T| = \exp(7.5)/4 \approx 452$. This was a quite credible result in itself, as well as being appreciably

more parsimonious than the AIC_1's $R = \exp(\frac{14}{2}) \approx 1097$ (where AIC denotes the Akaike information criterion), and far less parsimonious than the BIC_1's $R \approx \exp(-\frac{0.5881}{2}) \approx 0.745$ (where BIC dentoes the Bayesian information criterion). It can now additionally be viewed as the consequence of an approximately and asymptotically Bayesian hypothesis test.

3. An extension to the case of small samples

So far, we have only been considering the case where the sample size on which the significance test was based was large. When it is only of a comparatively modest size, the T-statistic must be replaced by Student's t-statistic. The number of degrees of freedom, ν, upon which that t-statistic is based, is also relevant.

We next need to find a small sample formula for the ABIC_1 corresponding to that for the large-sample (1.1). The distribution of Student's t-statistic is

$$f(t) = \frac{A_\nu}{(2\pi)^{1/2}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

where $A_\nu = (2/\nu)^{1/2}\Gamma((\nu+1)/2)/\Gamma(\nu/2)$ is an expression that tends to 1 asymptotically from below as $\nu \to \infty$. (Since A_ν is greater than 0.9 from $\nu = 3$ upwards, it can be ignored for practical purposes, so it has not been used either in (3.1) or in the construction of Table 4, below.) A small sample formula that corresponds to (1.1) and tends to it asymptotically as $\nu \to \infty$ is therefore

$$\begin{aligned}\mathrm{ABIC}_{1t} &= (\nu+1)\ln\left(1 + \frac{t^2}{\nu}\right) - \ln t^2 - (\nu+1)\ln\left(1 + \frac{1}{\nu}\right) \\ &= \ln\left(1 + \frac{t^2}{\nu}\right)^{\nu+1} - \ln t^2 - \ln\left(1 + \frac{1}{\nu}\right)^{\nu+1}. \qquad (3.1)\end{aligned}$$

Equation (1.1) supplies the requirement that when $|T| = 1$, ABIC_1 takes the value 0. Equation (3.1) indicates that the same value is approached asymptotically as $\nu \to \infty$, since $(1 + (x/\nu))^{\nu+1} \to \exp x$ for all x. In other words, when the observation made is exactly or asymptotically a single estimated standard deviation away from the null hypothesis value, the information criterion indicates either an exact equivalence or an asymptotic approach to indifference between H_0 and H_1, a desirable outcome in its own right.

Further, the first term on the right in (3.1) tends to $\ln(\exp(t^2))$, or simply to t^2, and consequently also to T^2, as $\nu \to \infty$. Similarly, the third term tends to -1. The second term obviously tends to $-\ln T^2$.

Table 4 displays the values of the ABFDR corresponding to the ABIC_{1t} values defined by (3.1). It will shortly be used to show that Fisher's two-tailed p has a valuable role to fill as a conservative statistic which, *when suitably transformed*, supplies an upper bound to the ABFDR. To demonstrate this, we need first to consider the behaviour of the ABFDR itself when the sample size is small.

Table 4 indicates that if any value of two-tailed p is calculated using the formula relevant to Student's t with the appropriate number of degrees of freedom, it will always result in a smaller value of the ABFDR than would have been the case had the ABFDR been calculated on the assumption that the relevant statistic was actually T, or, equivalently, that the relevant number of degrees of freedom was infinite.

For example, if we are in the situation where there are 29 degrees of freedom and we make an observation which corresponds to $p = 0.01$ or 1%, Table 4 indicates that the ABFDR is

TABLE 4: Reference Bayesian ABFDRs for given values of two-tailed p and degrees of freedom.

Two-tailed p	$\nu = 1$	$\nu = 2$	$\nu = 4$	$\nu = 9$	$\nu = 29$	$\nu = 99$	$\nu = \infty$
25%	0.4142	0.4602	0.4804	0.4893	0.4931	0.4942	0.4946
10%	0.2361	0.3076	0.3583	0.3885	0.4050	0.4100	0.4121
5%	0.1353	0.1940	0.2486	0.2875	0.3107	0.3182	0.3213
2%	0.0508	0.0916	0.1316	0.1674	0.1919	0.2004	0.2040
1%	0.0305	0.0487	0.0747	0.1019	0.1226	0.1301	0.1334
0.2%	6.24×10^{-3}	0.0103	0.0172	0.0268	0.0359	0.0396	0.0412
0.1%	3.13×10^{-3}	5.16×10^{-3}	8.84×10^{-3}	0.0144	0.0201	0.0225	0.0236
0.02%	6.28×10^{-4}	1.04×10^{-3}	1.82×10^{-3}	3.20×10^{-3}	4.88×10^{-3}	5.67×10^{-3}	6.05×10^{-3}
0.01%	3.14×10^{-4}	5.19×10^{-4}	9.18×10^{-4}	1.65×10^{-3}	2.61×10^{-3}	3.08×10^{-3}	3.30×10^{-3}
0.002%	6.28×10^{-5}	1.04×10^{-4}	1.85×10^{-4}	3.50×10^{-4}	5.94×10^{-4}	7.24×10^{-4}	7.89×10^{-4}
0.001%	3.14×10^{-5}	5.20×10^{-5}	9.28×10^{-5}	1.78×10^{-4}	3.11×10^{-4}	3.85×10^{-4}	4.22×10^{-4}

TABLE 5: $ABIC_1$ comparing the ABFDRs with published lower bounds to the FDRs.

Two-tailed p	$\|T\|$	ABFDR	Berger–Delampady lower FDR bounds
10%	1.645	0.4121	0.3916
5%	1.960	0.3213	0.2904
1%	2.576	0.1334	0.1093
0.1%	3.291	0.0236	0.0179

0.1226, which (as we have already indicated) is far from being significant. If, however, we had treated the observation as coming from a large sample, but still used the same p-value, we would have read off the ABFDR as 0.1334, which is even further away from being significant. So if a claim of significance is valid using the large sample ABFDR, that claim will remain valid even if, later, the small sample size is taken into account.

There is a choice here. On the one hand there is a reward (and a substantial one if ν is small) in calculating the ABFDR using the correct value of ν. In Table 4, we have taken advantage of this option and calculated the relevant ABFDR values accordingly. On the other hand, however, if it is substantially more convenient to use only the values of p that would have been relevant had ν been infinite, or if only those values are accessible, the ABFDR remains as an upper bound to the true FDR. It is even a reasonably good approximation as well, if ν is appreciable (say 29 or more), and more particularly so if p is not unduly small (say, $p > 0.1\%$).

4. A comparison between the ABFDRs and Berger and Delampady's lower bounds

The ABFDR values displayed in Table 4 are not only upper bounds, but are also reasonably close to the lower bounds obtained by Berger and Delampady (1987), especially for marginal and submarginal levels of significance. Berger and Delampady had used unimodal symmetric normal distributions to define their alternative prior distributions. This allowed them to arrive at lower bounds for the false discovery rates (FDRs). Their largest attained values are shown in the last column of Table 5. The ABFDRs of Table 4, used also in Table 5, follow from the complete ignorance priors of Figure 1.

5. Summary of conclusions for Part 2

It is possible to form a Bayesian hypothesis test that corresponds exactly to the single-parameter $ABIC_1$ of Part 1. However, to do this, it is first necessary to discard the notion that the appropriate statistics to be obtained from these tests need to be Bayes *factors*. The reference posterior odds of Equation (4.1) in Part 1, which converge to nonzero finite limits as the prior probability that the null hypothesis is true tends to zero, are more useful.

To allow those reference posterior odds to converge to finite limits, it was found to be convenient to use measures that resembled probabilities in every respect, except for the requirement that they sum or integrate to unity. The prior measure for the null hypothesis (H_0) was then a finite one, and the prior measure *density* for the alternative hypothesis (H_1) was also finite, but since it integrated to an infinite measure, it enabled H_0 to be assigned a finite prior measure, which could then also be interpreted as implying an infinitesimal but nevertheless meaningful prior probability.

Additionally, once even a single observation had been made, the posterior measure for H_1 also became finite, and this implied finite posterior probabilities for both hypotheses.

The precise formulation required to ensure that this hypothesis test corresponded exactly with the ABIC_1 also required that the Bayesian prior correspond to an extended form of Benford's (originally empirical) law of numbers, which extension in turn corresponded to 'complete ignorance'. We consider that these multiple correspondences point to something more than mere coincidence, and in Part 3 of this article (Brewer *et al.* (2012)) we shall be providing empirical evidence that supports this conclusion.

Appendix A.

The result obtained in Appendix A of Part 1 (that in the large sample case the ABIC_D is intermediate between its corresponding BIC_D and AIC_D values for almost all practical purposes) can to a large extent be extended to the small sample situation. Specifically, an information criterion can be formulated which is for nearly all practical purposes *asymptotically* intermediate between the BIC_D and the AIC_D as the number of degrees of freedom, ν, increases.

Since $(1 + (x/\nu))^{\nu} \to \mathrm{e}^x$ for all x and $t \to T$ as $\nu \to \infty$, ABIC_{1t} also tends as $\nu \to \infty$ to $T^2 - \ln T^2 - 1$, which is the same expression for the ABIC_1 as has already appeared in Equations (4.3) and (A.1) of Part 1 and (1.1).

Equation (3.1) further requires that when $|t| = 1$, $\mathrm{ABIC}_{1t} \approx 0$. This implies that when the observation is one estimated standard deviation away from the null hypothesis value, the small sample information criterion approximates indifference between the null and alternative hypotheses, a useful outcome in its own right.

Finally we can generalize the ABIC_{1t} of (3.1) into ABIC_{Dt}. In (3.1) itself, the first term is the small sample single-parameter case of $2 \ln L$, and the other two terms make up the penalty. The generalization of the first of those two penalty terms, $-\ln t^2$, to the multiparameter case, can be seen to be $D \ln((2 \ln L)/D)$, in line with the generalization of the corresponding penalty term in Appendix A of Part 1. Hence a complete multiparameter formula for ABIC_{Dt} may be written as

$$\mathrm{ABIC}_{Dt} = 2 \ln L - D \ln\left(\frac{2 \ln L}{D}\right) - D \ln\left(1 + \frac{1}{\nu}\right)^{\nu+1}. \tag{A.1}$$

As $\nu \to \infty$, the first two terms in (A.1) remain at $2 \ln L$ and at $-D \ln((2 \ln L)/D)$ respectively, and the remaining term tends to $-D$; hence ABIC_{Dt} tends asymptotically to the ABIC_D of Equation (A.2) of Part 1.

Further, since this ABIC_D has already been shown to be intermediate between its corresponding BIC_D and AIC_D values for almost all practical purposes, it follows that the ABIC_{Dt} of (A.1) is also asymptotically intermediate between those same BIC_D and AIC_D values for almost all practical purposes. Finally, since it is generally accepted that the BIC is overly parsimonious and the AIC is insufficiently so, once again, this intermediate property is a useful one.

References

BENFORD, F. (1938). The law of anomalous numbers. *Proc. Amer. Philos. Soc.* **78,** 551–572.

BERGER, J. O. AND DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2,** 317–352.

BREWER, K. R. W. (1999). Testing a precise null hypothesis using reference posterior odds. *Rev. R. Acad. Cienc. Exact. Fis. Nat. (Esp.)* **93,** 303–310.

BREWER, K. R. W. (2002). The Lindley and Bartlett paradoxes. *Pakistan J. Statist.* **18,** 1–13.

BREWER, K. R. W. AND HAYES, G. (2011). Understanding and using Fisher's p. Part 1: Countering the p-statistic fallacy. *Math. Scientist* **36,** 107–116.

BREWER, K. R. W., HAYES, G. AND GILLISON, A. N. (2012). Understanding and using Fisher's p. Part 3: Examining an empirical data set. To appear in *Math. Scientist* **37**.

GILLISON, A. N. *et al.* (2010). An extensive biodiversity survey in neotropical and palaeotropical landscape mosaics. Available at http://www.cbmglobe.org/publications.htm.

HAYES, G. AND BREWER, K. R. W. (2012). Understanding and using Fisher's p. Part 4: Do we even need to specify a prior measure at H_0? To appear in *Math. Scientist* **37**.

LAVINE, M. AND WOLPERT, R. (1995). Comment in the discussion of O'Hagan (1995). *J. R. Statist. Soc. B* **57,** 132.

NEWCOMB, S. (1881). Note on the frequency of the different digits in natural numbers. *Amer. J. Math.* **4,** 39–40.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc. B* **57,** 99–138.

SELLKE, T., BAYARRI, M. J. AND BERGER, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statistician* **55,** 62–71.