A Note Concerning a Selection "Paradox" of Dawid's

Stephen SENN

This article briefly reviews a selection "paradox" of Dawid's, whereby Bayesian inference appears to be unchanged whether or not treatments have been selected for inspection on the basis of extreme values. The problem is recast in terms of a hierarchical model. This offers an alternative explanation of the paradox but also reveals a disturbing dependence of inference on prior specification. The example may also be used to deepen students' understanding of the implications of using conjugate nonhierarchical priors in Bayesian analysis. To illustrate, some simulations are presented.

KEY WORDS: Bayesian inference; Hierarchical models; Prior distributions; Selection paradox.

1. INTRODUCTION

In a penetrating but also rather disturbing analysis, Dawid has drawn attention to apparent sharp disagreements between Bayesian and frequentist approaches to the analysis of a particular type of problem (Dawid 1994). This problem occurs when examination of a dataset leads to further investigation of a subset concerning which statistical inference is then required. In frequentist approaches it is generally accepted that the selection process has implications for the further analysis. For Bayesian approaches, however, Dawid pointed out that, "Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data" (p. 211). Dawid concluded that there is nothing illogical in such a Bayesian analysis but that, since the conclusion may be unwelcome, Bayesians nevertheless may be wise to think carefully about the sort of prior distribution that makes it possible. In particular, he suggested that, "conjugate priors for multivariate problems typically embody an unreasonable determinism property" (p. 211).

In this article I consider one particular example of Dawid's and offer a heuristic explanation of the phenomenon. A referee has pointed out to me that, since the phenomenon is perfectly logical from the Bayesian point of view, it is not a paradox. Nevertheless, it points to a sharp difference with frequentist practice and possibly with some intuition that many Bayesians will share. Hence, the fact that posterior distribution does not change when selecting the largest mean can be regarded as a paradox within a wider framework. Furthermore, there is some precedent to using the word "paradox" in such cases. For example, Lindley's famous demonstration (Bartlett 1957; Lindley 1957) that moderately significant p-values are more indicative of the truth of the null hypothesis when the sample size is large is entitled "A Statistical Paradox" but is not paradoxical if one considers the Bayesian formulation which justifies it as correct. In the rest of the article I shall refer to *Dawid's selection paradox*. This carries with it no implication that there is something wrong with Bayesian inference, although examination of it will support the view that prior distributions must be chosen with care.

This heuristic explanation offered for Dawid's selection paradox immediately suggests an alternative hierarchical formulation of priors that may partially resolve it or at the very least provide a useful way of looking at it. The conclusions may be simply illustrated using simulations and thus may make a suitable teaching example to deepen students' understanding of Bayesian methods and the implications of certain choices of prior distribution.

2. THE SPECIFIC EXAMPLE

Dawid considers a number of common examples of selection in statistics. The first and simplest of these will be examined here. Dawid describes an agricultural trial in which p varieties are tested, with the aim of choosing that having the highest mean yield. This general sort of problem arises elsewhere. In what follows, I shall consider instead the analogous case of choosing the most active among a group of similar pharmaceuticals in a preclinical experiment, for no other reason than that it is a practical setting with which I have some familiarity. Note, however, that using a conjugate prior is not necessarily the most appropriate way to model this and for alternatives the reader might look, for example, at Westfall, Johnson, and Utts (1997).

Suppose, following Dawid, that we select for future use the most apparently active pharmaceutical i^* associated with the largest sample mean, X_{i^*}. We suppose that each $X_i \sim N\left(\mu_i, \sigma^2\right)$. (Of course, each mean may not in practice be measured with equal precision, since, for example, the number of replicates may not be identical. It is, however, sufficient for the purpose of illustrating the point to be made here if this simplifying assumption is made.) Having selected the given treatment,

Stephen Senn is Professor of Statistics, Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, UK (E-mail: *stephen@stats.gla.ac.uk*). I am grateful to the UK Engineering and Physical Research Council for funding through the Simplicity, Complexity and Modelling project and to Novartis for further funding. I thank the referees and editors for helpful comments.

  DOI: 10.1198/000313008X331530

we wish to make inferences about μ_{i^*}, which we label μ^* for simplicity.

Now suppose that we take the proper prior $\mu_i \sim N\left(\theta, \tau^2\right)$ independently. Since θ and τ are known parameters (else we cannot express our prior distribution) we may, without loss of generality, proceed to measure everything in terms of standardized units of the prior for which revised formulation we have $\mu_i \sim N(0, 1)$. Suppose, therefore, that in our previous formulation this has been done. Hence σ^2 is the ratio of the data variance for a given treatment mean to the prior variance and we may set $\theta = 0,\ \tau^2 = 1$. We assume that σ^2 is known. Let y_i be the posterior mean corresponding to data mean x_i and let q_i^2 be the posterior variance. Hence, we have from standard Bayesian results, whereby the posterior mean is the precision-weighted linear combination of prior and data means and the posterior precision is the sum of prior and data precisions ("precision" here being the reciprocal of the variance; Box and Tiao 1992),

$$\begin{aligned} y_i &= \left(\frac{1}{1+\sigma^2}\right) x_i, \\ q_i^2 &= \left(1+\frac{1}{\sigma^2}\right)^{-1}, \\ \mu_i &\sim N\left(y_i, q_i^2\right). \end{aligned} \tag{1}$$

All we need to do to obtain the corresponding inference for μ^* is to substitute the largest observed mean, x^* for x_i and write y^* for the posterior mean. The difference between the data mean and the posterior mean is

$$-\left(\frac{\sigma^2}{1+\sigma^2}\right) x^*.$$

We thus see that there is indeed a shrinkage that takes place and that this will be larger the greater the posterior mean and also larger the greater the data variance, σ^2. However, shrinkage does not depend on the number, p, of treatments examined, nor does it depend on their mean $\bar{x} = \sum_{i=1}^{p} x_i / p$.

We now illustrate this using some simulations. The purpose of these requires a little explanation. These simulations are not designed to test how well Bayesian estimation works, in the sense of calibrating a Bayesian estimation procedure against a possible constellation of parameters in order to check frequentist properties, as for example in Lambert et al.'s examination of prior distributions for random effect variances (Lambert et al. 2005). Nor are these simulations an exercise in practical inference in the sense of Markov chain Monte Carlo estimation (Gelfand and Smith 1990). The simulations are an empirical representation of the knowledge embodied in two pieces of information we assume known: the prior distribution of μ_i and the conditional distribution of X_i given μ_i. Given these two pieces of information the bivariate distribution of μ_i, X_i is defined and can be empirically represented by a simulation. From such a simulation the regression of X_i on μ_i can be estimated. The regression coefficient which it estimates is nothing less than the Bayesian shrinkage factor, which is given by (1). Of course, this

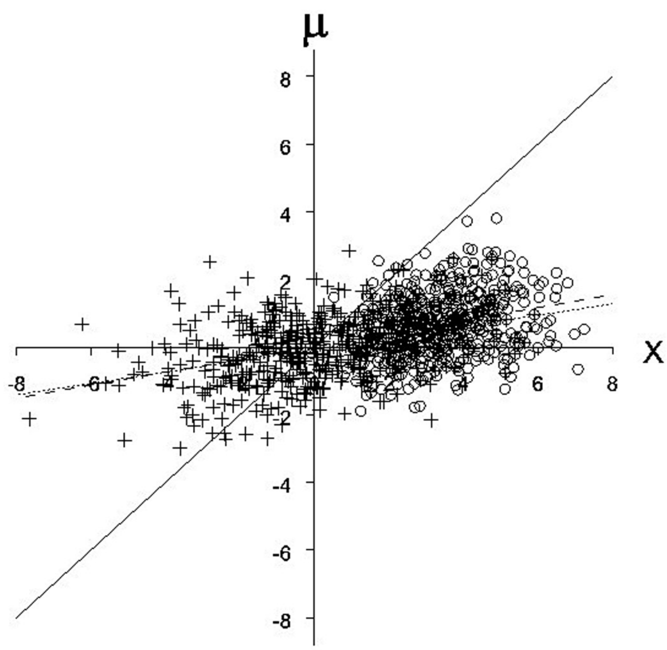

Figure 1. Simulation of the nonhierarchical case. "True" means, μ, and observed means X. The 500 unselected pairs are represented by $+$ signs and the 500 pairs selected because the observed mean is the highest of 10 are represented by $\circ$. The two regression lines of true on observed means are shown and are virtually identical and at approximately 0.2 are much less than 1, showing strong Bayesian shrinkage compared to exact equality, which is shown as a thick diagonal line.

means the simulation is superfluous. However, an empirical representation can be revealing for the student. Since it is estimated from a simulation, the estimated shrinkage factor is subject to sampling variation, but if the number of runs is large it will be very close to the theoretical value given by (1).

Consider now Figure 1. Two different setups have been simulated. In both a Normal population of treatment means μ is assumed to exist with mean $\theta = 0$ and variance $\tau^2 = 1$. Such a population is a frequentist analogue of the prior distribution discussed earlier. A series of $n = 10$ "true" treatment means $\mu_i,\ i = 1 \ldots n$ are simulated and to go with each an observed mean $X_i = x_i \sim N\left(\mu_i, \sigma^2\right),\ \sigma^2 = 4$ is also simulated. This is repeated 1,000 times and for each of 500 runs a pair (μ_i, x_i) is drawn at random from the ten simulated during that run. These 500 "random" pairs are plotted using the symbol $+$ with x_i on the horizontal axis and μ_i on the vertical axis. The estimated regression line of μ on x is plotted as a dotted line. For each of the other 500 runs a pair (μ_i, x_i) is chosen such that x_i is the largest among the ten generated. These "selected" pairs are plotted as circles on the same figure and the regression of μ on x is plotted as a dashed line. (Where color representation is available, much larger numbers of simulation runs can be represented to make the diagram more convincing.)

What the simulation shows is that whereas, of course, the selected x values are higher than the randomly chosen x values and in consequence the expected value of μ is higher where x is selected, the conditional expectation of μ for given observed

x is the same in both cases. This is illustrated by the regression lines being nearly on top of each other. In fact expression (1) says that the regression slope should be

$$\frac{1}{1+\sigma^2} = \frac{1}{1+4} = 0.2.$$

The simulation produces an estimated slope of 0.19 for the unselected group and 0.20 for the selected group. These are chance differences and for a classroom demonstration one might like to increase the run size.

3. A HEURISTIC EXPLANATION

A heuristic explanation of the conditional distribution being unaffected by selection can be offered that is quite revealing. The explanation is in line with the determinism that Dawid mentions. Consider what it means to be able to specify the proper (conjugate) prior that was used in Section 2. It implies that the population of means μ_i is known. It may seem at first that this is preposterous. Such prior distributions are commonly used with large values of τ^2 to represent very *imprecise* prior knowledge. However, a little thought suffices to show that, first, given this prior distribution the full distribution of all possible predictive means can be produced, second, that simulating from such a distribution can be approximated using an empirical distribution based on a large collection of means sampled from it and third, that however many means are examined the posterior mean of the distribution *as a whole* is unaffected. This gives the explanation as to why the values of the other means in the experiment have no influence. They are simply a random subset of the infinity of means from which this one has been drawn and with which it is *already implicitly being compared*. Given the prior distribution, which describes the distribution of all means, the particular sample of means observed is uninformative except to the extent that each x_i is separately informative about the relevant and corresponding μ_i.

4. A HIERARCHICAL MODEL

This paradox is entirely an artifact of having selected a prior distribution with independent μ_i. Instead of supposing that the treatments are drawn independently from the same population, suppose instead that they are drawn by some process analogous to cluster sampling. This might, indeed, be an appropriate analogy for the example of testing pharmaceuticals that we considered in Section 2, where identification of a particular chemical structure, a so-called "lead compound," suggests a number of similar compounds that may be examined together. Together the treatments form a "compound class." We suppose now that the following model holds. The μ_i are now no longer independent given the original prior but are conditionally independent (Dawid 1979) given their "true" class mean ϕ. We now write the model in terms of a two-stage hierarchy of parameters as follows:

$$\begin{aligned} \phi &\sim N\left(0, 1-\gamma^2\right) \\ \mu_i &\sim N\left(\phi, \gamma^2\right) \end{aligned}$$

with the data distribution conditional on μ_i as before. Note that for a single observed data mean x_i, the two stages are irrelevant with regards to what we can learn about μ_i: the parameter γ^2 is chosen so that the marginal distribution of X_i has variance $\left(1-\gamma^2\right)+\gamma^2+\sigma^2 = 1+\sigma^2$ as before. However, the hierarchical structure induces a correlation and it is no longer the case, where a sample of means have been observed, that our inference about any one of them is unaffected by the values of the others.

We can construct an appropriate inference by proceeding in the spirit of Dawid's prequential analysis (Dawid 1984). Suppose that we first observe $p-1$ means, which we label $x_1, x_2, \ldots x_{p-1}$ and then construct a posterior distribution for ϕ. We can use this to obtain a predictive distribution for μ_p. This distribution contains all the information about μ_p not contained in x_p. It can therefore act as a prior distribution for x_p. Finally, having observed x_p, we can update the distribution of μ_p appropriately. Given exchangeability, the order in which the means occur does not change our inference about any of the parameters once all the data are in. (This is central to the fact that the prequential approach can be used to mimic many standard inferences by acting as if the data have been added bit by bit.) We can, without loss of generality, assume that the largest mean we observe is the last.

Note that conditional on ϕ, the data means are independent with variance $\gamma^2+\sigma^2$. Therefore the data variance for the mean of the $p-1$ means, $\bar{X}_{p-1}$, is $\left(\gamma^2+\sigma^2\right)/(p-1)$. However, the prior variance for ϕ is $\left(1-\gamma^2\right)$ and its prior mean is 0. Hence the posterior distribution for ϕ is $N\left(z, r^2\right)$ with

$$\begin{aligned} z &= \left[\frac{(p-1)\left(1-\gamma^2\right)}{(p-1)\left(1-\gamma^2\right)+\left(\gamma^2+\sigma^2\right)}\right]\bar{x}_{p-1}, \\ r^2 &= \frac{\left(1-\gamma^2\right)\left(\gamma^2+\sigma^2\right)}{(p-1)\left(1-\gamma^2\right)+\left(\gamma^2+\sigma^2\right)}. \end{aligned} \tag{2}$$

Note that as $\gamma^2 \to 1$ the posterior distribution of ϕ collapses around 0. This is as it should be since this effectively eliminates the stage of the hierarchy describing the distribution of ϕ and as was already noted in Section 3 under such circumstances the *distribution* of the μ_i is known with certainty a priori even if any given individual value is not. However, for $\gamma^2 = 0$, the posterior mean and variance in (2) are

$$z = \frac{(p-1)}{(p-1)+\sigma^2}\bar{x}_{p-1}, \; r^2 = \frac{\sigma^2}{(p-1)+\sigma^2}. \tag{3}$$

Now we may use (2) to obtain the "predictive" distribution for μ_p (which, however, will not be observed). This will have the same expectation but variance $s^2 = r^2+\gamma^2$. We may use this as a prior for μ_p and then update using the standard Bayesian approach for inference about Normal distributions with conjugate priors already described to obtain—after some tedious but elementary operations—an expression for the posterior mean of μ_p as

$$\frac{(p-1)\left(1-\gamma^2\right)\left(\bar{x}_{p-1}\sigma^2+x_p\gamma^2\right)+x_p\left(\sigma^2+\gamma^2\right)}{\left[(p-1)\left(1-\gamma^2\right)+\left(1+\sigma^2\right)\right]\left(\sigma^2+\gamma^2\right)}. \tag{4}$$

Note that as $\gamma^2 \to 1$, the posterior mean given by (4) approaches that given by (1) whatever the value of p, but if $p = 1$, then it is the same as (1) whatever the value of γ^2. If we divide top and bottom of (4) by $(\sigma^2 + \gamma^2)$ and multiply and divide the rightmost term in the numerator by $(1 + \sigma^2)$, then we can write the posterior mean as

$$\frac{(p-1)\left(1-\gamma^2\right)\dfrac{\left(\bar{x}_{p-1}\sigma^2 + x_p\gamma^2\right)}{\left(\sigma^2+\gamma^2\right)} + \left(1+\sigma^2\right)\dfrac{x_p}{1+\sigma^2}}{(p-1)\left(1-\gamma^2\right) + \left(1+\sigma^2\right)}. \tag{5}$$

Thus, (5) is a weighted average of

$$\frac{\left(\bar{x}_{p-1}\sigma^2 + x_p\gamma^2\right)}{\left(\sigma^2+\gamma^2\right)}$$

and

$$\frac{x_p}{1+\sigma^2}$$

so that if the former is the same as the latter we have $\frac{x_p}{1+\sigma^2}$, which is not dependent on p (and the solution for the nonhierarchical case). Solving for this condition yields

$$\gamma^2 = 1 - \frac{\bar{x}_{p-1}}{x_p}\left(1+\sigma^2\right). \tag{6}$$

Expression (4) would be the same, of course, if x_p were any arbitrary mean among the p means. It is the relationship between x_p and $\bar{x}_{p-1}$ that influences the adjustment, and given these values whether x_p is the largest mean or some other mean is irrelevant. If $x_p = \bar{x}_{p-1}$, then (4) becomes

$$\frac{\left[(p-1)\left(1-\gamma^2\right)+1\right]x_p}{(p-1)\left(1-\gamma^2\right)+\left(1+\sigma^2\right)}.$$

5. A NUMERICAL ILLUSTRATION

The behavior of expression (4) is illustrated for a given numerical example in Figure 2. It is assumed that $\sigma^2 = 1/2$, $\bar{x}_{p-1} = 1$ and $x_p = 3$. A horizontal line is drawn at

$$\frac{x_p}{1+\sigma^2} = 2.$$

This is the posterior mean for the nonhierarchical case, but also happens to be the posterior mean, irrespective of γ^2 for $p = 1$. Further values of $p = 2, 5, 100$ are illustrated. Application of (2) shows that for these values the posterior means of ϕ given $\bar{x}_{p-1} = 1$ (but without knowledge of x_p) are 0.667, 0.889, and 0.995, respectively, when $\gamma^2 = 0$. These are the values to which the posterior means for μ_p given by (4) are "attracted" for the case when $\gamma^2 = 0$ and they are progressively closer to the observed value of $\bar{x}_{p-1} = 1$, the larger the value of p.

The behavior with respect to γ^2, however, is rather more interesting. It can be seen that for $p = 1$, which is the trivial case where the largest mean is the only mean obtained, we have the same inference whatever the value of γ^2. However, for other values of p the behavior is rather more complex. For $\gamma^2 = 0$

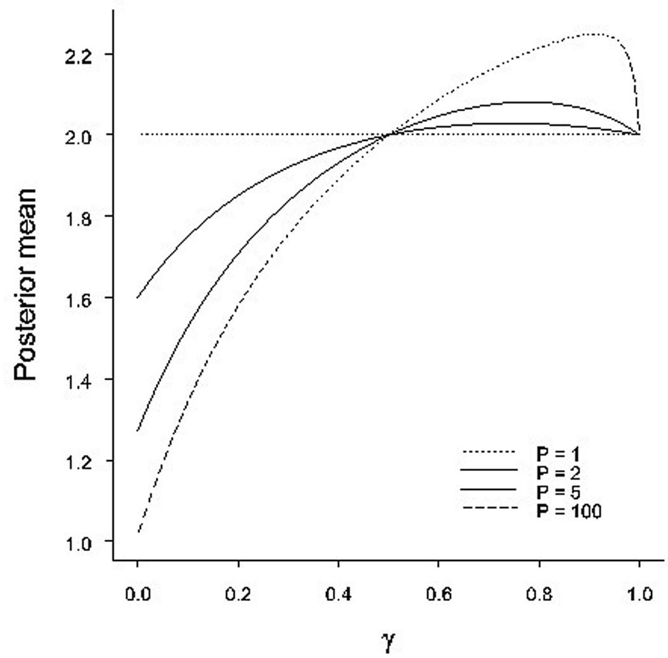

Figure 2. Hierarchical case. Posterior mean as a function of the hierarchical variance γ^2 when the data mean is 3. Various cases are illustrated depending on the number of means, p, from which the observed mean has been selected as the largest.

the values start near to the posterior mean of ϕ, which is less than the value of the posterior mean for μ_p of 2 for the nonhierarchical case. At $\gamma^2 = 0.5$, we have the same posterior mean for μ_p , whatever the value of p; that is to say, however many compounds in the class we have studied.

This is because for our numerical example (6) yields the result $\gamma^2 = 0.5$.

For $0.5 < \gamma^2 < 1$ we have that the posterior mean of μ_p is actually higher than for the nonhierarchical case.

6. A SIMULATION OF THE HIERARCHICAL CASE

Figure 3 shows an analogous simulation to Figure 1 but with an extra stage to illustrate hierarchical sampling. The value of γ^2 was set to 0.5. As before, the values for θ were 0 and for σ^2 were 4. Using these values, for each of 1,000 runs, a value ϕ for the group true mean was simulated. From each group distribution 10 true means μ_i were simulated and for each μ_i an observed value x_i was also simulated. Again, for 500 simulations runs a pair (μ_i, x_i) was chosen at random from the ten available and for another 500 runs a pair (μ_i, x_i) was chosen such that the value of x was the largest available. Again unselected pairs have been labeled + and selected pairs have been labeled with ○. Both regression lines have been plotted and these will now be seen to be quite different.

7. DISCUSSION

These results support Dawid's examination. The choice of prior is a delicate matter and has considerable influence on posterior inferences. Given our hierarchical prior, then the poste-

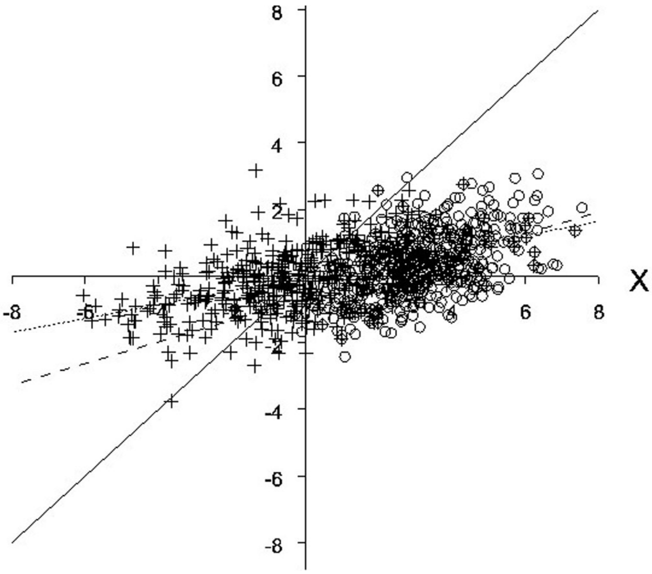

Figure 3. Simulation of the hierarchical case. "True" means, μ, and observed means X. The 500 unselected pairs are represented by $+$ signs and the 500 pairs selected because the observed mean is the highest of 10 are represented by $\circ$. The two regression lines of true on observed means are shown but selected (dash) and unselected (dot) are no longer identical. However, the regressions are still much less than 1 (shown by the thick diagonal line).

rior expected value of the most extreme mean in a collection is affected by both the number of treatments studied and their overall mean.

This is not the same as what happens in classical frequentist inference, where the inferential context is not, as in the Bayesian case, all experience and knowledge to date but some presumed infinitely repeated behavior such as, for example, repeatedly studying p treatments and making inferences about the mean of the largest.

Of course, Bayesian inference for hierarchical datasets is well developed. The program BUGS (Lunn, Thomas, Best, and Spiegelhalter 2000) was developed specifically to handle such inferences and there is a great deal of applied work in this area (Ashby 2006). The example considered is not, however, hierarchical in the traditional sense. The dataset considered, that of a single collection of means, is *not* hierarchical by nature, at least not obviously. What is being illustrated is the analysis of a single "galaxy" of means and not a collection of galaxies. The hierarchy is conceptual. In this respect the example is a little disturbing. The implication for the analysis of clinical trials, for example, would be the following: the same hierarchical model would have to apply for the analysis of a single clinical trial as would apply to a meta-analysis, the only difference being that in the former case one of the components of variation (trial by treatment interaction) would have to be supplied entirely by the prior. For recent discussion of the impact of the choice of prior distribution in meta-analysis (see Lambert et al. 2005; Senn 2007; Lambert et al. 2008; Senn 2008). A practical example of the latter was given recently in a regulatory context. See the discussion of Hung, Wang, and O'Neill (2005) and the hierarchical nature of the problem by Senn (2005).

To return to the example of preclinical screening, however, consider some process of elicitation that reveals a value for σ^2 (this, in our formulation, being the ratio of the data variance to the prior variance, which is standardized at 1). Inference is to be made about a single mean. As such γ^2 does not enter into the equation. As soon as a second treatment is studied, however, γ^2 must become part of the inference and the choice of the value may be crucial. Note that in theory several levels of hierarchy may be necessary to capture appropriate prior beliefs. But even if only two treatments have been studied in this compound class, the class itself may be similar to some other compound classes so that further hierarchical considerations may apply.

In short, the Bayesian may find it difficult to escape from prior experience when seeking to make a valid inference but find it equally difficult to recognize exactly what that prior experience is.

[Received November 2006. Revised May 2008.]

REFERENCES

Ashby, D. (2006), "Bayesian Statistics in Medicine: A 25 Year Review," *Statistics in Medicine*, 25, 3589–3631.

Bartlett, M. S. (1957), A Comment on D.V. Lindley's "Statistical Paradox," *Biometrika*, 44, 533–534.

Box, G. E. P., and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis* (Wiley Classics Library Edition ed.), New York: Wiley.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory" (with discussion), *Journal of the Royal Statistical Society*, Series B, 41, 1–31.

Dawid, A. P. (1984), "Statistical Theory—the Prequential Approach," *Journal of the Royal Statistical Society*, Series A, 147, 278–292.

——— (1994), "Selection Paradoxes of Bayesian Inference," in *Multivariate Analysis and its Applications* (Vol. 24), eds. T. W. Anderson, K. A.-T. A. Fang and I. Olkin, Philadelphia, PA: IMS.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Hung, H. M. J., Wang, S. J., and O'Neill, R. (2005), "A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-Inferiority Trials," *Biometrical Journal*, 47, 28–36.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005), "How Vague Is Vague? A Simulation Study of the Impact of the Use of Vague Prior Distributions in Mcmc Using Winbugs," *Statistics in Medicine*, 24, 2401–2428.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2008), Comments on "Trying to Be Precise About Vagueness" by Stephen Senn, *Statistics in Medicine* 2007; 26, 1417–1430, *Statistics in Medicine*, 27, 619–622. Author reply, 622–614.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "Winbugs—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility," *Statistics and Computing*, 10, 325–337.

Senn, S. (2005), " 'Equivalence Is Different'—Some Comments on Therapeutic Equivalence," *Biometrical Journal*, 47, 104–107.

——— (2007), "Trying to Be Precise About Vagueness," *Statistics in Medicine*, 26, 1417–1430.

——— (2008), Reply to Sutton et al.," *Statistics in Medicine*, 27, 622–624.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian Perspective on the Bonferroni Adjustment," *Biometrika*, 84, 419–427.