Anomalous Behaviour of the Welch-Satterthwaite Approximation

Anthony O'Hagan
University of Sheffield, UK

Maurice Cox
National Physical Laboratory, UK

Liam Wright
National Physical Laboratory, UK

December 15, 2021

Abstract

The Welch-Satterthwaite approximation is a simple and well-known approach to the Behrens-Fisher problem of inference about a sum of normal means with unequal variances. It consists of approximating the distribution of the corresponding sum of sample means divided by its estimated standard error by a t distribution, with 'effective degrees of freedom' given by the Welch-Satterthwaite formula. This approximation may then be used to construct an approximate confidence interval. However, a paper published by M. Ballico in 2000 shows that the interval can become narrower when one of the variances increases.

Ballico was working and publishing in the field of metrology, where the Welch-Satterthwaite approximation is widely used to construct confidence intervals, but this anomalous behaviour seems to be unremarked in the mainstream statistics literature. We prove that the anomaly can arise whenever one sample size is less than seven. The result has serious implications for metrology, where small sample sizes are common, and we believe it deserves to be more widely known wherever the Welch-Satterthwaite formula is used.

1 Introduction

Given $m > 1$ samples of data x_{ij}, $i = 1, \ldots, m$, $j = 1. \ldots, n_i$, assumed to be independent and normally distributed as $x_{ij} \sim \mathrm{N}(\mu_i, \sigma_i^2)$, with unknown μ_i and σ_i, consider frequentist inference about a linear combination of means $\theta = \sum_{i=1}^{m} a_i \mu_i$. By scaling the variables, we can set $a_i = 1$ without loss of generality and define

$$\theta = \sum_{i=1}^{m} \mu_i \ .$$

The standard frequentist estimator $t = \sum_{i=1}^{m} \bar{x}_i$ is unbiased with variance estimated by

$$u^2 = \sum_{i=1}^{m} u_i^2 \ , \tag{1}$$

where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, $u_i^2 = s_i^2/n_i$ and $s_i^2 = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/(n_i - 1)$.

However, construction of a confidence interval for θ is more problematic and several approaches have been proposed [20]. A simple proposal [17, 18, 19], known as the Welch-Satterthwaite approximation (W-S), is to treat $(t - \theta)/u$ as having a Student's t distribution with 'effective degrees of freedom'

$$d = u^4 \left(\sum_{i=1}^{m} \frac{u_i^4}{d_i} \right)^{-1} \ , \tag{2}$$

where $d_i = n_i - 1$ is the degrees of freedom of the i-th sample. Then an approximate frequentist 95 % confidence interval for θ is $t \pm k(d)u$, where $k(d)$ is the upper 97.5 % point of the Student t distribution with d degrees of freedom.

The validity of W-S has been called into question by Ballico [2], who found that in some circumstances increasing one of the sample variances s_i^2 led to a reduction in the width of the confidence interval for θ. We refer to this behaviour as anomalous because increasing the uncertainty in the data logically should not reduce uncertainty in θ. Hall and Willink [12] acknowledge the anomaly but claim that even when it arises W-S remains a good approximation in the sense that the coverage of the nominally 95 % confidence interval is close to 95 %.

We present the examples from these authors in Section 2 and provide what we believe to be the first theoretical study of the W-S anomaly by proving explicitly in the case $m = 2$ that it arises whenever one sample size is less than 7 and the variance parameter for the other sample is sufficiently small. We further show that the anomaly arises for any m whenever any sample size is less than 7. We also consider the defence of W-S by Hall and Willink and discuss why, while it may be true that the coverage of intervals obtained using W-S is close to nominal, the existence of the anomaly remains a serious concern.

Both these authors were writing in the context of metrology, where W-S is widely used, and where the half-width of the W-S 95 % confidence interval

$$U = k(d)u$$

is referred to as the *expanded uncertainty*. Also, following terminology in metrology, we will refer to the individual mean parameters μ_i as *inputs* and to θ as the *measurand*. In Section 3 we introduce metrology and the importance of W-S in that field. We consider the practical significance of the anomaly for metrology, and point out that in many applications, including the Ballico and Hall and Willink examples, the assumptions of W-S do not apply.

Section 4 concludes that the anomaly is an intrinsic problem for W-S and in our opinion renders it unfit for purpose.

2 The Welch-Satterthwaite anomaly

2.1 The Ballico examples

To introduce the W-S anomaly, Ballico presents the example which first drew it to his attention. His example has $m = 5$ inputs and consists of two cases, the second of which differs from the first only in that the uncertainty u_i for two inputs is much larger than in the first case, seven times larger for the third input and ten times larger for the fifth. His data are shown in Table 1.

Table 1: Ballico's 5-input example

Input i	d_i	u_i Case 1	u_i Case 2
1	3	12	12
2	8	2	2
3	20	1	7
4	50	0.5	0.5
5	50	0.3	3

Applying the W-S approximation for Case 1 we have $u = 12.22$, $d = 3.23$, $k(d) = 3.06$ and hence $U = 37.40$. In Case 2 we have $u = 14.36$, $d = 6.05$, $k(d) = 2.44$ and $U = 35.08$. Thus, increasing uncertainty in inputs 3 and 5 has nevertheless led to a smaller expanded uncertainty for the measurand. The larger u_i values necessarily lead to an increase in u, but the root of the anomaly lies in the W-S formula, which gives a higher effective degrees of freedom d and hence a smaller $k(d)$. The reduction in $k(d)$ is sufficient to offset the increase in u.

We are not aware of this anomalous behaviour having been reported in the mainstream statistics literature.

Ballico pointed to the fact that input 1 had the greatest uncertainty but had only 3 degrees of freedom and asserted that it is in circumstances like these that the anomaly arises. He went on to present a second example with $m = 2$ input quantities, to highlight the anomaly in a simpler context. In this

example, $u_1 = 1$, $d_1 = 3$ or 4 and d_2 is large, effectively infinite. He plotted the value of U as a function of u_2. The plot for $d_1 = 3$ is shown in Figure 1. We see that U decreases as u_2 increases for all u_2 less than about 0.6.

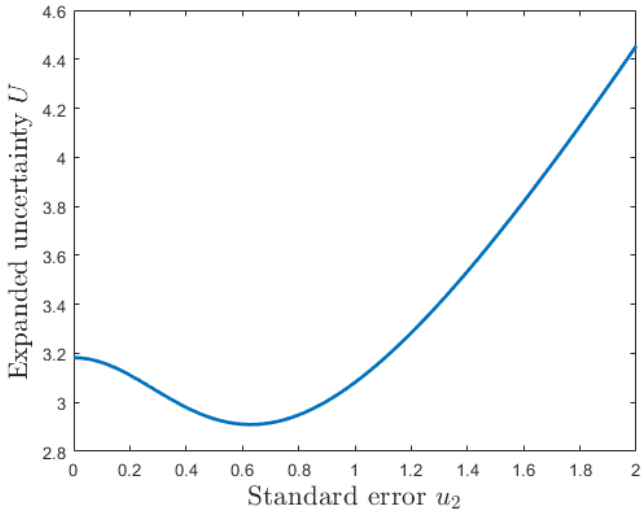

Figure 1: Anomalous behaviour of expanded uncertainty in a two-sample example

Ballico suggests, presumably on the basis of exploratory computations, that this anomaly occurs whenever one input has large uncertainty but low degrees of freedom, while the other has small uncertainty but high degrees of freedom. However, he does not go beyond these qualitative conclusions and presents no theoretical arguments.

2.2 The Hall and Willink example

Ballico's second example, as shown in Figure 1, demonstrates the anomalous behaviour as u_2 increases from zero, but in practice both u_1^2 and u_2^2 are random variables whereas it is the underlying population variances σ_1^2 and σ_2^2 that are fixed. In their example, Hall and Willink compute the expected value of the expanded uncertainty, which we denote by $M = E(U)$, by sampling from the underlying normal distributions, for various values of σ_2^2.

In their example, $m = 2, \sigma_1 = 2$, $d_1 = 3$, d_2 is infinite and σ_2 varied from 0 to 2. Table 2 gives their computed values of M and the coverage probability of the implied W-S 95 % confidence interval for various values of σ_2. We see the anomalous behaviour with M decreasing as σ_2 increases from 0 to 0.2 and 0.4.

Table 2: Hall and Willink's example

σ_2	M	Coverage
0.0	2.93	0.9504
0.2	2.86	0.9403
0.4	2.77	0.9353
0.6	2.80	0.9384
0.8	2.87	0.9424
1.0	3.04	0.9450
1.2	3.27	0.9473
1.4	3.53	0.9489
1.6	3.82	0.9493
1.8	4.13	0.9499
2.0	4.46	0.9499

The coverage of the W-S interval is shown to be less than the nominal 95 % for all $\sigma_2 > 0$, reaching a low point of about 93.5 %. Hall and Willink argue that this is sufficiently close to the nominal coverage for W-S to be acceptable as a practical method for obtaining an approximate confidence interval. We will return to this claim after examining the circumstances under which the anomaly arises.

2.3 Characterising the anomaly

We first consider the case $m = 2$, and note that in Figure 1 the expanded uncertainty decreases as u_2 increases from zero to some point and then increases thereafter. The same behaviour is seen in Table 2 as σ_2 increases, and we have observed this when the anomaly arises in our own numerical explorations. We will begin by deriving an explicit condition for when the derivative is negative at the origin in the case $m = 2$.

When $m = 2$ we have

$$\theta = \mu_1 + \mu_2 \ , \qquad u^2 = u_1^2 + u_2^2 \ , \qquad d = u^4 \left(\frac{u_1^4}{n_1 - 1} + \frac{u_2^4}{n_2 - 1} \right)^{-1} \ .$$

The expanded uncertainty

$$U = k(d)(u_1^2 + u_2^2)^{1/2}$$

is a function of the two random variables u_1^2 and u_2^2, which have independent chi-square distributions: $u_i^2 \sim \{d_i(d_i + 1)\}^{-1}\sigma_i^2\chi_{d_i}^2$.

We will write $u_2^2 = w_2^2\sigma_2^2$ to show explicitly the dependence on the population variance σ_2^2. As σ_2 increases from zero for fixed w_2 we effectively study Ballico's anomalous behaviour of U as u_2 increases, while by taking expectations with respect to the distributions of u_2^2 and w_2^2 we study Hall and Willink's anomalous behaviour of M as σ_2 increases.

The following results are proved in the Appendix.

$$U^* = k(d_1)u_1 \ , \tag{3}$$

$$M^* = k(d_1)[d_1(d_1 + 1)]^{-1/2}\sigma_1 r_0 \ , \tag{4}$$

$$\frac{\partial}{\partial\sigma^2}U^* = \tfrac{1}{2}\,(w_2^2/u_1)[k(d_1) + 4d_1k'(d_1)] \ , \tag{5}$$

$$\frac{\partial}{\partial\sigma^2}M^* = \tfrac{1}{2}[d_1(d_1 + 1)]^{1/2}[\sigma_1(d_2 + 1)]^{-1}(d_1 - 1)^{-1}r_0[k(d_1) + 4d_1k'(d_1)] \ , \tag{6}$$

where

$$r_0 = \sqrt{2}\,\frac{\Gamma((d_1 + 1)/2)}{\Gamma(d_1/2)}$$

and the superscript * denotes a quantity evaluated at $\sigma_2^2 = 0$.

Of particular note here are equations (5) and (6). Since $k(\nu)$ is a decreasing function of ν, $k'(d_1)$ is the one negative term in these expressions, and in particular both $\partial U^*/\partial\sigma^2$ and $\partial M^*/\partial\sigma^2$ have the same sign as $k(d_1) + 4d_1k'(d_1)$. We find that this sign is negative for $d_1 = 1, 2, 3, 4, 5$, but not for $d_1 \geq 6$. Thus, the derivative of both U and M will be negative at $\sigma_2^2 = 0$ whenever $n_1 < 7$. By symmetry, the derivative with respect to σ_1^2 will be negative at $\sigma_1^2 = 0$ when $n_2 < 7$.

Notice that when $d_1 < 7$, $\partial U^*/\partial\sigma^2 < 0$ for *all* values of u_1 and w_2. The fact that $\partial M^*/\partial\sigma^2 < 0$ follows immediately, and there is essentially no difference between Ballico's exposition with these data fixed, focusing on the behaviour of U, and Hall and Willink's focus on the behaviour of M by averaging over u_1 and w_2.

Following Ballico, Figure 1 is plotted against u_2 rather than u_2^2, or equivalently against σ_2 rather than σ_2^2. This is why it does not show a negative derivative at the origin. For any function f,

$$\frac{\partial}{\partial\sigma_2}f = 2\sigma_2\frac{\partial}{\partial\sigma_2^2}f \implies \frac{\partial}{\partial\sigma_2}f^* = 0 \ .$$

However,

$$\frac{\partial^2}{(\partial\sigma_2)^2}f = 2\frac{\partial}{\partial\sigma_2^2}f + 4\sigma_2^2\frac{\partial^2}{(\partial\sigma_2^2)^2}f \implies \frac{\partial^2}{(\partial\sigma_2)^2}f^* = 2\frac{\partial}{\partial\sigma_2^2}f^* \ .$$

Therefore, although the first derivative in Figure 1 is zero, the anomaly is seen because the second derivative is negative.

Hall and Willink's data in Table 2 were computed by Monte Carlo sampling. For their example, equation (4) gives $M^* = k(3)r_0/\sqrt{3}$, where $d(3) = 3.18245$ and $r_0 = \sqrt{2}\Gamma(2)/\Gamma(1.5) = 2\sqrt{(2/\pi)}$. Thus, $M^* = 2.932$, agreeing with the number in Table 2.

We now address the case of $m > 2$. Let

$$\mu_{2+} = \sum_{i=2}^{m} \mu_i \ , \qquad u_{2+}^2 = \sum_{i=2}^{m} u_i^2 \ ,$$

and

$$d_{2+} = u_{2+}^4 \left(\sum_{i=2}^{m} \frac{u_i^4}{d_i} \right)^{-1} \ . \tag{7}$$

Then it is straightforward to show that

$$\theta = \mu_1 + \mu_{2+} \ , \qquad u^2 = u_1^2 + u_{2+}^2 \ , \qquad d = u^4 \left(u_1^4/d_1 + u_{2+}^4/d_{2+} \right)^{-1} \ .$$

The case $m > 2$ can thereby be reduced to that of $m = 2$. Thus, if $n_1 < 7$, the anomaly will arise whenever u_{2+} is sufficiently small.

Ballico asserted that the anomaly arises when one sample size is sufficiently small, without specifying how small it had to be, and also in his examples the other sample size(s) were large. We have shown that it arises when one of the m sample sizes is less than 7, regardless of the other sample sizes.

2.4 W-S coverage

The value of U at $\sigma_2^2 = 0$, equation (3), is the expanded uncertainty for a single sample. It does not rely on the W-S approximation and provides an exact 95 % confidence interval for θ. (The number 0.9504 in Table 2 suggests that Hall and Willink's Monte Carlo sample size was not large enough to guarantee accuracy to the full number of digits quoted.) The result of the anomaly, then, is that as σ_2^2 increases from zero, we have more uncertainty in the data-generating process but the expected W-S interval is shorter. Therefore, the coverage of this interval must be less than the nominal 95 %. And as σ_2^2 continues to increase the coverage continues to decrease as long as the interval is shrinking. Table 2 shows not only this effect but also that the coverage remains below the nominal 0.95 for a substantially wider range of σ_2^2 values

Hall and Willink give only the one example, in which the coverage does not fall below 93.5 % but Guthrie [11] finds that it can be lower than 88 %. We have also found instances in which the coverage is less than 90 %. We do not have theory to show how far below the nominal 95 % the true coverage can be.

We also note that if either sample size is less than seven the user will *know* that the coverage of the W-S interval may be less than 95 %. The related phenomenon of *relevant* or *recognizable subsets* [16], whereby the user knows that when the data fall into some subset of the sample space a confidence interval has less or more than the nominal coverage conditionally, is regarded as a highly undesirable property.

Overall, we find Hall and Willink's defence of W-S unconvincing.

3 Welch-Satterthwaite in metrology

3.1 Metrology and the GUM

Metrology is the science of measurement. From the national metrology institutes, which conduct research into novel and improved measurements, to some 100 000 accredited testing and calibration laboratories worldwide, metrologists are concerned with making accurate measurements of all kinds. Measurement is an essential part of human activity. Estimates of quantities are required for a diverse range of applications and for each of these estimates a statement is needed about its quality. Such a statement is usually

made with measurement uncertainty. The two components, the estimate and the associated uncertainty, together constitute a common way of reporting a measurement result [7].

Measurement uncertainty plays an important role in many areas such as assessing compliance with regulation, and in the calibration of measuring systems. Accredited calibration and testing laboratories are obliged [1] to state the uncertainty of their measurement results, so that recipients can take that uncertainty into account when evaluating their own results.

The Joint Committee for Guides in Metrology (JCGM) is responsible for maintaining and promoting the use of the *Guide to the expression of uncertainty in measurement* (GUM) [5] and the *International vocabulary of basic and general terms in metrology* (VIM) [7]. The GUM (JCGM 100) has for a long time been the authoritative document concerned with the evaluation and expression of measurement uncertainty that attempts to meet this objective:

> This Guide establishes general rules for evaluating and expressing uncertainty in measurement that can be followed at various levels of accuracy and in many fields — from the shop floor to fundamental research. Therefore, the principles of this Guide are intended to be applicable to a broad spectrum of measurements ...

The GUM and its statistical methods for assessing measurement uncertainty are used daily in thousands of laboratories around the world. Although little known in the mainstream statistical community, metrology is a major application area for statistics.

The GUM uses the concept of *standard uncertainty*, which is specifically defined [5, clause 2.3.1] as the 'uncertainty of the result of a measurement expressed as a standard deviation'.

The GUM also advocates the use of *coverage intervals* having a specified probability of covering the true value of the measurand. Although it is sometimes reasonable to suppose that the range $x \pm 2u$, where x is the estimate and u the standard uncertainty, will have approximately 95 % coverage, particularly when the measurement is based on a large sample of data, small samples are routinely used in practice. The GUM introduces the *expanded uncertainty* [5, clause 6.2] $U = ku$, where k is a *coverage factor* usually chosen such that $x \pm U$ is a 95 % interval.

The GUM treats measurement as in general involving a *measurement model* relating the *measurand* (the quantity intended to be measured) Y to input quantities X_i:

$$Y = f(X_1, \dots X_N).$$

Knowledge of Y can be determined given f and knowledge of the X_i. Typically, the GUM itself requires estimates x_i of the X_i, associated standard uncertainties $u(x_i)$ and possibly covariances between the X_i.

We first consider a measurement that fits the model of Section 1. That is, the measurand Y is a linear combination of m quantities X_i, which are themselves evaluated from independent normal samples. Formally, we identify the value of the quantity X_i with the mean parameter μ_i of its sample. As in Section 1, we assume without loss of generality that the linear combination is a simple sum, so that we identify the value of the measurand Y with the sum of means θ.

For a linear measurement model, the GUM [5, clause G.4.1] recommends the use of the Welch-Satterthwaite approximation as in Section 1 to obtain an approximate expanded uncertainty

$$U = k(d)u$$

for the measurand. The significance of U in metrology goes beyond its role in determining a 95 % coverage interval. The authors [9] have argued that a more meaningful expression of uncertainty in the measurand than the standard uncertainty u is the *characteristic uncertainty*, which in this case is defined as half the expanded uncertainty. Thus, U may be seen as having an even more fundamental role in determining a measure of uncertainty.

The routine use of the W-S approximation in metrology is not restricted to measurements that fit the model of Section 1, namely the case of a linear measurement model and input quantities evaluated from normal samples.

- If the measurement model is not linear in the X_i the GUM [5, clause 5.1.2] recommends linearising the model by a first order Taylor series expansion.

- If the individual sample observations x_{ij} are not normally distributed, the GUM [5, clause G.2.1] invokes the Central Limit Theorem (CLT) to treat the sample means $\bar{x}_i$ as normal.
- If inputs are evaluated by means other than through samples, the W-S approximation may still be used as discussed in Section 3.3.

Although the JCGM has produced two documents on the propagation of probability distributions in models that are not necessarily linear and with data that are not necessarily normally distributed (JCGM 101 [8] for a single measurand and JCGM 102 [6] for multivariate measurands), by far the most common practice in metrology is to use the recommendations of the GUM itself.

3.2 Significance of the anomaly in metrology

The W-S approximation is therefore very widely used in metrology and, because of the expense involved in taking each individual observation, sample sizes as small as three or four are very common. Ballico reports that the anomaly was drawn to his attention in a practical measurement problem with five input quantities. In this instance, two related measurands were being measured with the same measurement model and with shared evaluations of some of the inputs, a situation that made it possible for the anomaly to be noticed.

In practice, the anomaly will typically go unnoticed, but when one sample size is less than seven and uncertainty in another input is sufficiently small its consequences are nevertheless real:

- the W-S 95 % interval will have less than 95 % coverage, and
- the characteristic uncertainty will be too small, giving an over-optimistic expression of uncertainty in the measurand.

Also, as noted in Appendix A.2, these consequences will follow even if the W-S approximation is used in the context of non-normal observations by appealing to the CLT.

Furthermore, it is common in metrology for the measurement model to include both one or more inputs that have been evaluated using small samples and other inputs that are corrections for potential biases or rounding which are judged to have small uncertainty (in the sense of Type B evaluations, discussed in Section 3.3). Therefore, the conditions for the W-S anomaly to arise are particularly prevalent in this field, leading to the true coverage of the W-S 95 % interval being less than 95 % and the implied uncertainty in the measurand being understated.

We know [11] that the coverage can be as low as 88 %, but even if, as Hall and Willink argue, the coverage is generally not far short of the nominal 95 %, the fact that we know in these circumstances that it has less than the stated coverage means that in our opinion the use of W-S should be deprecated on principle.

3.3 W-S with Type B evaluations

In the GUM a distinction is made between Type A and Type B evaluation of inputs to a measurement model. A Type A evaluation uses statistical methods to analyse observational data, and the most common such evaluation in practice is the case we have been considering of a sample of observations that are either assumed to be normally distributed or else the CLT is used to assume that the distribution of the sample mean is normal. A Type B evaluation, however, uses a knowledge-based probability distribution for an input quantity. The GUM [5, clauses E.3, G.4.2] asserts that both Type A and Type B evaluations can be combined to give the estimate t and the standard uncertainty u of the measurand θ as in Section 1, and furthermore that the W-S approximation may still be used to assign the expanded uncertainty U. In order to do this for a quantity X_i that is subject to Type B evaluation, the metrologist makes judgements, based on available data sources and his or her knowledge and expertise, to assign an estimate of X_i that is treated as $\bar{x}_i$, a standard deviation that is treated as u_i and a degrees of freedom d_i.

The GUM approach of combining Type A and Type B evaluations is enshrined in many procedures and much software for measurement uncertainty evaluation since the GUM was first published in 1993. This procedure is, however, highly controversial [10, 13, 14, 15]. Our position on the controversy is that Type B

evaluation must be treated as being based on subjective probability, and that if Type A evaluations are conducted using frequentist analysis of the observations then there is no formal justification for combining Type A and Type B evaluations. The only logical and coherent way to proceed must be by using Bayesian analysis in Type A evaluation, because then both the posterior probability distribution from a Type A evaluation and the metrologist's subjective probability distribution that constitutes Type B evaluation can be combined using the standard laws of probability. W-S would have no place in such an approach [3, 4]. However, there is resistance in the metrology community to adopting a fully Bayesian approach [3] and attempts have been made to justify the GUM's recommendations, including by arguing that Type B evaluations can be treated as if they represented frequentist probability statements. This is the position taken by Hall and Willink in the example we discuss in Section 2.2. They suppose that in a Type B evaluation of the input X_2 it has been assigned a normal distribution, which is why they assign an infinite degrees of freedom d_2. Input X_1, however, is evaluated from a sample of $n_1 = 4$ normal observations with variance $\sigma_1^2 = 4$.

Their results shown in Table 2 were simulated by Monte Carlo, wherein at each iteration a sample of 4 normal observations is drawn for X_1, but for X_2 they draw a single normal observation from its Type B distribution. We fail to see how this simulation can be justified, for two reasons. First, their sampling does not fit the statistical model from which W-S is derived, which with infinite d_2 would require in principle a sample of infinite size. Instead, they take a single observation with variance σ_2^2, call this $\bar{x}_2$ and assign $u_2 = \sigma_2$. Second, $\bar{x}_2$ should be fixed, being the metrologist's considered estimate of X_2. It is not a random quantity that can be sampled. By treating a Type B evaluation as if it were somehow the result of a random process, Hall and Willink are ignoring the nature of Type B evaluation and the principles of frequentist inference.

4 Conclusions

If we increase the error variance of observations, keeping all other elements of a statistical problem unchanged, it does not make sense for inference about a parameter in that problem to become more precise, in the sense that a confidence interval contracts. Yet exactly this anomalous behaviour can arise when using the Welch-Satterthwaite approximation with small samples.

In support of observations of this anomaly that have been made in some practical circumstances, we prove that the anomaly can occur whenever one sample size is less than seven.

When the anomaly arises, W-S 95 % intervals will have true coverage less than 95 %. Although Hall and Willink assert that the shortfall in coverage is minor, they present only one numerical example in support of their claim.

The width of a confidence interval is an important indicator of uncertainty, particularly in the field of metrology where the problem was identified, and in this sense the anomaly results in an understatement of uncertainty.

We suggest that the Welch-Satterthwaite formula should not be used when one of the contributing samples has size less than seven and, since it is approximate, used with caution otherwise.

5 Acknowledgements

This work was supported by an ISCF (Industrial Strategy Challenge Fund) Metrology Fellowship grant provided by the UK government's Department for Business, Energy and Industrial Strategy (BEIS). Alistair Forbes made a valuable review of a draft of this paper.

A Appendix

A.1 Expected expanded uncertainty

In the case of two normal samples we have

$$\theta = \mu_1 + \mu_2 \ , \qquad u^2 = u_1^2 + u_2^2 \ , \qquad d = u^4 \left(u_1^4/d_1 + u_2^4/d_2 \right)^{-1} \ ,$$

where d is the W-S 'effective degrees of freedom' for θ and $d_i = n_i - 1$ is the degrees of freedom for the i-th sample. The half-width of the resulting W-S 95 % confidence interval is the expanded uncertainty:

$$U = k(d)(u_1^2 + u_2^2)^{1/2} \ ,$$

where $k(d)$ is the upper 97.5 % percentage point of the Student's t distribution with d degrees of freedom. Our interest is in the behaviour of U and of $M = E(U)$, its expected value with respect to the independent random variables $u_i^2 \sim [d_i(d_i+1)]^{-1}\sigma_i^2\chi_{d_i}^2$.

In general, since U is a complex function of the u_i^2, it is not possible to obtain a closed-form expression for M. However, we can obtain explicitly its value and derivatives at $\sigma_2^2 = 0$.

To simplify the algebra, let

$$\phi = \sigma_2^2 \ , \qquad w_2^2 = u_2^2/\phi \ .$$

Then U is expressed explicitly as a function of ϕ through

$$U = k(d)(u_1^2 + \phi w_2^2)^{1/2} \ ,$$

and

$$d = (u_1^2 + \phi w_2^2)^2 \left(u_1^4/d_1 + \phi^2 w_2^4/d_2 \right)^{-1} \ . \tag{8}$$

We will denote any quantity evaluated at $\phi = 0$ with a superscript *, and hence

$$U^* = k(d_1)u_1 \ .$$

When $X \sim \chi_k^2$, $E(X^m) = 2^m\Gamma(m + k/2)/\Gamma(k/2)$. Therefore

$$M^* = k(d_1)[d_1(d_1+1)]^{-1/2}\sigma_1 r_0 \ ,$$

where

$$r_0 = \sqrt{2}\ \Gamma((d_1+1)/2)/\Gamma(d_1/2) \ .$$

A.2 First derivative

We introduce the differential operator $\Delta = \frac{\partial}{\partial\sigma_2^2} = \frac{\partial}{\partial\phi}$.

Rationalizing and differentiating expression (8),

$$\Delta d \left(u_1^4/d_1 + \phi^2 w_2^4/d_2 \right) + 2\, d\phi w_2^4/d_2 = 2\, w_2^2(u_1^2 + \phi w_2^2).$$

Evaluating at $\phi = 0$ gives

$$\Delta d^* = 2 d_1 w_2^2/u_1^2 \ ,$$

and we note that this is always positive. The fact that increasing uncertainty in the data can paradoxically increase the effective degrees of freedom is the root of the W-S anomaly, because increasing d will decrease $k(d)$.

Continuing the differentiation,

$$\Delta U = \tfrac{1}{2}\, k(d) w_2^2 (u_1^2 + \phi w_2^2)^{-1/2} + k'(d)(\Delta d)(u_1^2 + \phi w_2^2)^{1/2} \ ,$$

where $k'(\nu) = \frac{\partial}{\partial\nu} k(\nu)$. Therefore,

$$\Delta U^* = \tfrac{1}{2}\, (w_2^2/u_1)[k(d_1) + 4 d_1 k'(d_1)] \ .$$

Since the operations of expectation and differentiation commute, we can now obtain the derivative of M at $\phi = 0$:

$$\Delta M^* = \tfrac{1}{2}[d_1(d_1+1)]^{1/2}[\sigma_1(d_2+1)]^{-1} r_1[k(d_1) + 4d_1 k'(d_1)] \ ,$$

where

$$r_1 = 2^{-1/2}\,\Gamma((d_1-1)/2)/\Gamma(d_1/1) = (d_1-1)^{-1} r_0 \ .$$

Note that the above holds for $d_1 > 1$, and hence for $n_1 > 2$. If $n_1 = 2$, r_1 is infinite.

If the distribution of the observations x_{ij} with which the W-S approximation was introduced in Section 1 is not normal, then the u_i^2 will not have chi-square distributions. However, as long as the distribution of u_2^2 admits of a scale parameter σ_2^2, so that the distribution of w_2^2 is independent of σ_2^2, then the above results will hold in the following sense: both U^* and M^* will be multiples of $k(d_1)+4d_1k'(d_1)$, for fixed u_1^2 and w_2^2 or averaged with respect to those random variables, respectively.

References

[1] ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories.

[2] Ballico, M. Limitations of the Welch-Satterthwaite approximation for measurement uncertainty calculations. *Metrologia 37*, 1 (2000), 61–64.

[3] Bich, W., Cox, M., and Michotte, C. Towards a new GUM—an update. *Metrologia 53*, 5 (2016), S149.

[4] Bich, W., Cox, M. G., Dybkaer, R., Elster, C., Estler, W. T., Hibbert, B., Imai, H., Kool, W., Michotte, C., Nielsen, L., Pendrill, L., Sidney, S., van der Veen, A. M. H., and Wöger, W. Revision of the 'Guide to the Expression of Uncertainty in Measurement'. *Metrologia 49*, 6 (2012), 702–705.

[5] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.

[6] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Models with any number of output quantities. Joint Committee for Guides in Metrology, JCGM 102:2011.

[7] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. International Vocabulary of Metrology — Basic and General Concepts and Associated Terms. Joint Committee for Guides in Metrology, JCGM 200:2012.

[8] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology, JCGM 101:2008, 2008.

[9] Cox, M. G., and O'Hagan, A. Meaningful expressions of uncertainty in measurement. Tech. Rep. MS 27, National Physical Laboratory, 2021.

[10] Elster, C. Bayesian uncertainty analysis compared with the application of the GUM and its supplements. *Metrologia 51*, 4 (2014), S159.

[11] Guthrie, W. F. Should $(T_1 - T_2)$ have larger uncertainty than T_1? In *8th International Conference on Temperature: Its Measurements and Control*, vol. 2, pp. 887–892.

[12] Hall, B. D., and Willink, R. Does "Welch-Satterthwaite" make a good uncertainty estimate? *Metrologia 38* (2001), 9–15.

[13] Kacker, R., and Jones, A. On use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent. *Metrologia 40* (2003), 235–248.

[14] Lira, I. The GUM revision: the Bayesian view toward the expression of measurement uncertainty. *European Journal of Physics 37*, 2 (2016), 025803.

[15] Lira, I., and Wöger, W. Comparison between the conventional and Bayesian approaches to evaluate measurement data. *Metrologia 43* (2006), S249–S259.

[16] Robinson, G. K. Conditional Properties of Statistical Procedures. *The Annals of Statistics 7*, 4 (1979), 742 – 755.

[17] Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin 2*, 6 (1946), 110–114.

[18] Welch, B. L. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika 29*, 3/4 (1938), 350–362.

[19] Welch, B. L. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika 34*, 1/2 (1947), 28–35.

[20] Wikipedia. Behrens-Fisher problem — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Behrens%E2%80%93Fisher%20problem&oldid=993405382, 2021. [Online; accessed 20-April-2021].