Applied Stochastic Models in Business and Industry

(wileyonlinelibrary.com) DOI: 10.1002/asmb.1966

Published online in Wiley Online Library

# Discussion on 'Five examples of assessment and expression of measurement uncertainty' by Antonio Possolo

## 1. Introduction

The five examples in this paper present a varied and interesting collection of contexts in which models and methods for assessing measurement uncertainty are used. Consistent with Dr. Possolo's assertion that 'probability distributions are well suited to express measurement uncertainty', his paper presents and advocates the use of probability models to guide simulations whose variability expresses the uncertainty connected with a measurement.

Actually, although this is not always clear in Dr. Possolo's discussion, there is more than one kind of uncertainty involved with measurement. We must differentiate between *pre-data uncertainty* and *post-data uncertainty.* I say more about this distinction in Section 2 of this discussion, and this distinction serves as a guiding theme for my comments on Dr. Possolo's examples (and the methodology used in these examples) in Section 3. Section 4 concludes with some remarks about the question of how best to communicate information about uncertainty of measurement.

## 2. Pre-data uncertainty and post-data uncertainty

Pre-data uncertainty is concerned with the variability about the true value of the measurand of the values produced by a particular method of measurement—it is uncertainty about the accuracy of the measurement methodology. In statistics, indices of variability of point estimators or coverage probabilities of confidence intervals are used in the *frequentist paradigm* of inference to select procedures that will give the most accurate measurement, resulting in minimal uncertainty. Distributions of the point estimators provide not only these indices but also more detailed insight into the performance of the measurement method. Examples of such an approach in Dr. Possolo's paper are the arsenic-in-oyster-tissue example of Section 3 and the viscosity example of Section 4. In the arsenic-in-oyster-tissue example, the parametric bootstrap is used to estimate the distribution of the maximum likelihood estimator of the consensus value $\mu$ under the assumption that $\mu$ has the value $\hat{\mu}$ attained by the maximum likelihood estimator. One could instead simulate the *exact distribution* of the estimator when a particular value of the measurand holds. If the value specified for the measurand is correct, the distribution will be the correct pre-data uncertainty of measurement. The bootstrap distribution thus estimates the correct pre-data distribution of the measurement rather than giving it exactly. There is an unfortunate tendency in the literature to regard this bootstrap distribution as a post-data expression of uncertainty because the data are involved in choosing which distribution to simulate, but this is not the case—what is obtained is an estimate of the correct pre-data distribution of values for the measurement.

When expressed in terms of distributions, post-data uncertainty commonly takes the form of a distribution for the measurand (not the measurement) conditional upon observed measurement values. To obtain such a distribution, we start with a prior distribution for the measurand that expresses our uncertainty about the value of the measurand before the data are obtained. The posterior distribution is then obtained by Bayes theorem from the prior distribution and the conditional distribution of the data given the value of the measurand. Comparing the posterior distribution with the prior distribution reflects the change in our uncertainty (or belief) about the measurand value due to the information provided by the data. A follower of the Bayesian paradigm of statistical inference will make inferences about the value of the measurand based on the posterior distribution.

Although both a prior distribution for a measurand and a distribution for a measurement are valid probabilities over the same space of possible values, they are conceptually very different entities. Different people can have different prior distributions (which is why these distributions are said to be *subjective*), but the distribution of the data given the true value of the measurand is unique and can be observed by everyone (and is therefore said to be *objective*). It is the arbitrary (and thus subject to manipulation) nature of prior distributions that has made the Bayesian paradigm controversial among statisticians, but if a prior distribution is honestly constructed as a model of one's belief, it can aid in clarifying communication between observers, giving values to subjective judgements that otherwise may have gone unreported in classical scientific experiments.

The Bayes paradigm is also a natural way to learn from a series of measurements or experiments because each posterior distribution from an earlier measurement or experiment is used as the prior distribution for the next measurement or experiment. In parametric model contexts, it can be shown that as data accumulates, the prior distribution held by an individual at the start of the process becomes less and less consequential. In the end, as long as each individual's prior distribution gives positive probability to an arbitrarily small interval containing the true value $m_0$ of the measurand, the posterior distributions of all individuals will (as data accumulates) tend to the same distribution concentrated at $m_0$ even though their prior opinions differed considerably.

One of the difficulties I had in reading Dr. Possolo's examples is that it was not always clear whether he was using pre-data or post-data distributions to generate his simulations. In his Section 2, which involves a situation of the type covered by the 'Guide to the expression of uncertainty in measurement' (GUM), he tries to cover all bases and in his sixth paragraph warns his readers that his results are only completely valid when all variances or uncertainty distributions are of the posterior distribution type. Otherwise, his procedures 'fail to propagate all the uncertainty in play'. The general situation considered by the GUM assumes that some of the inputs (type A) are measured (are estimated from data) and come with (estimated) pre-data variances as indices of uncertainty, whereas other inputs (type B) have no data associated with them (but can be assigned 'prior' distributions using expert opinion or posterior information from previous experiments). The GUM advocated using a Taylor series expansion of the function relating the inputs to the measurand, along with the given pre-data variances of the type A inputs and the 'prior' variances of the type B inputs, to estimate the variance of the measurand as an index of uncertainty. Such a confusion of the two types of uncertainty distribution concerned some of the statisticians at the National Institute of Standards and Technology and led to an attempt on my part to reconcile the GUM approach with statistical theory [1]. In the example treated by Dr. Possolo in Section 2 of his paper, his inputs are assumed to be all type B. No measurements are indicated, no distributions for the measurements given the correct value of the inputs are specified, and Dr. Possolo himself indicates that the angles are given somewhat arbitrarily chosen distributions. In this case, the uncertainty distribution simulated for the index $n$ of refraction of the prism is of the post-data (subjective) type and represents our uncertainty about the value of the measurand $n$. Contrastingly, if *all* of the input distributions were pre-data (objective, type A); the resulting distribution would then also be pre-data (objective) but would reflect our uncertainty about the accuracy of the measurement (estimate) of $n$, not of the value of the measurand. If one follows the GUM by mixing pre-data distributions for some inputs with post-data (or prior) distributions for other inputs in the simulations, the result would be neither mathematically correct nor scientifically meaningful.

## 3. Comments on Dr. Possolo's examples

I will discuss Dr. Possolo's five examples in order of their appearance.

### *3.1. Refractive index*

I have already commented on this example in Section 2, but let me note here that I find the joint distribution used for the apex angle and total deviation to be a bit unrealistic, in that the apex angle $\alpha$ clearly must be acute and thus takes values only in a proper subset of the set of values where the von Mises density is positive. Also, the extent of deviation $\delta$ is restricted by the value of $\alpha$ (and thus cannot be statistically independent of $\alpha$). The distribution for the triple $(m, \alpha, \delta)$ used by Dr. Possolo is unlikely to be a posterior distribution, although it might be a possible prior distribution for someone unfamiliar with the restrictions on values imposed by the experimental context. It would be instructive to see this example worked out in full with specification of the following: (i) a more realistic joint prior distribution for the inputs (including the inputs to the estimate of $m$) and (ii) the pre-data frequentist distribution of the measurements.

### *3.2. Arsenic in oyster tissue*

This example illustrates how uncertainty about a certain type of measurement (pre-data) or about the value of a measurand (post-data) must depend on the level of generality of the context. Here, there is a difference between the uncertainty (pre-data or post-data) resulting from a measurement taken at a specific laboratory and the corresponding type of uncertainty when measurements can come from any of several labs. Thus, my doctor prefers that I use the same laboratory for all of my medical tests because there is an extra level of uncertainty that results if I obtain my tests at any of several (possibly randomly selected) laboratories.

The pre-data (frequentist) mixed-model assumptions that Dr. Possolo uses have a rich history at the National Institute of Standards and Technology [2,3]. Here, the data refer to only one measurand measured at several laboratories. A greater level of generality would be studies of several measurands each measured at several laboratories or, even more generally, several measurands each measured by more than one method (e.g., wet chemistry versus dry chemistry) at each of several laboratories. When the measurands are different levels of the same chemical, W.G. Cochran proposed a specialization of this model in which the laboratory bias was assumed to be linear in the level of the chemical measured (and the slopes and intercepts of these linear functions differed randomly from laboratory to laboratory). These generalizations were studied by my Ph.D. student, Marc Sylvester, in his dissertation [4].

In this example, Dr. Possolo notes that the estimated lab biases deviate from a Gaussian distribution but does not explain why these *estimates* of the true biases should be expected to have a Gaussian distribution if the heteroscedastic, Gaussian, linear mixed model holds for the data. The alternative model that Dr. Possolo suggests is reasonable and may indeed be a better fit to the data but requires considerably more computation to analyze. Indeed, to fit this model to the data, Dr. Possolo has to adopt prior distributions for the measurand, the variance of the random lab biases, the error variances of the laboratories, and a degree of freedom parameter, none of which he necessarily believes (although he argues that his choice will not affect the posterior distribution very much). The uncertainty distribution obtained for the measurand is of the post-data (posterior) type in contrast to the (estimated) pre-data uncertainty distribution of the maximum likelihood estimator obtained by the parametric bootstrap. In this respect, Dr. Possolo illustrates the comment that many applied statisticians make about the frequentist Bayesian controversy: ‘We use the approach that seems to work best for our problem’.

### *3.3. Viscosity*

Assuming that the concentrations $c$ are known (or have negligible uncertainty), the model assumed for the viscosity example is just classical linear regression, and the measurand $\kappa$ of interest is the ratio of the slope to the square of the intercept. The joint distribution of the least squares (and maximum likelihood) estimates of intercept and slope is bivariate Gaussian, and thus it is a straightforward exercise to find the distribution of the ratio of the estimated slope to the square of the estimated intercept. This distribution is likely to be a little bit messy but, at worst, can be expressed as a one-dimensional integral and solved numerically by quadrature. Similarly, one can obtain the variance of the estimate (a type A measure of uncertainty) as a function of the true intercept and slope and then plug in the estimates of intercept and slope into this function to estimate the pre-data uncertainty of the maximum likelihood estimate of $\kappa$. The parametric bootstrap is an alternative way to compute this same estimate but does require the choice of the number $M$ of bootstrap replications to achieve a desired accuracy of computation. The delta method (Dr. Possolo’s Equation (1)) provides a different, and usually less accurate, estimator but requires the least computation and has the advantage of indicating the relative contributions of the estimated intercept and estimated slope to the (pre-data) uncertainty. I am not sure that the viscosity example was the best way to introduce modeling of joint distributions through the copula because the appropriate copula for a distribution with normal marginals is precisely the one that yields the correct bivariate normal distribution of the estimated intercept and slope. In general, the joint distribution of estimates is determined by the modeling assumptions for the observed data. Imposing a joint distribution as illustrated in this example can at best estimate the true joint distribution, with the closeness of estimation unknown.

Copulas do have their use in determining uncertainty—namely in posing more realistic joint prior distributions for inputs (and/or parameters) than that of statistical independence. Bayesians have suggested good ways to elicit prior opinion about individual quantities; these methods can be used to elicit marginal prior distributions. It is also often possible to elicit opinions about dependency of pairs of quantities in terms of (squared) correlations. Copulas provide a good way to combine this information into a joint distribution that may, hopefully, approximate an individual’s true prior beliefs. Also, using a copula depending upon just a single correlation, as well as varying that correlation while holding marginal distributions fixed, provides a useful way to determine sensitivity of the posterior distribution of a measurand to dependency in the prior distribution of inputs.

I remark that bootstrap ‘confidence intervals’ for ratios of regression parameters have zero confidence for any sample size, making such an interval estimator of doubtful use as a measure of uncertainty in this example [5,6].

### *3.4. Fukushima*

This is perhaps Dr. Possolo’s most interesting example. It presents a novel design and methodology to evaluate various components of pre-data uncertainty, as measured by the relative standard deviations due to the following: (i) interpolation methods; (ii) interpolations; and (iii) measurement error for the measurements of radioactivity proper. The component of uncertainty due to the interpolation method is estimated as a main effect on the basis of the measurements taken at the various locations. Because the smoothing methods did not allow for replications at the locations, the component of uncertainty for measurement error had to be taken from the operating manual. In other applications, it would be determined from replications. Finally, the fact that the choice of locations was not entirely random leads to the use of designed cross-validations to estimate the interpolation error. With these components of the total pre-data uncertainty treated as statistically independent, their estimated variances are combined using root mean square to yield an estimated pre-data uncertainty. The dominant source of uncertainty is the interpolations, whereas the method of interpolation and the measurement error of radioactivity contribute approximately equal amounts to the uncertainty. It should be noted that comparison of interpolation methods was equivalent in this case to comparison of models for the data because the two methods compared made different assumptions about the distribution of the data. In other analyses of pre-data uncertainty, a component of uncertainty that usually needs exploration is the model assumed for the data. This would be the case, for example, when doing an analysis of the arsenic-in-oyster-tissue measurement uncertainty.

### *3.5. Deepwater Horizon*

An analysis similar to that of the Fukushima example can be performed on post-data uncertainty measures. The components of uncertainty would include the choice of prior distribution as well as the choice of (frequentist) models for the data. The Deepwater Horizon example illustrates one aspect of such an analysis—the contribution of the choice of prior distribution to the uncertainty. The uncertainty component considered here is actually not of the choice of prior distribution, but rather of two ways to combine the prior (subjective) uncertainty distributions of five experts: (i) simple averaging of the densities or (ii) the geometric mean of the densities (logarithmic averaging).

The simple averaging of the densities always produces a density, whereas the geometric average is not a density because the total area under the curve (or total probability in general) is less than 1. Multiplying the geometric average by a constant of proportionality produces a density $\pi(\mu)$. Because the accept/reject sampling method does not involve the constant of proportionality, it follows that using the accept/reject method on the geometric average of the experts' densities to draw a sample yields observations from $\pi(\mu)$ . When the experts' densities are all Gaussian in form (as Dr. Possolo assumes), the geometric average of the experts' densities is proportional to a Gaussian density with mean and variance respectively given by

$$\hat{r} = \frac{\sum_{i=1}^{5} \left(m_i / v_i^2\right)}{\sum_{i=1}^{5} \left(1 / v_i^2\right)} \quad \text{and} \quad u^2(\hat{r}) = \frac{1}{\frac{1}{5} \sum_{i=1}^{5} \left(1 / v_i^2\right)} ,$$

where $m_i$ and $v_i^2$ are the means and variances of the individual experts' densities, $i = 1, 2, 3, 4, 5$. There is no need to sample from this density to obtain a probability interval; the minimum-length 95% probability interval obtained from this density is well known to be $[\hat{r} - 1.96\, u(\hat{r}),\ \hat{r} + 1.96\, u(\hat{r})]$ .

Note that the mean and variance of the density yielded by the (arithmetic) average of the experts' densities are

$$\hat{q} = \frac{1}{5} \sum_{i=1}^{5} m_i \quad \text{and} \quad u^2(\hat{q}) = \frac{1}{5} \left[ \sum_{i=1}^{5} v_i^2 + \sum_{i=1}^{5} (m_i - \hat{q})^2 \right] ,$$

respectively. Comparing the variances of the densities produced by the two pooling methods, we can easily see from the arithmetic mean/harmonic mean inequality that $u^2(\hat{q})$ is larger than $u^2(\hat{r})$ even when all of the experts' means are the same. In the Deepwater Horizon example, the means are not the same. Also, whereas the geometric mean of the experts' densities produces a unimodal density, the arithmetic mean of the densities has two peaks.

The method of taking a geometric average of the prior densities of several experts or observers is said to yield a 'group Bayesian' method of Bayesian analysis, in the sense that this 'prior' (apart from the constant of proportionality) yields a posterior distribution that (again, apart from the constant of proportionality) is the same as the geometric average of the posterior distributions of the individual experts. Thus, with the use of the group Bayesian methodology, the pool of experts behaves like a single Bayesian in learning from repeated experiments. In contrast, the posterior distribution obtained from the average of the experts' priors is not the simple average of their individual posterior distributions but instead is a weighted average, where the weights reflect how well the data 'agree' with the experts' prior opinions. This property makes weighted averages of distributions useful for model selection purposes, but not particularly useful for learning from data.

## 4. Communication of uncertainty of measurement

A measurement, particularly of a standard, can be used for many purposes by different individuals. The contexts in which such measurements are used dictate which components of the uncertainty of such measurements come into play. Our uncertainty tends to be greater (more dispersed), the broader the context of usage is. Thus, as has already been noted, the uncertainty of a measurement when it is always to be taken from a particular laboratory by a particular method will not involve components of uncertainty due to choice of laboratory or choice of method of measurement. Evaluating the uncertainty of measurement involves the often different prior beliefs of users of the measurement. At a minimum, when measurement is only used by the individual taking the measurement, the information needed to evaluate uncertainty involves only the models for the pre-data distribution of any data input used for type A inputs, a prior distribution reflecting what information the individual brings to the analysis of the measurement (including information about type B inputs), and the type A input data. When multiple users are involved, communication between the experimenter and the users becomes more complicated. Although it is tempting to use a consensus or pooling of user beliefs, the Deepwater Horizon example shows that different methods of pooling prior information can produce quite different results. In the end, each user of the measurement information will want to use his or her own prior distribution to evaluate uncertainty. The current response of Bayesians to this problem is to talk about robust Bayesian methodology, but such a methodology can only be used once, and repeated usage of the same methods in a series of measurements does not produce the ‘prior to posterior to prior to posterior’ consistent method of learning, which is one of the great advantages of the Bayesian paradigm. Instead, what needs to be communicated for type A inputs to measurement is the *likelihood function* (the joint density of the data given the unknown parameters or measurands expressed as a function of these unknown quantities). Such a function is not always available in a parsimoniously expressed form suitable for publication or communication, and research needs to be carried out on general ways to provide an approximation to the likelihood function that can be easily communicated and for which the contribution of the errors of this approximation to the uncertainty of measurement can be accurately evaluated. Central limit and saddle-point approximations are possible approaches to this problem, but my own research and that of my students [7, 8] have been to use the computer to fit spline approximations to the logarithm of the likelihood to achieve these goals.

I congratulate Dr. Possolo for a very stimulating paper.

LEON J. GLESER
*Professor of Statistics*
*University of Pittsburgh*
*Pittsburgh, PA, USA*
E-mail: *gleser@pitt.edu*

## References

1. Gleser LJ. Assessing uncertainty in measurement. *Statistical Science* 1998; **13**(3):277–290.
2. Rukhin AL, Vangel MG. Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association* 1998; **93**:303–308.
3. Rukhin AL, Sedransk N. Statistics in metrology: international key comparisons and interlaboratory studies. *Journal of Data Science* 2007; **5**:303–312.
4. Sylvester MA II. Estimation of a common mean through a series of similar interlaboratory experiments. *Ph.D. dissertation*, University of Pittsburgh, 2001.
5. Gleser LJ, Hwang JT. The nonexistence of 100(1 − alpha)% confidence sets of finite expected diameter in errors-in-variables and related models. *Annals of Statistics* 1987; **15**(4):1351–1362.
6. Gleser LJ. Comment on “Bootstrap Confidence Intervals” by Thomas J. DiCiccio and Bradley Efron. *Statistical Science* 1996; **11**(3):219–221.
7. Sezer A. Representing uncertainty by spline function approximation of log-likelihood. *Ph.D. dissertation*, University of Pittsburgh, 2006.
8. Song MS. Numerical approximation of the likelihood of correlation matrices. *Ph.D. dissertation*, University of Pittsburgh, 2010.