Comments on 'Bayesian estimate of the Newtonian constant of gravitation' with an alternative analysis

Meas. Sci. Technol. **18** (2007) 2275–2280 doi:10.1088/0957-0233/18/7/063

COMMENT

Comments on ‘Bayesian estimate of the Newtonian constant of gravitation’ with an alternative analysis

R Willink

Industrial Research Ltd, PO Box 31-310, Lower Hutt 5040, New Zealand

E-mail: r.willink@irl.cri.nz

Received 19 February 2007, in final form 23 March 2007
Published 19 June 2007
Online at stacks.iop.org/MST/18/2275

Abstract
This paper provides critical comments on the methodology employed in ‘Bayesian estimate of the Newtonian constant of gravitation’ (Dose 2007 *Meas. Sci. Technol.* **18** 176–82), where many independent estimates of G and their quoted uncertainties are combined to form a single estimate and standard uncertainty. It is argued that the use of the non-Gaussian likelihood functions is unwarranted and that some of the prior distributions used are inappropriate. Another criticism is that the uncertainty of the final estimate is interpreted differently from the uncertainties of the contributing estimates. A non-Bayesian means of analysis is presented for problems of this type. The corresponding estimate is $G = 6.674\,08(12) \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, with the figures in parentheses indicating standard uncertainty.

Keywords: inconsistent data, likelihood model, outlier detection, random effects model, robust estimation

1. Introduction

In [1], an estimate of the Newtonian constant of gravitation and the uncertainty of this estimate are obtained by combining a number of independent estimates in a Bayesian statistical analysis. Here, we make a number of observations about the methodology employed in that analysis. We argue that the estimate and its uncertainty are derived using several assumptions that cannot be justified[1].

A fundamental concept of Bayesian statistics is the idea that strength of belief can be quantified. In science, such quantification of belief is to be a rational reflection of the information available. The way in which this information is propagated in response to the acquisition of data is through Bayes’ rule, which can be summarized by ‘*posterior probability* $\propto$ *prior probability* $\times$ *likelihood*’ or, less succinctly, by

> ‘the probability attributed to Hypothesis A being true after analysing a dataset’ $\propto$ ‘the probability attributed to A being true before analysing the dataset’ $\times$ ‘the probability of observing that particular dataset on the condition that A is true’.

So, according to the basic tool of Bayesian statistics, a meaningful and correct result can be obtained from an acceptable quantitative representation of prior belief and an accurate description of the probabilistic mechanism generating the data (i.e. the form of the likelihood function). Conversely, if either of these descriptions is inadequate then we cannot claim that our result is trustworthy. The arguments given in this comment suggest that both these elements are missing from the analysis in [1].

Section 2 discusses the likelihood models used in [1] and the related concept of robustness to outliers. Section 3 considers the descriptions of prior probability in [1]. Section 4 comments on the posterior distributions obtained and on matters of interpretation. Section 5 considers the weighted-

[1] This author refereed an earlier form of [1] and recommended rejection in that form.

mean least-squares estimator and the criticisms of (the use of) this estimator given in [1]. Section 5.1 shows how the adoption of a natural model leads to a modified form of weighted mean being a suitable estimator of G, and presents a corresponding estimate. Section 5.2 contains some comments about the treatment of data that appear inconsistent.

2. Likelihood models

The problem of estimating the Newtonian constant of gravitation G addressed by Dose in [1] involves combining a set of n independent measurement estimates $\{d_i\}$ quoted with standard uncertainties $\{\sigma_i\}$. Initially, the figure σ_i is reasonably interpreted as representing the standard deviation of the distribution from which d_i was drawn and each of the n distributions is reasonably taken to have mean equal to the unknown value G. The idea of forming an acceptable likelihood function is therefore the idea of accurately describing the forms of these distributions.

The first distribution considered by Dose is the Gaussian distribution. There is a sound basis for assuming that d_i is drawn from such a distribution; the elemental errors occurring throughout the process of measurement can be regarded as adding independently. An approximation to normality ensues from the central limit theorem (which does not require the variables to be identically distributed). If we were to depart from the normal model then we might favour a blunter distribution, such as the uniform distribution, because often non-normal components of error will be modelled as uniform random variables and one of these components might be dominant. Dose, however, neither explicitly nor implicitly associates the likelihood functions used in [1] with the data-generating process. He gives a different justification, the principle of maximum entropy, for using the Gaussian distribution and considers two other (less blunt) distributions—the Laplace distribution and the 'hyperbolic-cosine' distribution—chosen apparently because they would be less sensitive to values of d_i 'far off the mainstream' [1, p 177].

As implied, the central limit theorem provides a strong theoretical reason for supposing the data to be drawn from distributions that are approximately normal. But what is the reason for supposing the data to be drawn from approximate Laplace distributions or hyperbolic-cosine distributions? Moreover, what is the reason for assuming that the same form of non-normal distribution is applicable with each of the data points? That is, why should the likelihood have the same form for different experiments carried in 1982 and 2003—unless there is some unifying mechanism like the central limit theorem?

Thus, Dose does not seem to see the likelihood function as reflecting the data-generating process. Instead, he appears to base his analysis around the concept of robustness to outliers.

2.1. Robustness to outliers

In [1, p 178], we read that 'The aim of the present paper is to make data censoring and ad hoc adjustments superfluous by developing sufficiently robust methods'. But what does 'sufficiently robust' mean, and what evidence is there that the Gaussian, Laplace and hyperbolic-cosine distributions are sufficiently robust? The forms of the Gaussian and Laplace densities are $\exp(-|x|^a)$ with $a = 2$ and $a = 1$, respectively. So, if the Laplace density is chosen because it is more robust than the Gaussian density then why not also consider the density with $a = 0.5$ or $a = 0.25$?

Robustness is about maintaining accuracy when one or more of the model assumptions is violated. So, in any full study of robustness there should be an examination of the insensitivity of the results to changes in the sampling model. Moreover, in a Bayesian analysis there should also be an examination of sensitivity to the choice of prior probability model. However, often the only concept of robustness addressed is insensitivity to data that do not appear to have arisen under the assumed sampling distribution, i.e., 'outliers'. This is the concept of robustness referred to in [1].

The tolerance to outliers that Dose seeks is insensitivity to data pairs (d_i, σ_i) that seem inconsistent with the other data[2]. The implied premise is that a method of analysis that is insensitive to outliers will be more accurate. But this premise is not correct; for example, *predetermining* the estimate of G to be some convenient constant is a procedure that is completely insensitive to an outlier! By this *reductio ad absurdum*, we see that we must also examine the effect of choosing a 'robust' procedure on the influence of correct data as well; the procedure might downweight such data unjustifiably.

The idea of invoking a 'robust' method implies that we accept that some data might have been generated under a mechanism different from the model. Surely, some of these unwanted data might *not* appear as outliers, and surely we would also want our method to be insensitive to the values of these data. This idea is relevant to Dose's examination of the sensitivity of results to the value of a fictitious measurement [1, figure 2]. He shows that when this fictitious value is varied while *distant from the bulk of measurement values* the estimates obtained using his two non-Gaussian likelihoods are altered less than the estimate obtained using the Gaussian likelihood[3]. However, it can be seen from his diagram that when the fictitious measurement has a value near the middle of the cluster of measurements on the x-axis the sensitivity of the estimates with the non-Gaussian likelihoods is in fact greater than with the Gaussian likelihood. That is, the non-Gaussian likelihoods are less robust than the Gaussian likelihood to a datum that does not appear as an outlier but which should be discounted.

3. Prior distributions

The first prior distribution encountered in [1] is a flat distribution for the description of belief held about G prior to observing the data. It would be difficult to argue for the use of any other prior distribution given that some of the data were known as far back as 1982.

The second prior distribution used in [1] is assigned to a common factor by which each standard uncertainty is

[2] We may distinguish between the concepts of insensitivity to the *presence* of an outlier and insensitivity to the *value* of the outlier [3].

[3] Dose claims that the figure shows 'a clear saturation' [1, p 181]. But the word 'saturation' implies the existence of an asymptote—and there are no asymptotes.

supposed to have been underestimated or overestimated, α. This prior distribution increases without bound towards the origin, as if it is *a priori* thought most likely that the correct uncertainties have been overestimated by an infinite common factor! Presumably, if those contributing the data $\{\sigma_i\}$ are competent then any prior distribution for α should have most of its density near unity. That is, the natural origin for the prior distribution is 1 not 0—so the prior distribution used in [1], which is a common 'non-informative' prior for an unknown scale parameter, is inappropriate.

And what possible mechanism could exist by which all experimenters underestimate or overestimate their assessment of precision by an unknown equal factor? Where in the measurement process might such a mistake be made? If we are to assume that some equal misjudgement is made by every laboratory then this misjudgement presumably represents some improperly assessed source of uncertainty common to all experiments, and it seems far more realistic to suppose that such an error acts *additively*; see section 5.1. Dose acknowledges that the assumption of a constant multiplier α is 'rather unrealistic' for this problem [1, p 182].

The third prior distribution described in [1] is the discrete uniform distribution used to weight equally the three models associated with the different likelihoods. There seems no reason why the three distributions chosen—the Gaussian, Laplace and hyperbolic-cosine distributions—are representative of all possible likelihood functions. So, why should our attention be restricted to them and why in equal measure?

Dose has felt sufficiently informed to restrict the likelihood models envisaged to these three forms (and he has not given a justification). Yet he has felt sufficiently uninformed to be unable to prefer any one of these forms to any other. But his introduction of the Gaussian likelihood uses the idea that the Gaussian distribution (uniquely) maximizes the relevant figure of entropy. If this principle of maximum entropy carries any weight at all then the normal likelihood must be the best candidate, by definition!

4. Posterior distributions and interpretation

As Dose implies [1, p 177], the fundamental output of a Bayesian analysis is the posterior distribution. He chooses to summarize the posterior distribution of G by its mean and variance, which are undoubtedly the best summary figures if the value of G is to be included in another analysis. However, no percentile of a general distribution can be calculated using only the mean and variance so, unless symmetry is assumed, no probability statement about G itself can subsequently be made, with the exception of a weak statement corresponding to the (Bienaymé–)Chebyshev inequality. That is, almost all the quantified 'belief' about G itself is lost when summarizing the posterior distribution by its mean and variance.

Our only comment about the specific results obtained in [1] relates to the posterior distributions of α in the final analysis under the Gaussian, Laplace and hyperbolic-cosine likelihoods. These distributions have means of 3.1, 2.4 and 2.3 [1, table 4]. Thus, according to the analysis, each group of scientists contributing the data has underestimated the standard deviations of the error distributions by a factor as large as 2! Who could accept an analysis that implies this? The idea that there is a sizeable constant scaling factor affecting every quoted uncertainty would be unacceptable to most critical readers. Yet the results of the analysis rest on this idea.

Dose writes that his final result is 'characteristic for the given data set' but that the same least-squares results can be obtained from a infinite number of datasets [1, p 181][4]. It is unclear what 'characteristic for the given data set' means and what the relevance of this idea would be. Clearly, a dataset in which the difference of each d_i from the final estimate was reversed in sign would give exactly the same posterior distributions in this analysis, and obviously the results depend heavily on the likelihood models and prior distributions assumed.

We now come to the issue that might be regarded as the most troubling. As noted, the uncertainties attributed to the input data $\{d_i\}$ are regarded as being (multiples of) the *standard deviations of the distributions from which the input data are drawn*. Yet the uncertainty attributed to the final result is the *standard deviation of the posterior distribution calculated for the measurand*. We can see a basic change in the interpretation or description of 'uncertainty' during the analysis. Further evidence of this ambiguity is given by Dose's statement at the end of his paper that the quoted uncertainties should rather be taken as 'the point estimates of the true uncertainties' [1, p 182]. He, therefore, seems to leave the Bayesian interpretation and return to the classical understanding that a quoted uncertainty is an estimate of the standard deviation of the distribution from which the input datum was drawn. If such internal inconsistency is acceptable to readers then one might wonder if the concept of uncertainty has any meaning at all.

5. Weighted mean

Our focus now shifts to comments in [1] about the (inverse-variance) weighted mean

$$\tilde{x} = \frac{\sum_i d_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

as an estimate of G and about the figure

$$\sigma_{\tilde{x}} = \left(\frac{1}{\sum_i 1/\sigma_i^2}\right)^{1/2}$$

as the standard uncertainty of $\tilde{x}$.

Dose calls $\tilde{x}$ the least-squares result but does not clearly state the corresponding model[5]. The model under which $\tilde{x}$ becomes the (weighted-)least-squares solution and $\sigma_{\tilde{x}}$ becomes the standard error is that

> 'each d_i is drawn independently from a distribution with mean G and standard deviation σ_i'.

Under this model $\tilde{x}$ is the realization of a random variable that is the minimum-variance linear-unbiased estimator (MVLUE) of G and the quantity $\sigma_{\tilde{x}}$ is the standard error

[4] Could it be argued that an analysis that gives the same results from an infinite number of datasets must be 'robust'?!

[5] An unambiguous model generally encourages an appropriate means of analysis. Conversely, if a certain means of analysis has become a 'pet tool' [1, p 181] then perhaps there has been no habit of specifying a model.

of this estimator (i.e. the standard deviation because of unbiasedness)[6]. These properties hold whether or not the distributions are Gaussian and whether or not the distributions have the same form [2]. So, whether the criticism in [1] is directed at the usefulness of $\tilde{x}$ or at the usage of $\tilde{x}$, the assertion that '[$\tilde{x}$ and $\sigma_{\tilde{x}}^2$] rests on the strong assumption that $\{d_i\}$ are samples from Gaussian distributions' is misleading.

The failure to state a model also relates to other comments made about $\tilde{x}$. In particular, it relates to the observation that its standard uncertainty, $\sigma_{\tilde{x}}$, is not altered if the difference between each datum and $\tilde{x}$ is multiplied by a common factor [1, p 176]. This property is not a failure of $\tilde{x}$, which finds its *raison d'être* in the model above, but is a failure of the observer who considers that $\tilde{x}$, or any other estimate, might be used without specifying a model. If we envisage that the data might exhibit a greater amount of spread than is implied by $\{\sigma_i\}$—and that the measure of uncertainty should reflect this greater spread—then we should state a model that allows for such extra variation. If such a model was stated then it would involve at least one additional parameter, and it would be obvious that $\tilde{x}$ and $\sigma_{\tilde{x}}$ ought to be discarded in favour of figures that depended on that parameter.

This idea that the uncertainty $\sigma_{\tilde{x}}$ is not altered when the difference between each datum and $\tilde{x}$ is adjusted prompts Dose to write 'the uncertainty [$\sigma_{\tilde{x}}$] ... associated with the estimate [$\tilde{x}$] ... depends only on the set of data uncertainties but not on the scatter of the data' [1, p 176]. He goes on to state that his result 'provides a rigorous uncertainty estimate which relies on both the scatter of the data and the quotation of their experimental uncertainties' [1, p 181]. With this in mind, we consider the effect on his results of his assumption that there is a constant scaling factor α for all the uncertainties and his choice of prior distribution for α. On taking the limit, which in effect is a step taken in [1], this prior distribution [1, equation (14)] becomes the improper prior $p(\alpha) \propto 1/\alpha$. For any known constant k, the corresponding prior distribution of $k\alpha$ has the same form, $p(k\alpha) \propto 1/(k\alpha)$, which means that we consider ourselves to have the same prior knowledge of $k\alpha$ as we do of α. This implies that if all the quoted uncertainties were multiplied by $1/k$ then the posterior distribution of $\hat{G}$ will be unaltered[7]! Thus, Dose has obtained an estimate for which the *uncertainty depends on the scatter of the data but not on the scale of the set of experimental uncertainties!*

5.1. *A modified weighted mean and an alternative analysis*

We now show how the weighted mean can be modified in a reasonable way for use in this problem. Figure 1 shows the data used by Dose to obtain his final result numbered by the row in which they appear in his table [1, table 3]. There is no need to carry out any statistical procedure to see that there is an excessive amount of spread in the data in relation to the quoted uncertainties. So the data cannot be thought of as arising under the basic model given above. One or more of the data pairs (d_i, σ_i) must, in some sense, be erroneous.

Clearly, there is some component of measurement error incompletely represented in one or more of the σ_i values. If we acknowledge that this error can be large enough to make the corresponding d_i appear as an outlier then we should also acknowledge that there might be smaller errors not represented in the standard uncertainties of other estimates. (In fact, these smaller errors seem more likely to occur than the larger ones.) Thus, like Dose with the multiplier α, we are led to modify the model by including a parameter to describe such an effect. As implied in section 3, it seems natural for this effect to act additively rather than multiplicatively. So our model is that

> 'each d_i retained is drawn independently from a normal distribution with mean G and variance $\sigma_i^2+\sigma^2$'

and our ability to estimate G is not affected markedly by the introduction of the single unknown parameter σ^2. The implied assumption of normality for the acknowledged errors has the backing of the central limit theorem. The implied assumption that the extra errors for the data retained are drawn from a single normal distribution is supported by arguments given in [5, sections 4.2, 4.3]. In particular, the extra error for a laboratory is taken to be the sum of one or more subtle random variables of which that laboratory is unaware when carrying out the measurement. If metrologists of every laboratory are regarded as being equally skilful and equally careful then the same distribution should be assumed for each of these errors. The requirement that this extra distribution be normal will be much less important than the requirement that the original distributions be normal because we expect the variance of this extra distribution to be comparatively small (which the results suggest it is in this case).

The maximum-likelihood estimate $\hat{G}$ of G under the model is obtained in two stages. We minimize

$$\sum_i \log\left(\sigma^2+\sigma_i^2\right) + \sum_i \frac{\left(d_i - \frac{\sum_k d_k/\left(\sigma^2+\sigma_k^2\right)}{\sum_k 1/\left(\sigma^2+\sigma_k^2\right)}\right)^2}{\left(\sigma^2+\sigma_i^2\right)}$$

[4, equations (6)–(9)] [5, equation (9)] to find the maximum-likelihood estimate $\hat{\sigma}^2$ of σ^2 and then calculate the estimate

$$\hat{G} = \frac{\sum_i d_i/\left(\sigma_i^2+\hat{\sigma}^2\right)}{\sum_i 1/\left(\sigma_i^2+\hat{\sigma}^2\right)},$$

which is seen to be a modified weighted mean. The standard uncertainty of this estimate may be taken as

$$\sigma_{\hat{G}} \approx \left(\frac{1}{\sum_i 1/\left(\sigma_i^2+\hat{\sigma}^2\right)}\right)^{1/2}.$$

(Note also that if we take σ^2 to be equal to $\hat{\sigma}^2$ then $\hat{G}$ is the MVLUE of G and $\sigma_{\hat{G}}$ is the corresponding standard uncertainty even if we relax the assumptions of normality.)

Each datum in turn is tested for compatibility with the model fitted to the other data by the iterative method described in [5, step 1]. A datum identified as incompatible is excluded

[6] The property that the weighted mean is the MVLUE is associated with the Gauss–Markov theorem [2]. This property is analogous to a well-known property of the sample mean when estimating the mean of a distribution from a random sample. Other properties of the sample mean are listed in [3]. It might be expected that the weighted mean will have properties analogous to some of these.

[7] For example, suppose we estimate G from an independent sample drawn from some distribution with unknown mean G and variance $\lambda \equiv \alpha\sigma$, where σ is known. If G and α are given the usual non-informative joint prior density $p(G, \alpha) \propto 1/\alpha$, as in [1], then G and λ have the same prior density $p(G, \lambda) \propto 1/\lambda$. Therefore, whatever the likelihood model, the analysis leading to the posterior distribution of G is the same as if we had no knowledge of the variance at all. Adopting the marginal prior $p(\alpha) \propto 1/\alpha$ has, in effect, ignored the information provided by our knowledge of σ. Thus, the posterior distribution of G is unaffected by the value of σ.

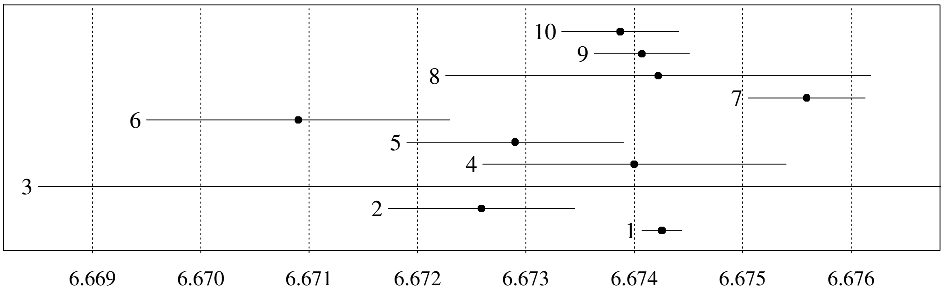

Figure 1. Estimates of G in multiples of 10^{-11} m^3 kg^{-1} s^{-2}. The data are indexed by $i = 1, \ldots, 10$, where i is the row number in [1, table 3]. A dot indicates d_i and the corresponding bar indicates the interval $[d_i - 2\sigma_i, d_i + 2\sigma_i]$.

from the final analysis. Such a datum may be interpreted as having been subject to an extra error drawn from a distribution with variance larger than σ^2. The method is constructed so that the probability of data being excluded when none should be excluded is kept at or below 0.05.

When this method is applied to the dataset of figure 1, points 7, 6 and 2 are removed from consideration in turn, the corresponding 'p-values' being approximately 10^{-5}, 0.001 and 0.006. The results obtained from the remaining seven points are $\hat{G} = 6.674\,08 \times 10^{-11}$ m^3 kg^{-1} s^{-2} and $\sigma_{\hat{G}} \approx 0.000\,12 \times 10^{-11}$ m^3 kg^{-1} s^{-2}, with $\hat{\sigma} = 0.000\,13 \times 10^{-11}$ m^3 kg^{-1} s^{-2}.

The method we have employed takes into account the quoted uncertainties and also the scatter of the estimates. It is 'robust' through the identification and removal of outliers. The analysis treats each estimate on the same basis—even if some are ultimately excluded from the final equation. In the particular case studied here, the proportion of data excluded from the analysis, 3 out of 10, is large. Consequently, a more complicated but *defensible* model might be sought through further consideration of the experiments that generated the data. One possible modification would be to allow different values of σ^2 for different subsets of measurements grouped according to the experimental technique employed [4, section 5].

5.2. On the issue of inconsistent data

The analysis in section 5.1 involved the exclusion of three data pairs to leave a set that is internally consistent. This step was influential in allowing the analysis to give a result with a standard uncertainty half the size of that obtained by Dose. Many readers will, presumably, not accept the figure of uncertainty obtained because they will see the excluded data as being meaningful. The existence of different approaches to the treatment of inconsistent data may reflect different opinions about the very meaning of such inconsistency[8].

Dose's understanding is that 'once an experiment has been identified as valid and state of the art, the numerical result of its measurements, point estimate and uncertainty should be accepted' [1, p 178], which we suppose means not subject to further scrutiny[9]. But can we be so confident in our auditing of the individual experiments as to conclude that no problem is present simply because no problem is seen? It seems improper not to carry out a subsequent audit of the resulting dataset to determine consistency.

The dataset above, for example, will fail this test of consistency; the data are informing us that there is a problem and we cannot pretend otherwise[10]. There has obviously been some unknown effect at work in the generation of this data. If each datum is to contribute to the final estimate of G and if no justifiable model can be developed to describe this effect then perhaps we ought to quote an artificially large standard uncertainty to accommodate this effect. Remedial action of some sort is required, be it (i) the meaningful and legitimate modification of the model, such as in section 5.1, (ii) the exclusion of data by some statistically defensible procedure, such as in section 5.1, or (iii) the explicit quotation of a 'catch-all' uncertainty, as perhaps in the 'arbitrary enhancement of the calculated uncertainty by the CODATA committee' [1, p 178]. A full dataset that is discordant under all reasonable models, such as the dataset of figure 1, cannot legitimately be treated as consistent through the choice of an ad hoc model. Inconsistent data cannot be combined without loss of scientific meaning.

6. Conclusion

The claim is made in [1] that the methodology employed is rigorous. The *mathematical* analysis may be rigorous, but many arguments given in this comment suggest that the logic behind the assumptions involved is not. The maintenance of logic and meaning in the analysis seems secondary to (i) a reluctance to censor inconsistent data and (ii) an associated emphasis on achieving robustness to outlying observations. Both the likelihood functions and prior distributions used in [1] seem constructed with this narrow view of robustness in mind. Meaning and accuracy are compromised by admitting likelihoods that do not possess any clear connection with a mechanism through which the data might have been created and by weighting these likelihoods as strongly as the Gaussian likelihood, which has prior reason to be favoured.

The non-Gaussian likelihoods were *chosen* from an infinity of possibilities. Bayesian statistics is *not about choice*;

[8] For example, what understanding of the nature of data permits Dose to write 'Clearly the least squares result overestimates the precision' [1, p 180]? What is the basis for this assertion and how is it clear?

[9] The requirement that the experiment is 'state of the art' seems unnecessary for the numerical results to be trustworthy; are estimates of uncertainty obtained in earlier experiments invalid?

[10] Bayesian statisticians claim that their methods utilize all the available information (though it is common to find that every prior distribution used is designed to be 'non-informative'). Here, there is information contained in the fact that the data set is inconsistent.

it is about realistic likelihoods and genuine belief (or, in practice, approximations to genuine belief). Where there is no genuine belief, the quantification of prior ignorance or impartiality might be thought possible through the concept of so-called 'non-informative' prior distributions, but the use of such distributions, especially improper distributions, is controversial.

Many people believe that Bayesian methodology is preferable to non-Bayesian methodology. However, any superiority of the theory of Bayesian statistics does not seem to translate into practice, where prior distributions used are often not genuine and where analysts seem too much at liberty to choose the distributions that will affect the results. In this author's opinion, the analysis in [1] suffers from such inadequacies. The analysis uses (i) an inappropriate non-informative prior for a parameter α that cannot be said to represent any existing quantity, (ii) a non-informative prior distribution for weighting a restricted (i.e., heavily 'informed') set of likelihood functions chosen without proper justification and (iii) two different interpretations of uncertainty. Consequently, the results have no clear meaning and should be disregarded.

References

[1] Dose V 2007 *Meas. Sci. Technol.* **18** 176–82

[2] Odell P L 1983 Gauss–Markov theorem *Encyclopedia of Statistical Sciences* vol 3 ed S Kotz, N L Johnson and C B Read (New York: Wiley) pp 314–6

[3] Willink R 2007 *Metrologia* **44** 105–10

[4] Willink R 2002 *Metrologia* **39** 343–54

[5] Willink R 2006 *Metrologia* **43** S220–30