ACTA IMEKO
ISSN: 2221-870X
December 2015, Volume 4, Number 4, 30-37

Type A evaluation of measurement uncertainty when the sample size is not predetermined

Robin Willink

16 Rochdale Avenue, Glendowie, Auckland 1071, New Zealand

ABSTRACT
The archetypal procedure in Type A evaluation of measurement uncertainty involves making *n* observations of the same quantity, taking the sample figure s^2 to be an unbiased estimate of the underlying variance and quoting the figure $s/\sqrt{n}$ as the relevant standard uncertainty. Although this procedure is theoretically valid when the sample size *n* is fixed, it is not necessarily valid when *n* is chosen in response to the growing dataset. In fact, when the experimenter makes observations until a certain level of uncertainty in the mean is reached, the bias in the estimation of the variance can be as much as -45 %. Likewise, the usual nominal 95 % confidence interval can have a level of confidence as low as 88 %. This issue is discussed and techniques are suggested so that Type A evaluation of uncertainty becomes as accurate as is implied. The 'objective Bayesian' approach to this issue is discussed and an associated unacceptable phenomenon is identified.

Section: RESEARCH PAPER

Keywords: bias; propagation of errors; propagation of uncertainty; stopping rule; unbiased estimation

Citation: Robin Willink, Type A evaluation of measurement uncertainty when the sample size is not predetermined, Acta IMEKO, vol. 4, no. 4, article 8, December 2015, identifier: IMEKO-ACTA-04 (2015)-04-08

Section Editor: Franco Pavese, Torino, Italy

Received April 26, 2015; **In final form** November 7, 2015; **Published** December 2015

Corresponding author: Robin Willink, e-mail: robin.willink@gmail.com

1. INTRODUCTION

The assessment of the variance of a component of measurement error by statistical means is known as Type A evaluation of measurement uncertainty. The principal form of Type A evaluation discussed in the *Guide to the Expression of Uncertainty in Measurement*, (the GUM) [1], [2], relates to the estimation of a quantity by the averaging of n individual results regarded as having been independently drawn from a normal distribution with mean equal to the quantity of interest. The analysis there does not explicitly acknowledge that it assumes the sample size n to be determined without reference to the data, as if it had been fixed before the analysis. Yet many laboratory experiments involve sample sizes that are not known in advance. Typically, an experimental scientist will continue to make observations until he or she is satisfied with the results, so that the final sample size is affected by whether the growing dataset makes sense and appears adequate for the experimenter's purpose. It is therefore important to study whether the accuracy of the GUM's process of Type A evaluation is affected by such practice.

There are several relevant results from the theory of classical statistics. Let X_i be the random variable for the ith value drawn independently from the normal distribution with unknown mean μ and unknown variance σ^2. Define

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

So $\bar{X}$ and S^2 are the random variables for the sample mean and sample variance when an independent sample of fixed size *n* is taken. The results of fundamental importance are italicized in the following statements.

1. *The distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n,* (and these expressions for the mean and variance of $\bar{X}$ are correct even if the distribution is not normal).

2. *The expected value of* S^2 *is* σ^2, (and this is the case even if the distribution is not normal). So S^2 is called an 'unbiased estimator' of σ^2. Consequently, the realization or outcome of S^2, which is a number denoted s^2, is called an unbiased estimate of σ^2.
3. *The sample-mean random variable* $\bar{X}$ *is independent of every function* $g(X_1, \cdots X_n)$ *for which* $g(X_1, \cdots X_n) = g(X_1 + a, \cdots X_n + a)$ *for every constant* a [3].
4. The random variable $(X - \mu)/(S/\sqrt{n})$ has the t-distribution with $n - 1$ degrees of freedom. So, with $t_{\nu,p}$ denoting the *p*-quantile of the t-distribution with ν degrees of freedom, *the random interval*

$$Q \equiv \left[\bar{X} - \frac{t_{n-1,0.975}S}{\sqrt{n}}, \bar{X} + \frac{t_{n-1,0.975}S}{\sqrt{n}}\right]$$

has probability 0.95 of covering μ *whatever the values of* μ *and* σ^2. Therefore, this random interval is a '95 % confidence interval' for μ. The corresponding numerical interval

$$\left[\bar{x} - \frac{t_{n-1,0.975}s}{\sqrt{n}}, \bar{x} + \frac{t_{n-1,0.975}s}{\sqrt{n}}\right]$$

can be called a 'realized 95 % confidence interval for μ' or an 'evaluated 95 % confidence interval for μ'. (Regrettably, there are two slightly different definitions of a confidence interval that can be found in authoritative encyclopedias of statistics. Following the *International Dictionary of Statistics* [4], we take the term confidence interval to refer to the estimator, not the estimate. Therefore the confidence interval itself is Q and the adjective 'realized' or 'evaluated' is added when referring to the numerical interval that is the resulting estimate of μ. In this, we differ with the *Encyclopedia of Statistical Sciences* [5], which takes the confidence interval itself to be the numerical interval.)

Result 1 tells us that the standard uncertainty to be associated with the observed sample mean $\bar{x}$ should be an estimate of $\sigma/\sqrt{n}$, which is the standard deviation of the measurement error. Result 2 is important because many measurement processes involve combining estimates of several 'input quantities', and the law of propagation of uncertainty involves summing the estimates of the variances of the errors. The ensuing estimate of the variance of the total error will be unbiased if the component estimates are unbiased, (notwithstanding the effect of approximating the measurement equation by a locally linear function). Result 3 implies that $\bar{X}$ is independent of any statistic that provides no information about μ. In particular, it implies the well-known result that $\bar{X}$ and S^2 are independent. Result 4 is important because it implies that 95 % of the '95 % confidence intervals' evaluated in a long series of problems will enclose the unknown actual values of the measurands. This gives a practical and unambiguous meaning to the output of an uncertainty analysis.

Now consider a situation where the sample size is not fixed but is somehow chosen by the experimenter in response to the data. The sample size n is now the numerical value taken by a discrete random variable N. The point estimators of μ and σ^2 become

$$\bar{X}_N \equiv \frac{1}{N}\sum_{i=1}^{N} X_i$$

and

$$S_N^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2,$$

and the confidence interval for μ (which is the interval estimator of μ) becomes

$$Q_N \equiv \left[\bar{X}_N - \frac{t_{N-1,0.975}S_N}{\sqrt{N}}, \bar{X}_N + \frac{t_{N-1,0.975}S_N}{\sqrt{N}}\right].$$

With the sample size being random, the estimator S_N^2 does not necessarily behave in the same way as S^2, so we cannot say that the expected value of S_N^2 is equal to σ^2. Similarly, the random intervals Q_N and Q are different entities, so while it is known from theory that Q has probability 0.95 of covering μ, the probability that Q_N will cover μ at different values of μ and σ is not known. To be more specific, we do not know the minimum probability over all possible values of μ and σ, which is the 'confidence coefficient' or 'confidence level' of Q_N. As shall be seen, the expected value of S_N^2 and the confidence coefficient of Q_N depend on the rule that the experimenter uses to decide when to stop making observations, which is called the 'stopping rule'. We will find that, if the stopping rule is particularly unfavourable, σ^2 can be underestimated by approximately 45 % on average and the confidence coefficient of Q_N can be reduced to approximately 88 %.

With the sample size being random, we must also recognise that the variance of the measurement error, which is equal to the variance of $\bar{X}_N$, is no longer a constant. However, our interest lies in estimating not the unconditional variance of the measurement error but the variance conditional on the observed sample size, i.e. conditional on the event $N = n$. We cannot take it for granted that this conditional variance is equal to σ^2/n, so we cannot assume that the conditional standard deviation of the measurement error is given by $\sigma/\sqrt{n}$. More importantly, we cannot even take it for granted that $\bar{X}_N$ is an unbiased estimator of μ and, therefore, we cannot assume that the sample mean is the preferred estimate of the quantity measured. (The estimator would not be unbiased, for example, if our rule was to stop sampling when the sample mean reached a certain level.) Thus there are some basic questions to answer about $\bar{X}_N$ before we examine the behaviours of S_N^2 and Q_N.

1.1. The conditional distribution of the measurement error

We have just hinted at the fact that the distribution of the sample mean $\bar{X}_N$ on the condition that the sample size is n could conceivably differ from the normal distribution with mean μ and variance σ^2/n. If so, using the sample mean as the estimate of the quantity measured and taking our estimate of $\sigma/\sqrt{n}$ as the standard uncertainty might be misguided. However, as now shown, there is no cause for alarm if our stopping rule depends only on the spread of the data and not on the location of the data with respect to any origin.

We would like our stopping rule to be such that the distribution of $\bar{X}_N$ on the condition that $N = n$ is the normal distribution with mean μ and variance σ^2/n. Result 3 above implies that, for a fixed sample size m, the sample mean random variable $\bar{X}_m$ is independent of every function of the X_i variables that gives us no information about location. So if the decision to stop is made according to the value of some such function then the conditional distribution of $\bar{X}_N$ at any point is

unaffected by whether sampling stopped at that point or not; i.e., knowing that our stopping criterion will be first met by a sample of size n would give us no extra information about the mean of those n observations. It follows that the distribution of $\bar{X}_N$ on the condition that $N = n$ is equal to the normal distribution with mean μ and variance σ^2/n, as required.

The only stopping rules we discuss in the rest of this paper meet this condition, and so our principal task remains the unbiased estimation of σ^2, from which we obtain a (conditional) unbiased estimate of σ^2/n simply by dividing by n. We would then seek an estimate of this to use as the square of the standard uncertainty, $u^2(\bar{x})$, for potential use in the 'law of propagation of error'. Consequently, our performance measure of relative bias in S^2/n retains its relevance.

1.2. Sequential statistics and stopping rules

The idea that the sample size depends in some way on the data is the key idea of *sequential statistics*. Perhaps the best example of a situation involving sequential statistics is where the efficacies of two treatments are to be compared in patients recruited to a medical trial. It is unethical to subject people to a treatment known to be sub-optimal, so the trial should be stopped as soon as the identity of the preferred treatment becomes clear. Yet the overall procedure must maintain its statistical rigour for the comparison to be scientifically acceptable. This requires the development of a formal stopping rule that preserves the statistical integrity of the conclusion. In our context of Type A evaluation, such integrity relates to the conditional distribution of the estimator of μ, (which we have just discussed), to bias in the estimator of σ^2 and to the probability that the procedure generates an interval enclosing μ, which is an event that can be called 'success'. Whatever stopping rule is put into place, our estimate of σ^2 should not be negatively biased and the minimum probability of success should be at least 95 %.

In addition to the idea of having a fixed sample size, there are at least two broad possibilities for the way in which an experimenter might decide to cease making observations:

1. stop when the result is deemed to be sufficiently accurate, i.e. when the measurement uncertainty is sufficiently low;
2. stop when the data suggests that the experimental procedure is behaving as expected.

As written, these two ideas are somewhat vague. We need to be more specific when describing a stopping rule because if there is no formal criterion by which an experimenter decides to finish sampling then there is no well-defined procedure whose statistical properties can be determined or claimed, either using theory or simulation. So let us now describe the basic situation in mathematical terms. Observations of a single quantity are made one by one, the results being $x_1, x_2, \cdots, x_n$, and after obtaining x_k the figures

$$\bar{x}_k \equiv \frac{1}{k}\sum_{i=1}^{k} x_i$$

and

$$s_k^2 \equiv \frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{x}_k)^2$$

are calculated. The process is stopped when the growing dataset of x_k, $\bar{x}_k$ and s_k^2 values satisfies some well-defined condition, one possibility being a condition that reflects a minimum required level of measurement accuracy and another possibility being a condition that relates to the general appearance of the dataset. The sample size n, which is the outcome of a random variable N, is the final value of k.

2. VARIOUS STOPPING RULES AND THEIR EFFECTS

Let us now examine the effects of different procedures on the aspects of performance described, which are (i) the extent of any bias in the estimation of σ^2 and (ii) the extent of any reduction in the probability that the confidence interval contains μ. We focus on the idea of stopping when the measurement is deemed sufficiently accurate for the experimenter's purposes, which in essence is when the figure $s_k/\sqrt{k}$ is sufficiently small. Other criteria for stopping are discussed only briefly.

2.1. Stopping at a predetermined level of accuracy

Consider the notion of making observations only until a predetermined level of accuracy is reached. When the quantity that is measured is just one input to a broader measurement, (which is the primary context of the GUM), this implies stopping when the corresponding figure of standard uncertainty, $s_k/\sqrt{k}$, first drops beneath a predetermined maximum value g. In contrast, when the measurement stands alone, it implies stopping as soon as $t_{k-1,0.975}s_k/\sqrt{k} \leq h$, where h is a predetermined maximum half-width for the 95 % uncertainty interval. These are different stopping criteria, and we will indicate them by G and H respectively.

The performances of these procedures will also depend on any initial sample size. With σ being unknown, the smallest possible sample size is two. However, we can also envisage a situation in which the experimenter is never prepared to use a sample of less than three, or maybe a larger number still. The rule of starting with a sample of size two and stopping as soon as $s_k/\sqrt{k} \leq g$ can be called $\mathrm{G}(2, g)$, while the rule of starting with a sample of size three and stopping as soon as $t_{k-1,0.975}s_k/\sqrt{k} \leq h$ can be called $\mathrm{H}(3, h)$. The general forms of these rules are:

$\mathrm{G}(n_1, g)$: stop if $k \geq n_1$ and $s_k/\sqrt{k} \leq g$

$\mathrm{H}(n_1, h)$: stop if $k \geq n_1$ and $t_{k-1,0.975}s_k/\sqrt{k} \leq h$.

In this paper, n_1 always indicates a predetermined minimum sample size.

The solid lines in Figure 1 show the relative bias of S_N^2 for various stopping rules of the G kind. The bias is a function of σ/g alone: the results were obtained by simulation where, without loss of generality, g was set to unity. For each rule, 10^5 replications were conducted at each value of σ/g indicated by the markers. Figure 1 shows that if we use $\mathrm{G}(2, g)$ and the true unknown value of σ is approximately $2.5g$ then there is a negative bias of 45 %, i.e. the expected value of S_N^2 is as low as $0.55\,\sigma^2$. If there were many uncertainty components and each was affected in this way then the result would be the quotation of an uncertainty interval that was 26 % shorter than the appropriate interval.

The solid lines in Figure 2 show the probability that Q_N encloses μ for stopping rules of the H kind. The probability is a function of σ/h alone. For each rule, 10^5 replications were conducted at each value of σ/h indicated by the markers.

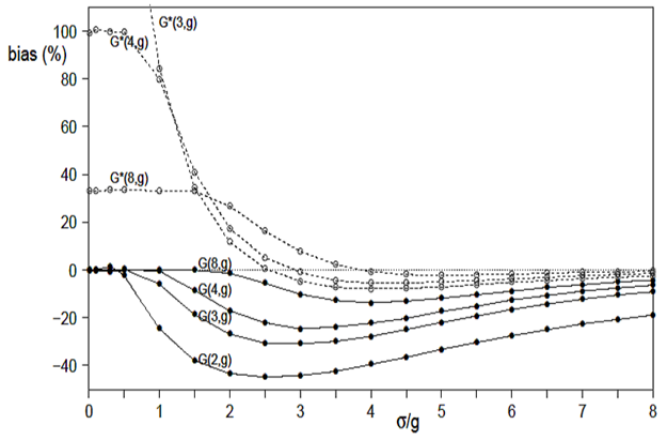

Figure 1. Bias of S_N^2 (solid lines) and $NS_N^2/(N-2)$ (dashed lines) in the estimation of σ^2 using stopping rules G(n_1, g) and G$^*(n_1, g)$ respectively.

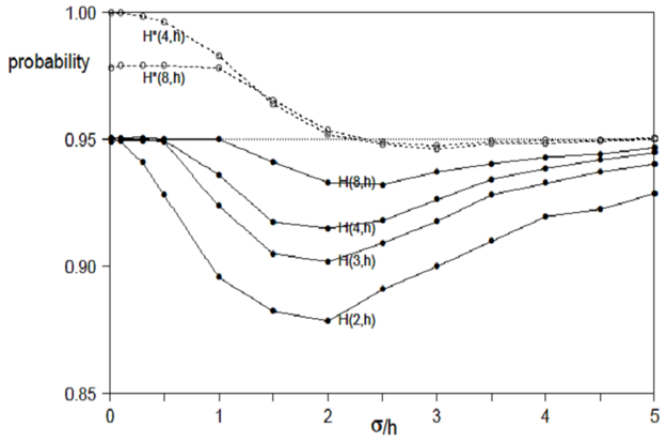

Figure 2. Probability that intervals Q_N (solid lines) and Q_N^* (dashed lines) enclose μ using stopping rules H(n_1, h) and H$^*(n_1, h)$ respectively.

Figure 2 exhibits the same general pattern as Figure 1, but the scale differs in the *x*-direction because h needs to be larger than g for comparable results. The success rate of Q_N with H$(2, h)$ can be as low as 88 %, and this occurs when $\sigma \approx 2h$.

These results should be of some concern to any scientist whose practice is to measure and average the results until the uncertainty is 'small enough'. That practice entails stopping at a point where the sample estimate of variance tends to be smaller than average at that sample size, and the validity of the statistical inference becomes affected. Performance levels can be improved, but are not able to be fully corrected, by using larger values of the initial sample size, n_1.

Restoring performance levels

Performance can be restored somewhat by involving an 'effective sample size' smaller than the actual sample size. Accordingly, the dashed lines in Figures 1 and 2 show the results obtained when using the figure $NS_N^2/(N-2)$ as the estimator of σ^2 instead of S_N^2. So the effective sample size for the estimation of the variance is deemed to be $N-2$ not N. The two corresponding rules are:

G$^*(n_1, g)$: stop if $k \geq n_1$ and $s_k/\sqrt{k-2} \leq g$

H$^*(n_1, h)$: stop if $k \geq n_1$ and $t_{k-3,0.975}s_k/\sqrt{k-2} \leq h$.

After applying either of the rules, the sample size n is equal to k, the standard uncertainty to be quoted with $\bar{x}_n$ as an estimate of the quantity measured is $u(\bar{x}_n) = s_n/\sqrt{n-2}$ and the realized 95 % confidence interval to be quoted for μ is

$$\left[\bar{x}_n - \frac{t_{n-3,0.975}s_n}{\sqrt{n-2}}, \bar{x}_n + \frac{t_{n-3,0.975}s_n}{\sqrt{n-2}}\right].$$

This is the outcome of the confidence interval

$$Q_N^* \equiv \left[\bar{X}_N - \frac{t_{N-3,0.975}S_N}{\sqrt{N-2}}, \bar{X}_N + \frac{t_{N-3,0.975}S_N}{\sqrt{N-2}}\right].$$

(In accordance with usual relationship between the sample size and the number of degrees of freedom, the method uses the multiplier $t_{n-3,0.975}$ instead of $t_{n-1,0.975}$.)

The dashed lines in Figures 1 and 2 show the performances of G$^*(n_1, g)$ and H$^*(n_1, h)$. The minimum possible sample size is three with a G* rule and four with an H* rule. Using G$^*(3, g)$ in place of G$(3, g)$ reduces the worst negative bias from -31 % to -8 % and using H$^*(4, h)$ instead of H$(4, h)$ increases the confidence coefficient from 91½ % to 94½ %. The conservatism at low values of σ/g and σ/h is of little consequence; at these points in the parameter space, the absolute bias and the width of the confidence interval will be small. (The relative bias at $\sigma/g \to 0$ approaches $100\{2/(n-2)\}$ %, so the dashed line for G$^*(3, g)$ on Figure 1 approaches the y-axis at the value 200 %.)

The rule G$^*(4, g)$ has a worst negative bias of 5½ % and attains a confidence coefficient of 94½ %, (which is not shown graphically). These levels of performance might be regarded as acceptable, so it is suggested that the rule G$^*(4, g)$ be used when measuring for a prescribed level of accuracy in this way. Thus, the minimum sample size suggested is four.

Conditional confidence and conditional bias

The confidence coefficient is the minimum probability that the procedure generates an interval containing μ. We have seen that the nominal 95 % estimation procedure involving H$(4, h)$ has confidence coefficient 88 %, which seems too low for the label '95 % uncertainty interval' to be acceptable. But perhaps we should be more interested in the minimum probability conditional on the observed value of the random variable N, i.e. conditional on the event $N = n$. In other words, perhaps we should prefer to generate 95 % 'conditional confidence intervals', where the relevant probability statement holds on the condition that N takes its observed value. That is, if a random interval $I(\bar{X}_N, S_N^2, N)$ satisfies

$$\Pr\{I(\bar{X}_N, S_N^2, N) \ni \mu | N = n\} = 0.95$$

then $I(\bar{X}_N, S_N^2, N)$ on the condition that $N = n$ is a '95 % conditional confidence interval for μ'. In this way, we would be making more use of the available information, and potentially generating a narrower uncertainty interval.

The availability of a conditional-confidence procedure seems to depend on the availability of an ancillary statistic [6], which is a statistic whose distribution does not depend on the value of the parameter being studied. With stopping rule G(n_1, g), for example, the variable N is not ancillary when estimating σ^2 because its distribution depends on σ/g. Further analysis of these ideas of ancillarity and conditional inference is outside the scope of this paper.

2.2. A two-stage procedure for a predetermined standard uncertainty

We now give two methods where the total sample of size n is formed from an initial sample of predetermined size n_1 and a

second sample of a size n_2 chosen using the data in the initial sample. So $n = n_1 + n_2$. The first method is found in the following paragraphs and the second method is given in Section 2.3. Both methods rely on the fact that the sample variance $s_{n_1}^2$ obtained after a fixed number of observations n_1 is an unbiased estimate of σ^2. Therefore, on the condition that the total sample size is n, the figure $s_{n_1}^2/n$ is an unbiased estimate of the figure sought, σ^2/n.

Suppose that the requirement of Section 1.1 is met. Also, as in Section 2.1, suppose that there is a predetermined figure g that would be a maximum acceptable value for the standard uncertainty of the measurement. Set $n^* = s_{n_1}^2/g^2$. Then, notwithstanding the fact that n^* will not be a whole number, we would set $n = n^*$ and the figure of standard uncertainty would be g. In practice we would set n equal to the value of n^* rounded up. Thus, an acceptable procedure is:

1. Make n_1 observations, calculate $s_{n_1}^2$ and $n_2 = \left\lceil s_{n_1}^2/g^2 \right\rceil - n_1$;
2. If $n_2 > 0$ then make n_2 more observations, calculate the overall mean $\bar{x}$ and quote $u(\bar{x}) = g$. If $n_2 < 0$ then calculate $\bar{x}$ from the first sample and set $u(\bar{x}) = s_{n_1}/\sqrt{n_1}$.

This procedure is analogous to Stein's fixed-width confidence interval [7], which has been an influential technique in sequential statistics. Stein showed that a procedure for realizing a 95 % confidence interval of a fixed width $2h$ for the mean μ of a normal distribution is as follows:

1. Make n_1 observations, calculate $s_{n_1}^2$ and $n_2 = \left\lceil \left(t_{n_1-1,0.975}\, s_{n_1}/h\right)^2 \right\rceil - n_1$;
2. If $n_2 > 0$ then make n_2 more observations and calculate the overall mean $\bar{x}$. If $n_2 \leq 0$ then calculate $\bar{x}$ from the first sample;
3. Quote the interval $[\bar{x} - h, \bar{x} + h]$ as a realized 95 % confidence interval for μ.

The drawback of these procedures is that neither makes use of the information about the variance σ^2 contained in the second sample.

2.3. A two-stage procedure with the standard uncertainty chosen after the first stage

So far we have considered the situation where the minimum required level of precision represented by g or h is known beforehand. That is somewhat unrealistic, for the experimenter is potentially committing himself or herself to making a huge number of observations depending on the ratio of σ/g or σ/h with σ unknown beforehand. So now we envisage that an acceptable value for the standard uncertainty is identified after taking the first sample and that the corresponding size of the second sample is then calculated.

The sample variance $s_{n_1}^2$ obtained after a fixed number of observations n_1 is an unbiased estimate of σ^2 based on $n_1 - 1$ degrees of freedom. Thus, on the condition that the overall sample has size $n_1 + n_2$ and the requirement of Section 1.1 is met, $s_{n_1}^2/(n_1 + n_2)$ is an unbiased estimate of the conditional variance of the overall sample mean, which is $\sigma^2/(n_1 + n_2)$. Therefore, a valid procedure is:

1. Make n_1 observations and calculate $s_{n_1}^2$;
2. Without paying attention to the mean of the first sample, choose n_2 so that $s_{n_1}/\sqrt{n_1 + n_2}$ is an acceptable figure of standard uncertainty;
3. Make n_2 more observations, calculate the overall mean $\bar{x}$ and define $u(\bar{x}) = s_{n_1}/\sqrt{n_1 + n_2}$.

This procedure can be improved upon if the experimenter accepts the possibility of the corresponding standard uncertainty being a little larger than the figure chosen. No matter how n_2 is chosen, the quantity

$$s_{two}^2 = \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n_1+n_2} \left(x_i - \sum_{i=n_1+1}^{n_1+n_2} \frac{x_i}{n_2} \right)^2 ,$$

which is the sample variance of the second sample, is an unbiased estimate of σ^2 based on $n_2 - 1$ degrees of freedom. Therefore the pooled estimate of variance,

$$s_{pool}^2 = \frac{(n_1 - 1)s_{n_1}^2 + (n_2 - 1)s_{two}^2}{n_1 + n_2 - 2}$$

is an unbiased estimate of σ^2 based on $n_1 + n_2 - 2$ degrees of freedom, and so has a smaller parent variance than $s_{n_1}^2$. Thus, provided that the choice is made beforehand, the third step above can be replaced by:

3. Make n_2 more observations, calculate the overall mean $\bar{x}$ and define $u(\bar{x}) = s_{pool}/\sqrt{n_1 + n_2}$.

Of the methods described so far, the method of this section is, perhaps, the method of choice for practical experimentation. It provides the experimenter with some control over both the final standard uncertainty and the final sample size.

2.4. Other stopping rules

In the course of this work a number of other stopping rules were examined, almost all of which appeared to lead to statistical exactness, as would be hoped for. The basic requirement for exactness, when sampling from a normal distribution, is that the decision to stop is not influenced by the observed values of the sample mean and sample variance. A method for preventing this, albeit one that most experimentalists would find unacceptable, would be as follows. Let a and b be constants whose values are chosen randomly and are not revealed until sampling has stopped. (We require $b \neq 0$.) Instead of observing $x_1, x_2, \cdots$ the experimenter observes $a + bx_1, a + bx_2, \cdots$. With a and b being completely unknown, the observations give no information about the sample mean and sample variance (of the recoverable x_i values) and hence no information about μ and σ. Therefore, no stopping rule can affect the validity of the statistical inference, and so the performance levels of the Type A evaluation procedure will not be affected.

Perhaps the value of that somewhat impractical idea is greatest when it is used in a thought process. If the experimenter can honestly say that they would have stopped taking observations at the same point if they had been observing $a + bx_i$ instead of x_i for all choices of a and b then the rule by which sampling was ceased does not introduce error.

Last, we consider the idea of measuring until forced to stop simply because of time constraints. In this case, the sample size is not influenced by any of the data, and so this stopping rule does not introduce error.

3. BAYESIAN ANALYSIS

Currently, there is some interest in the use of Bayesian statistics for the evaluation of measurement uncertainty. Indeed, the revision of the GUM is being carried out in line with some of the principles of Bayesian statistics [8], these principles being associated with 'objective-Bayesian' methods. So we now discuss a Bayesian understanding of the issue described in this paper.

The output of a Bayesian statistical analysis depends on the prior distributions for the unknown parameters and the likelihood function for the data observed. It does not depend on the stopping if the stopping rule is 'noninformative', which means that the value taken by N provides no information about the parameter being estimated beyond what is contained in the prior distributions and the likelihood function at that point [9], [10], [11]. This is a weak condition, and all the stopping rules discussed in this paper would be noninformative in a Bayesian analysis. Thus the Bayesian paradigm does not distinguish between the reliability of the output obtained with a fixed sample size and the output generated by one of the stopping rules we have examined. This presents the objective-Bayesian statistician with a number of related problems when considering Type A evaluation of uncertainty, as we now demonstrate. These problems can be added to a growing list of difficulties associated with the use of Bayesian ideas in the evaluation of measurement uncertainty [12]-[17].

3.1. An inconsistency and an invalid procedure

Our measurement problem involves random sampling from a normal distribution with mean μ and unknown variance σ^2. It is well known that when the standard prior distributions are used in an objective-Bayesian analysis of this problem, the posterior distribution of μ after observing a sample of size n is the distribution of $\bar{x}_n + Ts_n/\sqrt{n}$ where T is a random variable with the *t*-distribution with $n-1$ degrees of freedom. (This is also the distribution advocated to describe μ in Supplement 1 to the GUM [18].) The conclusion would be drawn that 'there is 95 % probability that μ lies in the interval $\left[\bar{x}_n - t_{n-1,0.975}s_n/\sqrt{n}, \bar{x}_n + t_{n-1,0.975}s_n/\sqrt{n}\right]$', which is seen to be the same interval realized by Q. This would be the case in an objective-Bayesian analysis under any of the stopping rules we have discussed.

Therefore, when the rule is to take a fixed number of observations, the Bayesian analysis generates the same interval as the classical methodology, which is known to be successful on 95 % of occasions. Yet when using rule $\mathrm{H}(2,h)$, for example, the analysis is successful on less than 95 % of occasions, especially when the reference set of problems involves small values of σ/h; see Figure 2. So the results of two procedures with different success rates are, in effect, always regarded by the objective-Bayesian statistician as being equally reliable!

3.2. An unacceptable principle

The idea that Bayesian inference does not depend on the typical stopping rule opens up the hypothetical possibility of abuse. For simplicity, suppose that σ is known exactly. After making k observations $x_1, \cdots, x_k$ the Bayesian statistician using the flat prior distribution for μ (and the person adhering to Supplement 1 to the GUM [18]) will assign to μ the normal distribution with mean $\bar{x}_k$ and standard deviation $\frac{\sigma}{\sqrt{k}}$. Whatever the value of $\bar{x}_k$ and however the observation process was terminated, the person will feel able to state '$Pr\left(\mu \in \left[\bar{x_k} - \frac{1.96\sigma}{\sqrt{k}}, \bar{x_k} + \frac{1.96\sigma}{\sqrt{k}}\right]\right) = 0.95$'. If a probability of 0.05 is used as a threshold of plausibility, the person will consider all values outside that interval to be implausible as values for μ. Suppose the person is prepared to make up to 10 observations and, for some reason, wishes to reach the conclusion that a certain value z is *not* a plausible value for μ. Simulations show that they will be able to do so with probability at least 0.195 even if z is equal to the true value of μ. That is, if I_k is the random interval

$$\left[\sum_{i=1}^{k} X_i/k - 1.96\,\sigma/\sqrt{k}\,, \sum_{i=1}^{k} X_i/k + 1.96\,\sigma/\sqrt{k}\right]$$

then

$$\Pr(I_1 \not\ni z \text{ or } \cdots \text{ or } I_{10} \not\ni z\,) \geq 0.195 \text{ for all } z$$

and, in particular, for $z = \mu$. If the person is prepared to make up to 30 observations then the figure becomes 0.29.

Worse behaviour is admitted by moving to large, albeit astronomical, sample sizes, as an informal argument now shows. After k observations the objective Bayesian statistician will calculate a symmetric 95 % credible interval for μ to be $\left[\bar{x}_k - \frac{1.96\sigma}{\sqrt{k}}, \bar{x}_k + \frac{1.96\sigma}{\sqrt{k}}\right]$. The (frequentist) probability that the interval calculated after k_1 observations will exclude μ is 0.05. Now let k_2 be a number so much greater than k_1 that the effect of the first k_1 results on the mean of the first k_2 results is negligible. Then, in effect, the probability that the interval calculated after k_2 observations will exclude μ is 0.05 independently of whether the interval after k_1 observations excludes μ. Thus, the probability that one or both of these two intervals will exclude μ is $1 - 0.95^2 \approx 0.1$. Now let k_3 be a number so much greater than k_2 that the effect of the first k_2 results on the mean of the first k_3 results is negligible, and so on. If there are q stages then the probability that at least one of the q intervals will exclude μ is $1 - 0.95^q$, which can be made arbitrarily close to 1 by making q sufficiently large. Moreover, this argument holds even if we carry out the analysis with 99 % intervals, or with any other threshold for implausibility. Thus, by measuring enough times, (albeit an astronomical number of times), the objective-Bayesian statistician can always conclude that the true value of μ is implausible as a value for μ. Evidently, this is also the case for any other value that they wish to rule out.

3.3. Poverty of objective Bayesian statistics

The set $\{x_1, x_1 + x_2, \cdots\}$ is a sample path of a stochastic process $\{X_1, X_1 + X_2, \cdots\}$. Under the assumption of independent sampling from a shared distribution with mean μ and standard deviation σ, the related stochastic process $\{A_k\} \equiv \left\{\sum_{i=1}^{k} (X_i - \mu)/\sigma\right\}$ obeys the *law of the iterated logarithm*, which states that

$$\lim_{k\to\infty} \sup \frac{A_k}{\sqrt{k \log\log k}} = \sqrt{2} \quad \text{with probability one.}$$

Therefore, $\lim_{k\to\infty} \sup A_k/\sqrt{k} = \infty$. This implies that

$$\lim_{k\to\infty} \sup \frac{\bar{X}_k - \mu}{\sigma/\sqrt{k}} = \infty$$

where $\bar{X}_k \equiv \sum_{i=1}^{k} X_i/k$ is the stochastic process of the sample mean. And, by symmetry, $\lim_{k\to\infty} \inf \frac{\bar{X}_k - \mu}{\sigma/\sqrt{k}} = -\infty$. Thus, by

sampling long enough, we can make the quantity $\frac{\bar{x}_k-\mu}{\sigma/\sqrt{k}}$ pass through any value we choose, and in particular we can reach values k_1 and k_2 such that the intervals $\left[\bar{x}_{k_1}-\frac{a\sigma}{\sqrt{k_1}},\bar{x}_{k_1}+\frac{a\sigma}{\sqrt{k_1}}\right]$ and $\left[\bar{x}_{k_2}-\frac{a\sigma}{\sqrt{k_2}},\bar{x}_{k_2}+\frac{a\sigma}{\sqrt{k_2}}\right]$ are disjoint for any finite predetermined value of the coverage factor a, say 1.96. If the distribution is normal then process $\{A_k\}$ is a Brownian motion, which has a self-similarity property. This means we can repeat the process *ad infinitum* to obtain new pairs (k_1,k_2) with this property. Importantly, we can choose k_1 arbitrarily and then wait until a suitable k_2 arises.

This phenomenon permits a whimsical but correct analysis of how profit can be guaranteed when betting against an objective-Bayesian statistician who trusts the figures of probability they calculate and behaves accordingly. A Bayesian acting as a 'rational agent' will always accept a bet in which they view the expected value of their profit as positive and will reject a bet in which they view the expected value of their profit as negative. In the current problem of estimating the mean of a normal distribution, we take a sample of arbitrary size (k_1) and offer the person 1 cent against 10 cents that μ lies outside the 95 % credible interval they calculate for it. They will necessarily accept, believing their expected profit to be $0.95\times 1-0.05\times 10=0.45$ cents. We then continue sampling until the corresponding 95 % credible interval is disjoint from the original one, which - as has been shown - must happen at some point (k_2), and we present the same offer with the new interval. Again, the person must accept if they are to be acting in a way consistent with their philosophy. Then we point out that they must be incorrect in at least one of the wagers, and we generously concede the other. Our profit is 9 cents. This process can be repeated until matters of morality and compassion dictate.

There is nothing in the Bayesian theory to invalidate this behaviour by the objective-Bayesian statistician. It is an example of the incoherence admitted by using an 'improper' prior distribution in a Bayesian analysis. The same result applies in the case of normal sampling with σ unknown because the sample standard deviation is a consistent estimator of σ.

4. TYPE A EVALUATION OF THE VARIANCE IN THE TECHNIQUE

This paper has addressed Type A evaluation of measurement uncertainty when all the data originates in the measurement at hand. Accordingly, σ^2 is deemed unknown *a priori* and is estimated only from the data obtained in the measurement. In this situation we are primarily concerned with obtaining an unbiased estimate of σ^2/n where n is the final sample size. The standard uncertainty to be quoted is then $\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}^2$ is some figure that in effect acts as an estimate of σ^2. Usually $\hat{\sigma}^2=s^2$.

However, there are many situations where σ^2 has been estimated beforehand in order to 'calibrate' or 'characterize' a measurement process and to establish 'the standard uncertainty of the measurement technique, $\hat{\sigma}$'. This is to be contrasted with 'the standard uncertainty in the particular measurement result $\hat{\sigma}/\sqrt{n}$'. For example, the GUM describes a situation in which the variability of one aspect of the procedure was estimated beforehand from 25 observations, so that the corresponding estimate of variance was based on 24 degrees of freedom [1], [2]. In the earlier measurement, the objective will have been to obtain an unbiased and sufficiently accurate estimate of σ^2, (which is a quantity that does not vary with n), rather than an unbiased and sufficiently small estimate of σ^2/n. If this estimate of σ^2 is subsequently used in the measurement at hand then the experimenter knows in advance how many observations must be taken to obtain a sufficiently accurate result. So the analysis given in Section 2 is not necessary and statistical exactness is maintained provided that the behaviour of the experimenter is not affected by the changing value of the sample mean.

5. CONCLUSION

The thought processes by which experimentalists decide that enough observations have been taken will be many and varied - and they might not be describable in rigorous form. However, one idea is that observations are made until the corresponding component of standard uncertainty associated with repeatability or reproducibility is, apparently, smaller than some predetermined level. Figures 1 and 2 show that if such a 'stopping rule' is applied in Gaussian sampling and if this figure of standard uncertainty is obtained from the observations at hand then the effect is to invalidate the statement of uncertainty. Without appropriate modification, the procedure gives, on average, too small an estimate of the measurement variance, the worst possible bias being -45 %. Similarly, the associated nominal 95 % confidence interval can have a confidence coefficient as low as 88 %. A simple modification described in Section 2.1 involving the use of an effective sample size two less than the actual sample size restores the performance to an acceptable level. An appropriate related stopping rule is $G^*(4,g)$.

Perhaps it is more realistic to envisage that the final number of observations made is chosen after observing the data from an initial set. Accordingly, Section 2.2 and Section 2.3 give two-stage procedures that do not compromise the validity of the Type A evaluation. The method of Section 2.3 seems preferable because it gives some control over both the sample size and the standard uncertainty of measurement.

One finding of this paper is that making observations until the corresponding standard uncertainty 'looks small enough' does not lead to a valid statistical procedure and hence does not lead to valid Type A evaluation of measurement uncertainty. An advocate of Bayesian statistical methods would disagree with that statement, but unacceptable results appear when the analysis is carried out using the standard objective-Bayesian prior distributions. This further calls into question the wisdom of adopting Bayesian ideas in the supplements to the GUM and in the revision of the GUM.

The material presented in this paper relates to Type A evaluation of uncertainty where the variability of the measurement technique is assessed during the measurement at hand. It does not relate to the situation where the estimate of the variance was obtained beforehand, as when the technique is calibrated.

REFERENCES

[1] Guide to the Expression of Uncertainty in Measurement, clause 4.2, (Geneva: International Organization for Standardization) 1995.

[2] JCGM 100:2008 Evaluation of measurement data - Guide to the expression of uncertainty in measurement, clause 4.2, http://www.bipm.org/en/publications/guides/
[3] J.K. Patel, C.B. Read, Handbook of the Normal Distribution, Section 5.2.3, Marcel Dekker, New York, 1982.
[4] J. Pfanzagl, Estimation: confidence intervals and regions, International Encyclopedia of Statistics (eds.) W.H. Kruskal, J.M. Tanur, The Free Press, Macmillan, 1978, pp. 259-267.
[5] G.K. Robinson, Confidence intervals and regions, Encyclopedia of Statistical Sciences, vol. 2, (eds.) S. Kotz, N.L. Johnson, C.B. Read, Wiley, 1982, pp. 120--127.
[6] J. Kiefer, Conditional inference, Encyclopedia of Statistical Sciences, vol. 2, (eds.) S. Kotz, N.L. Johnson, C.B. Read, Wiley, 1982, pp. 103-109.
[7] C. Stein, A two sample test for a linear hypothesis whose power is independent of the variance, Ann. Math. Statist. 16 (1945) 243-258.
[8] W. Bich, Revision of the 'Guide to the Expression of Uncertainty in Measurement'. Why and how. Metrologia 51 (2014) S155-S158.
[9] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Section 7.7.5, 2nd ed., Springer, New York, 1985.
[10] A. O'Hagan, Kendall's Advanced Theory of Statistics, volume 2B, Section 5.19, Bayesian Inference, Edward Arnold, 1994.
[11] P.H. Garthwaite, I.T. Jolliffe and B. Jones, Statistical Inference, Section 7.6, 2nd ed., Oxford University Press, 2002.
[12] B.D. Hall, Evaluating methods of calculating measurement uncertainty, Metrologia 45 (2008) L5-L8.
[13] R. Willink, Difficulties arising from the representation of the measurand by a probability distribution, Measurement Science and Technology 21 (2010) 015110.
[14] R. Willink, On the validity of methods of uncertainty evaluation, Metrologia 47 (2010), pp. 80-89.
[15] R. Willink, Measurement of small quantities: further observations on Bayesian methodology, Accreditation and Quality Assurance 15 (2010) 521-527.
[16] R. Willink, Measurement Uncertainty and Probability, Cambridge University Press, 2013.
[17] R. Willink, What can we learn from the GUM of 1995?, Measurement (in press).
[18] JCGM 101:2008 Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method, clauses 6.4.9 and 6.4.7, http://www.bipm.org/en/publications/guides/