ANOTHER POSTERIOR PARADOX

B. D. PUZA,∗ **
K. R. W. BREWER * AND
T. J. O’NEILL,* Australian National University

Abstract

An inference problem is presented in which the only datum is the sum of \( n \) independent and identically distributed normal random variables with mean \( \mu \), and where \( n \) itself has a known probability distribution. The choice of a flat prior for \( \mu \) over the whole real line, or even over a long finite interval, leads to a posterior for \( n \) which favours small values of that parameter. Consequently, the estimation of \( \mu \) is biased upwards. This paradox is resolved by adopting a proper prior for \( \mu \) that approximates \( f(\mu) \propto 1/|\mu| \) in its middle ranges. The chosen prior has some empirical support from Benford’s law of numbers. The connection with this law is discussed, as is the choice of a suitable prior in a number of related situations. This article warns of the possible dangers in blindly using a flat prior distribution to represent \textit{a priori} ignorance.

Keywords: Benford’s law; improper posterior; improper prior; noninformative prior; reference prior

2000 Mathematics Subject Classification: Primary 62F15

1. Introduction

When estimating the mean, \( \pi \), of a Bernoulli distribution, there are three prior distribution densities proportional to \( x^a(1-x)^a \) that have some claim to be regarded as reference priors. Setting \( a = -1 \) yields the improper Haldane prior, for which the resulting estimates of \( \pi \) are unbiased; a rarity in the context of Bayes estimation. Setting \( a = -0.5 \) yields the Jeffreys prior, which has certain desirable invariance properties, while setting \( a = 0 \) yields the uniform distribution over the range \([0, 1]\), which also has some intuitive appeal.

Although these three priors may look, and indeed are, very different, none of them contains much information. Indeed, if we start with a Haldane prior and observe a single 0 and a single 1, the resulting posterior is itself the uniform distribution; and it may be heuristically useful to think of the Jeffreys prior as arising from the observation of ‘half a 0 and half a 1’.

It is therefore not unreasonable to allow a statistical analyst to choose whichever of these three priors happens to be convenient, or better still to work in terms of a prior proportional to \( x^a(1-x)^a \) and allow the final user to choose the value of \( a \), if and when a choice becomes necessary.

But while this much latitude is frequently permitted when estimating the mean of a binomial distribution, no such freedom is commonly allowed in the estimation of the normal mean. (For simplicity, we will assume here that the corresponding variance is known.) On the contrary, it is often taken for granted that the improper uniform over the entire real line is the only appropriate
reference prior for that problem. However, there are conceivable circumstances in which that is not a useful reference prior. In Section 2 we introduce a thought experiment that exemplifies such a set of circumstances. In Section 3 we explain how a paradox then arises, which can seriously bias the estimation procedure. In Section 4 we show how it can be resolved using a different reference prior, and in Section 5 generalize the problem to a situation where \( n \) can take values greater than 2, and where its distribution is no longer known. In Section 6 we adduce Benford’s law of numbers as providing empirical support for that prior. Section 7 is concerned with the consequences of observing other possible transformations of the raw data. Section 8 contains a short summary of our findings.

2. The paradoxical experiment

Consider a random sample \( y_1, \ldots, y_n \) from a normal distribution with known variance \( \sigma^2 \) and a mean \( \mu \) that has a uniform prior distribution over the whole real line. Independently of that mean, the number of observations in the sample, \( n \), is decided by the toss of a fair coin (for each experiment separately) to be 1 for heads and 2 for tails. We are then informed only that the total of the observations \( y_1 + \cdots + y_n \) is \( \hat{y} \) (that is, we have a numerical value only for \( \hat{y} \)). What then is the posterior probability that \( n = 1 \)? Also, what bearing has that posterior probability on the estimation of \( \mu \)?

This specification may be summarised in mathematical form as follows:

\[
(y_1, \ldots, y_n | \mu, n) \sim \text{independent and identically distributed } N(\mu, \sigma^2),
\]

\[
f(\mu | n) \propto \begin{cases} 1, & -\infty < \mu < \infty, \\ \frac{1}{2}, & n = 1, 2, \end{cases}
\]

\[
y_1 + \cdots + y_n = \hat{y}, \quad \text{and only } \hat{y} \text{ and } \sigma^2 \text{ are known.}
\]

The only datum supplied in this context is \( \hat{y} \), and our primary task is to find \( f(n \mid \hat{y}) \), the posterior density of \( n \). Using the standard Bayesian algorithm, we obtain

\[
f(n \mid \hat{y}) \propto f(\hat{y}, n)
\]

\[
= \int f(\hat{y}, n, \mu) \, d\mu
\]

\[
= \int f(n) f(\mu | n) f(\hat{y} | \mu, n) \, d\mu
\]

\[
\propto \int 1 \cdot 1 \cdot f(\hat{y} | \mu, n) \, d\mu
\]

(\text{where the constant of proportionality does not depend on } n)

\[
= \frac{1}{n} \int_{-\infty}^{\hat{y}} \frac{1}{\sqrt{\pi} \sigma} \sqrt{2\pi} \exp\left\{- \frac{1}{2n\sigma^2} (\hat{y} - n\mu)^2 \right\} \, d\mu \quad \text{(since } (\hat{y} | n, \mu) \sim N(n\mu, n\sigma^2))
\]

\[
= \frac{1}{n} \int_{-\infty}^{\hat{y}} \frac{1}{\sqrt{\pi} \sigma / \sqrt{n}} \sqrt{2\pi} \exp\left\{- \frac{1}{2(\sigma^2/n)} \left( \mu - \frac{\hat{y}}{n} \right)^2 \right\} \, d\mu
\]

\[
= \frac{1}{n},
\]

(4)

because the value of the last integral is unity. So the posterior probability that \( n = 1 \) is \( f(n = 1 \mid \hat{y}) = \frac{1}{1 + \frac{1}{2}} \) or \( \frac{2}{3} \). This is counterintuitive, especially since it is not a function
Another posterior paradox

35 of \( \dot{y} \). No matter what value \( \dot{y} \) takes, and knowing full well that the choice of \( n \) was decided by a fair coin, the rules of Bayesian inference require us to affirm that the actual probability of heads on each toss was \( \frac{2}{3} \). Note that there are, in fact, two aspects to this paradox.

Firstly, the posterior density at (4) does not depend on the datum \( \dot{y} \), and so must be the same as the prior density at (3). But (3) and (4) are not the same functions of \( n \).

Secondly, the mere observation of \( \dot{y} \) makes \( n = 1 \) more likely than \( n = 2 \), regardless of the value of \( \dot{y} \).

Before proceeding further, it is pertinent to note that the problems being considered here do not appear in such sharp perspective only because the sample sizes involved are small. However large the potential sample sizes, say \( n = 500 \) and \( n = 1000 \) for each experiment, instead of only \( n = 1 \) and \( n = 2 \), the posterior probabilities for these two possibilities would still be \( \frac{2}{3} \) for the smaller sample and \( \frac{1}{2} \) for the larger.

Also, if the ratio between the potential sample sizes were more extreme, then so would be that of the paradox. For instance, if the potential sample sizes were \( n = 1 \) and \( n = 99 \), then the posterior probabilities would be \( \frac{99}{100} \) for the smaller sample and only \( \frac{1}{100} \) for the larger.

Finally, it is a simple matter to generalise the randomising of the sample size from only two possibilities to any finite number, provided that the experimenter knows the objective probabilities that have been assigned to them. Two cases where there is a finitely large number of such possibilities, including the case where the experimenter does not know the objective probabilities that have been assigned to them, are considered in Section 5.

3. Explaining the paradox

We might have guessed initially that the posterior probability for \( n = 1 \) would be the same as for the prior, namely \( \frac{1}{2} \), but as we have just seen, a rigorous Bayesian analysis yields the posterior probability \( \frac{2}{3} \), regardless of the value of \( \dot{y} \). The reason for this is that ‘almost all’ the prior probability is located at an infinitely large distance on either side of zero, so that all possible finite values of \( \dot{y} \) are automatically regarded as very small, and therefore more likely to be the value of a single observation than the sum of two.

Given this misappropriation of the posterior probabilities between \( n = 1 \) and \( n = 2 \), the estimation of \( \mu \) can be seriously biased upwards. In the original experiment, the overestimation is only by a factor of \( \frac{10}{9} \), but had the initial choice been between \( n = 1 \) and \( n = 99 \), each with the objective probability \( \frac{1}{2} \), the overestimation would have been by a factor of almost 2.

4. Resolving the paradox

So far we have a satisfactory explanation of the paradox, but not a successful resolution of it. It is not acceptable to have the posterior probability of \( n = 1 \) to be even approximately \( \frac{2}{3} \) for any and every possible value of \( \dot{y} \), and hence for every experiment. A successful resolution of the paradox would need to ensure that the posterior probability of \( n = 1 \) averaged out to \( \frac{1}{2} \) over all possible experiments.

One way of achieving this result would simply be to ensure that the prior distribution was proper. If it were flat over the range \( -c < \mu < c \), for any value of \( c \) much greater than \( \sigma \), then for values of \( |\dot{y}| \) close to or greater than \( |2c| \) the probability that \( n = 1 \) would be almost zero. However, when the averaging is done over all experiments, it is standard Bayesian practice to do so over all possible samples drawn from the prior distribution. It is only this kind of
averaging that would produce the desired result, namely that the expectations of the posterior probabilities for \( n = 1 \) and for \( n = 2 \) would each be close to \( \frac{1}{2} \).

This is not the same as averaging out over all possible samples drawn from the actual population. To get the posterior probabilities to be \( \frac{1}{2} \), both for \( n = 1 \) and for \( n = 2 \) in those circumstances, the true population values of \( \mu \) would need to be distributed exactly in accordance with the assumed prior, that is, they would themselves need to be distributed uniformly over the interval \(-c < \hat{y} < c\). This is a very strong assumption, and in almost all cases quite unrealistic. If \( c \) is chosen to be very large, so that nearly all of the true population values are much closer to 0 than to \( c \) or \(-c\), then the original paradox remains, with the posterior distributions of \( n = 1 \) being close to \( \frac{2}{3} \) for the great majority of potential replications of the experiment.

It follows that for a satisfactory resolution of the paradox it would be necessary to use a prior for which the mean, \( \mu \), was distributed reasonably in accordance with its actual distribution. We have already ruled out the choices of a flat prior over the whole real line and of a flat prior over any very large finite range, such as \(-c < \mu < c\) with \( c \geq 10^{100} \). It would seem that we are left only with a choice between two strategies. Either we must adopt a strictly subjectivist approach and demand that the prior on \( \mu \) reflect the experimenter’s own beliefs, or we must find a new reference prior distribution for the unknown \( \mu \) that would ensure that the posterior probabilities for \( n = 1 \) and for \( n = 2 \) are each close to \( \frac{1}{2} \). In order to do the latter rigorously we need to choose a prior distribution for \( \mu \) that ensures \( f(n \mid \hat{y}) \approx f(n) \) for all \( \hat{y} \).

Fortunately, this is not difficult. Under the model given by (1), (2) and (3), the posterior density of \( \mu \) given \( \hat{y} \) is provided by

\[
f(\mu \mid \hat{y}) = \frac{2}{3} \text{N}\left(\mu; \frac{\hat{y}}{1}, \sigma^2 \right) + \frac{1}{3} \text{N}\left(\mu; \frac{\hat{y}}{2}, \sigma^2 \right),
\]

where \( \text{N}(t; a, b^2) \) denotes the density at \( t \) of the normal distribution with mean \( a \) and standard deviation \( b \).

This distribution is a mixture of two normal distributions, where one has weight \( \frac{2}{3} \), mean \( \hat{y}/1 \) and variance \( \sigma^2 \), and the other has weight \( \frac{1}{3} \), mean \( \hat{y}/2 \) and variance \( \sigma^2/2 \). The first corresponds to the situation where \( n = 1 \) and the second to the case where \( n = 2 \). Figure 1 illustrates this situation with \( \sigma = 0.1 \). The solid line corresponds to the case described above. Also shown,
with a dashed line, is an equally weighted mixture of the same two normal distributions, with both weights equal to $\frac{1}{2}$.

Observe, by making a visual estimate of the area under each peak, that, for the solid curve, $\mu$ is a posteriori about twice as probable to be in the general vicinity of 1 as it is to be in the general vicinity of $\frac{1}{2}$. For the two posterior probabilities to be approximately equal, all that is needed is for the density of the prior distribution for $\mu$ in the vicinity of $\mu = \frac{1}{2}$ to be twice as high as its density in the vicinity of $\mu = 1$. A simple choice of joint prior on $n$ and $\mu$ that meets these requirements is given by

$$f(\mu \mid n) = \frac{n}{2c}, \quad -\frac{c}{n} < \mu < \frac{c}{n}, \quad (5)$$

and (3). The only limitations here on the choice of $c$ are that it must be positive, finite and larger than twice the observed $\hat{y}$.

It might be considered strange that the prior distribution of $\mu$ should depend on the realised value of the random variable $n$, but we should remember that this realised value of $n$ is never revealed to the final observer (that is, to the analyst). What that observer is implying, by the use of this prior distribution, is that if the realised value of $n$ is large then the value of $\mu$ is likely to be relatively small, and vice versa. Since the only datum he has is a single observation on a random variable that has the conditional expectation $n\mu$, this is, in fact, a very reasonable proposition.

If now, in the derivation of $f(n \mid \hat{y})$, the joint prior defined by (5) and (3) is substituted for that defined by (2) and (3), then we obtain

$$f(\mu \mid \hat{y}) = \frac{1}{2} N\left(\mu; \frac{\hat{y}}{1}, \frac{\sigma^2}{1}\right) + \frac{1}{2} N\left(\mu; \frac{\hat{y}}{2}, \frac{\sigma^2}{2}\right),$$

which corresponds exactly to the dashed line in Figure 1.

This choice therefore resolves the paradox completely. It is, however, instructive to generalise the situation to one where there is a wider choice of $n$.

5. A generalization to $n > 2$

If we now generalize the paradoxical experiment introduced in Section 2 to the situation where $n$ can take any of the values from 1 to $m$, and the distribution $f(n)$ is still both known and objectively determined, but not necessarily uniform over all values of $n$, we obtain (1),

$$f(n) = p_n, \quad n = 1, 2, \ldots, m, \quad \sum_{n=1}^{m} p_n = 1, \quad (6)$$

and (7). Combining (6) and (5), we obtain the following step function for the unconditional prior on $\mu$:

$$f(\mu) = \frac{1}{2c} \left\{ \begin{array}{ll}
1p_1 + \cdots + mp_m, & |\mu| < \frac{c}{m}, \\
1p_1 + \cdots + (m-1)p_{m-1}, & \frac{c}{m} < |\mu| < \frac{c}{m-1}, \\
\vdots & \vdots \\
1p_1, & \frac{c}{2} < |\mu| < \frac{c}{1}, \\
0, & |\mu| > c.
\end{array} \right. \quad (7)$$
If viewed as a two-dimensional stepped pyramid, the height of the $n$th step is $np_n/(2c)$ and its length is $2c/n$, so its area is $p_n$. The height of the pyramid up to the top of the $n$th step is $\sum_{i=1}^{n} p_i/(2c)$, and that height is proportional to the prior density for $\mu$ in the interval $-c/n < \mu < c/n$. The entire area occupied by the pyramid is one unit, corresponding to $\sum_{n=1}^{m} p_n = 1$. Figure 2 illustrates this for the case where $c = 100$ and $f(n) = \frac{1}{5}$, $n = 1, \ldots, 5$.

The unconditional prior suggested for $\mu$, given by (7), leads to the following posterior:

$$f(\mu | \hat{y}) = \sum_{n=1}^{m} p_n N\left(\mu; \frac{\hat{y}}{n}, \frac{\sigma^2}{n}\right),$$

which is a mixture of normal distributions with weights identical to the prior probabilities given by (6).

Suppose now that we are told nothing about the prior distribution of $n$. That is, we no longer know the $f(n)$ given in (6). In this case it seems reasonable to adapt a suggestion in relation to the ‘tramcar problem’ by Jeffreys (1961, Section 4.8) and take the prior on $n$ as

$$f(n) \propto \frac{1}{n}, \quad n = 1, \ldots, m,$$

where $m$ is ‘large’. In this case, $p_n = 1/(kn)$, where $k = \sum_{n=1}^{m} n^{-1}$, and the unconditional prior density $f(\mu)$ becomes approximately proportional to $1/|\mu|$. That is,

$$f(\mu) \propto \frac{1}{|\mu|}, \quad \mu \in \mathbb{R}.$$
Another posterior paradox

Figure 3: Unconditional prior density of $\mu$ (continuous line) when $c = 100$ and $f(n) \propto 1/n$, $n = 1, \ldots, 10$, and a prior proportional to $1/|\mu|$ (dashed line).

Figure 4: Posterior density of $\mu$ when $c = 100$, $f(n) \propto 1/n$, $n = 1, \ldots, 10$, and $\hat{y} = 1$.

is determined by the choice of $m$ to be (in this case) 10. Larger values of $m$ would sharpen the peak and shift it closer to $\mu = 0$. The reason why the special case defined by the prior (9) was singled out for special treatment was because of Benford’s (empirical) law of numbers (see Benford (1938) and Hill (1998)), which is closely related to Jeffreys’ tramcar problem. We now turn to a consideration of that law.

6. Benford’s law of numbers, and its relevance

There is, in fact, considerable empirical support for the specification $f(n) \propto 1/n$. Benford (1938), elaborating on an observation by Newcomb (1881), demonstrated that if a number $q$ is randomly selected from almost any large source (such as a tabulation of census data or a list of physical constants) then the probability that $q$ commences with the digit $d$ ($d = 1, 2, \ldots, 9$) is $\log_{10}(1 + d^{-1})$. As shown in Lee (1989, Section 3.8), this result would be implied by the density of $q$ itself being proportional to $q^{-1}$. In the same way, we might (in the absence of any evidence to the contrary) expect the prior density of $\mu$ to be roughly proportional to $|\mu|^{-1}$, at least on each half of the real line separately. However, since there could well be a substantial difference between these two halves, we now concentrate our attention on the distribution of $|\mu|$. Applying Lee’s result in this context, we have that the adoption of the prior distribution $f(\mu) \propto |\mu|^{-1}$ implies that Benford’s law holds for $|\mu|$.
There is an important reservation which needs to be kept in mind here. Although Benford’s law is implied by \( f(q) \propto q^{-1} \), the converse does not hold. Also, in practice, \( f(q) \) falls substantially below the values implied by \( f(q) \propto q^{-1} \), both when \( q \) is very large and when \( q \) is very small. The reason for this is simply that human conventions, following the requirements of human convenience, adopt units for measuring very small and very large quantities that are different from those used to measure quantities of intermediate size. For instance, we use Ångstrom units for certain subatomic distances; millimetres, metres and kilometres for terrestrial distances; parsecs and megaparsecs for interstellar and intergalactic distances. These conventions enable us to avoid the need for many zeros before or after the decimal points, without invalidating Benford’s law, but cause \( f(q) \propto q^{-1} \) to fail for very large and very small numbers. Fortunately, however, this failure conveniently corresponds to the two ranges of \( \mu \) for which its suggested prior distribution, given by (7), falls substantially below proportionality to \( |\mu|^{-1} \).

So far, we have only considered the relevance of Benford’s law in relation to \( |\mu| \), but, in situations where the sample size is not determined by a known random distribution, there is also a relevance to the distribution of \( n \). In those circumstances it makes sense to assume that Benford’s law holds for \( n \) (as well as for \( |\mu| \)) and, correspondingly, also that \( f(n) \propto n^{-1} \) except (in this case) for very large values of \( n \). Once again, that is consonant with the assumption already suggested in (5), namely that \( p_i = 2c/i \). There is, therefore, a close relationship between the step function for \( f(\mu) \) implied by that assumption, namely \( f(\mu) = n, -c/n < \mu < c/n, n = 1, \ldots, m \), and the middle region of \( f(\mu) \propto |\mu|^{-1} \) that implies Benford’s law. This relationship is illustrated by Figure 2.

For very small and very large values of \( |\mu| \), where \( f(\mu) \propto |\mu|^{-1} \) is not empirically supported by Benford’s law, the step function \( f(\mu) = n, -c/n < \mu < c/n, n = 1, \ldots, m \), takes (quite appropriately) values that are well below those implied by \( f(\mu) \propto |\mu|^{-1} \). For small values of \( |\mu| \), specifically for \( |\mu| < c/m \), the step function is flat at \( f(\mu) = m \). For values of \( |\mu| > c \), it takes the value zero. While there is some arbitrariness in the choices of \( c \) and \( m \), a possible rule of thumb would be first to set \( m \) at a number as large as the available computer can comfortably handle (in order to avoid unnecessary truncation of the summation in (8)), and then to set \( c \) equal to some convenient number larger than \( m \) (or in those more usual situations where individual observation values are available, a number larger than the largest of those observations).

What all this amounts to is that the assumption that \( f(n) \propto n^{-1} \) except for very large values of \( n \) leads to the conclusion that \( f(\mu) \propto |\mu|^{-1} \) holds for the middle ranges of \( |\mu| \), raising at least the possibility that Benford’s law is not merely an empirical discovery, but to some degree a logical necessity as well.

The relevance of Benford’s law therefore amounts to the following points.

1. Assuming that Benford’s law holds, at least for small and moderate \( n \), we are led to the conclusion that it should also hold in the middle ranges of \( |\mu| \).

2. Assuming that Benford’s law holds for the middle ranges of \( |\mu| \) (for whatever reason), also resolves the paradox outlined in Sections 2–5 (namely that the range of posterior distributions for \( n \) was incompatible with its known objective probability distribution).

3. The fact that Benford’s law does not imply \( f(\mu) \propto \mu^{-1} \) for very large or very small values of \( |\mu| \), encourages the use of a prior that approximates to \( f(\mu) \propto \mu^{-1} \) for moderate values of \( |\mu| \), that caps the size of \( f(\mu) \) for small values of \( |\mu| \), and which also implies that \( f(\mu) = 0 \) for very large values of \( |\mu| \).
4. The step function \( f(\mu) = n, \; -c/n < \mu < c/n, \; n = 1, \ldots, m \), with \( c \) and \( m \) appropriately chosen, provides a particularly convenient way of fulfilling the requirements set out in point 3.

It might, of course, be considered preferable to use the improper distribution \( f(\mu) \propto |\mu|^{-1} \) instead, in order to avoid the arbitrary elements in the choices of \( c \) and \( m \), but the following considerations militate against this.

1. There is no simple analytical form for \( \int u^{-1}e^{-u^2} \, du \), so numerical integration would be necessary.

2. The improper prior leads to an improper posterior, because the integral \( \int_{0}^{\infty} u^{-1}e^{-u^2} \, du \) is divergent.

3. When \( \sigma^2 \) is not small, the normal distributions in (8) can have appreciable heights at \( \mu = 0 \), at which value \( f(\mu) \) is indefinitely large, and this could disturb the approximation to an unlimited extent.

7. Other forms of data

The fact that the proper priors introduced in Sections 4 and 5 remove the paradox initially presented does not necessarily mean that they should be used in all situations. The paradox is only a sharp one where the observations are sparse, where also \( n \) is either a random variable with a known distribution (see Section 4) or subject to Benford’s law of numbers (see Section 5), and finally where, in addition, the observed datum is the sample total, \( \hat{y} \). In other situations, the use of a flat prior is simpler and can easily be justified. If, for instance, \( n \) is known beforehand, or the observed datum is \( (n, \bar{y}) \), then no such paradox arises. If the datum is something more complicated than \( \hat{y} \), for example if it is \( n\hat{y} \), not even the proper priors of Sections 4 and 5 can remove the paradox. However, since it is difficult to imagine a situation where \( n\hat{y} \) is known but \( n \) is not, such problems are unlikely to arise in real life.

The relevance of the paradox is therefore limited in practice; but it nevertheless teaches the important lesson that the flat prior over the entire real line is not uninformative. That prior implies, rather, that almost all the potentially observable data have indefinitely large absolute values. There is a close parallel here with the situation described in Section 1, where the Haldane, Jeffreys and uniform priors were very different but gave rise to very similar posteriors once the sample was of moderate size. The parallel between the Haldane prior and the flat prior over the real line is particularly close, in that both are improper, and both imply, almost certainly, extreme-valued data.

8. Summary

The paradox described is one that can arise when a uniform distribution, whether actually improper or merely flat over an unrealistically wide range, is used as a reference prior. The particular case investigated in Section 2 was that where \( y \sim N(\mu, \sigma^2) \) with \( \sigma^2 \) known but \( \mu \) unknown, where the only datum is \( \hat{y} = \sum_{i=1}^{n} y_i \), and where \( n \) has a known and objective random distribution. The use of a flat reference prior then leads to contradictory inferences for \( n \) and to a potentially serious overestimation of \( \mu \).

In Section 3, this was explained by the bulk of the prior probability being located in regions with extreme values of \( |\mu| \). More appropriate priors for \( \mu \) were then presented, in Section 4 for this original problem, and in Section 5 for the case where the distribution of \( n \) is unknown.
In Section 6 this last case was shown to be related to Benford’s law of large numbers, and Section 7 dealt with other forms of data. A major conclusion of this article is that care should be taken when contemplating the use of a flat reference prior.

The SPLUS® code used to generate Figures 1–4 is available on demand from the authors.

Acknowledgements

We are indebted to Professor D. V. Lindley for first drawing our attention to the highly informative nature of a uniform distribution over the whole real line, and also to Professor Sir David Cox for suggesting a connection with Jeffreys’ tramcar problem.

References