Don't forget the degrees of freedom: evaluating uncertainty from small numbers of repeated measurements

Blair Hall

b.hall@irl.cri.nz

Talk given via internet to the 35$^{\text{th}}$ ANAMET Meeting, October 20, 2011.

Motivation

Is there a need for accurate uncertainty statements about complex quantities when dealing with small numbers of repeated measurements?

The GUM introduced a calculation of *effective degrees of freedom* to deal with the case of small numbers of repeat measurements.

Effective degrees of freedom:

- extends a classical statistical concept
- needed to calculate coverage factors
- currently out of favour with metrology theorists
- GUM method does not apply to complex quantities

There is an *effective degrees of freedom* calculation for complex problems too.

No one seems to be using it.

Notes

The notion of effective degrees of freedom (DoF) was introduced in the GUM.

The basic tool for uncertainty propagation in the GUM is the Law of Propagation of Uncertainty (LPU), which evaluates a standard uncertainty for the measurement result. The LPU takes into account the standard uncertainties of all measurement influence quantities and the sensitivity of the measurement to each influence.

In classical statistics, the degrees of freedom is a measure of how reliable the sample standard deviation is as an estimate of the population standard deviation. The GUM adopted this notion, and expanded it to help quantify how good a combined standard uncertainty is as an estimate of the standard deviation of the distribution of measurement errors.

So, in addition to the LPU, there is a calculation in the GUM that propagates information about finite sample sizes. It is called the Welch-Satterthwaite formula and obtains an effective degrees of freedom for the result. That number, ν_{eff}, is used to calculate a coverage factor for the expanded uncertainty. Without an accurate value for ν_{eff}, the coverage probability of the uncertainty statement will be incorrect.

The GUM extended *classical* methods

Suppose you wish to measure the physical quantity

$$\theta = \mu_1 + \mu_2$$

You can observe

$$x_{1 \cdot i} = \mu_1 + \varepsilon_{1 \cdot i}, \quad x_{2 \cdot i} = \mu_2 + \varepsilon_{2 \cdot i}$$

where $\varepsilon_{1 \cdot i}$ and $\varepsilon_{2 \cdot i}$ are measurement errors in each observation.

$$\varepsilon_{1 \cdot i} \sim N(0, \sigma_1), \quad \varepsilon_{2 \cdot i} \sim N(0, \sigma_2)$$

- Samples of sizes n_1 and n_2 are taken.
- The sample means x_1, x_2 provide an estimate of

$$\theta \approx y = x_1 + x_2$$

- The standard error in the sample means estimates

$$\frac{\sigma_1}{\sqrt{n_1}} \approx u(x_1), \quad \frac{\sigma_2}{\sqrt{n_2}} \approx u(x_2)$$

Notes

To understand the thinking behind effective degrees of freedom, we need to consider classical statistics.

In this case, we imagine two independent noise sources that perturb observations of μ_1 and μ_2. These random errors have Gaussian distributions with zero mean. We do not know the respective variances, σ_1^2 and σ_2^2: they are estimated from our observations.

We collect n_1 observations for μ_1, and n_2 for μ_2. The sample statistics are

$$\overline{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1 \cdot i}, \quad \overline{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2 \cdot i}$$

$$s(\overline{x}_1)^2 = \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^{n_1} (x_{1 \cdot i} - \overline{x}_1)^2, \quad s(\overline{x}_2)^2 = \frac{1}{n_2(n_2 - 1)} \sum_{i=1}^{n_2} (x_{2 \cdot i} - \overline{x}_2)^2$$

In the conventional GUM notation, we would write x_1, instead of $\overline{x}_1$, and $u(x_1)^2$, instead of $s(\overline{x}_1)^2$, etc.

... extended GUM methods: Welch-Satterthwaite

The standard uncertainty associated with our estimate of θ is
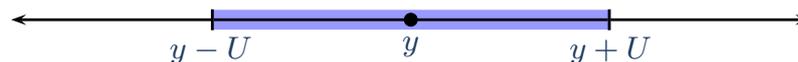
$$u(y) = \sqrt{u(x_1)^2 + u(x_2)^2}$$

The effective degrees of freedom ν_{eff} are found (approximately) using the Welch-Satterthwaite formula

$$\frac{u(y)^4}{\nu_{\text{eff}}} = \frac{u(x_1)^4}{n_1 - 1} + \frac{u(x_2)^4}{n_2 - 1} \, ,$$

A number of degrees-of-freedom ν_{eff} is an effective measure of the sample size. It is used to find a coverage factor k_p, for a given level of confidence p.

The expanded uncertainty is then

$$U = k_p \, u(y)$$

$$y - U \qquad\qquad y \qquad\qquad y + U$$

Notes

The last step in which the degrees of freedom are evaluated is approximate. In this (textbook) problem the distribution of error in y is also Gaussian, but it is difficult to obtain an expression for the associated sampling distribution of $u(y)$.

If we had more than two influence quantities, then there is no known solution.

The Welch-Satterthwaite formula is an approximate method that works well in most cases.

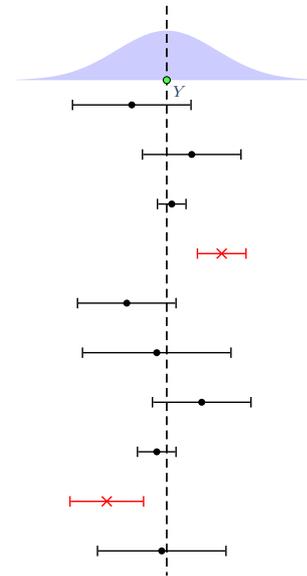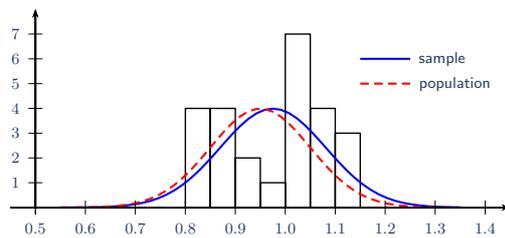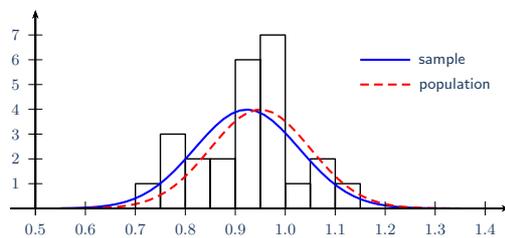The coverage factor is found from tables of the Student t-distribution

$$k_p = t_{\alpha, \nu_{\text{eff}}} \, , \quad \text{with } \alpha = (1 + p)/2$$

$t_{\alpha, \nu}$ is the $100\,\alpha^{\text{th}}$ percentile of the t-distribution with ν degrees of freedom

Motivation
Real quantities (GUM)
Complex quantities
Discussion

The sample mean and standard deviation are not equal to the parameters of the population (actual distribution)

Notes

The figures on the left illustrate the effect of sample size.

The figures show two independent sets of 25 'measurements' of a quantity X. A random error with a standard deviation of 0.1 has been added to X to simulate measurement errors.

In the top figure, the sample mean = 0.973 and the sd = 0.105

In the bottom figure, the sample mean = 1.025 and the sd = 0.105

Note that we can neither measure the value of the quantity, nor the amplitude of the error, exactly.

The figure on the right shows how the variability of sampling statistics leads to occasional uncertainty statements (expanded uncertainties) that do not cover the measurand. The coverage probability indicates the frequency, over many independent measurements, with which uncertainty statements will cover the measurand.

Yes, Welch-Satterthwaite does work!

- In the GUM, a number of effective degrees of freedom is needed to calculate the <u>expanded uncertainty</u> for a given coverage probability.

- An alternative, propagation of distribution (PoD) method does not use DoF. This is currently favoured by many metrology theoreticians and is described in the BIPM GUM supplements.

- Proponents of PoD assert its superiority. However, the view that PoD is correct by definition, which is held by many, leads to confusion among practitioners.

- The confusion stems from the lack of a clear, commonly held, understanding of *coverage probability*.

- Studies [1,2] show WS performing better than PoD (always) when coverage probability is interpreted in terms of many independent experiments.

[1] H Tanaka and K Ehara, *Validity of expanded uncertainties evaluated using the Monte Carlo Method*, in Advanced Mathematical and Computational Tools in Metrology and Testing (World Scientific, 2009) pp 326-329.

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

Notes

The PoD method is often referred to as the Monte Carlo Method. There are, however, many ways of using Monte Carlo methods for numerical computations. In fact, the *implementation* of PoD, as described in the GUM supplements, uses a Monte Carlo method.

The GUM approach to uncertainty calculation assumes that the underlying distribution of error in a measurement result is Gaussian. Then, it uses the effective DoF and the Student t-distribution to obtain a coverage factor k, which scales the extent of the expanded uncertainty.

Clearly, the Gaussian assumption will not always hold. However, in most well-designed measurement scenarios it is reasonable. That is why the GUM approach is so useful.

Some practitioners have expressed concern that the expanded uncertainty calculated by GUM methods was incorrect. References in [2] are an early example of this. However, such concerns do not seem to be well-considered.

Coverage probability has to be thought of as the long-run-average rate for uncertainty intervals to cover the measurand in repeated and independent experiments. While this criterion cannot readily be verified in practical measurements, it is the *only* reasonable interpretation of coverage probability that has been given to date.
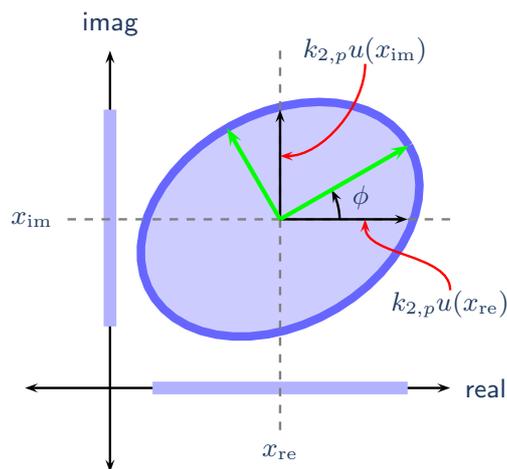
Complex uncertainty regions

The uncertainty region is an ellipse.

It's orientation and shape depend on the standard uncertainties in the real and imaginary components $u(x_{\mathrm{re}})$, $u(x_{\mathrm{im}})$ and on the correlation r between x_{re} and x_{im}

The size of the ellipse is set by a two-dimensional coverage factor $k_{2,p}$, which is a function of the degrees of freedom.

Notes

The standard uncertainties in the real and imaginary components are equal to the square root of the diagonal terms in the covariance matrix.

The degrees of freedom associated with the standard uncertainties in the real and imaginary components will be the same.

The principal axis lengths are not, in general, proportional to $u(x_{\mathrm{re}})$ and $u(x_{\mathrm{im}})$.

The length of the axes is however proportional to the eigenvalues of the covariance matrix.

Level of Confidence

A coverage factor can be calculated from ν and the coverage probability p

$$k_{2,p}^2(\nu) = \frac{2\nu}{\nu - 1} F_{2,\nu-1}(p) \,,$$

- $F_{2,\nu-1}(p)$ is the upper $100p^{\mathrm{th}}$ percentile of the F-distribution
- ν is a measure of the sample size (c.f. $\nu = N - 1$)
- when ν is infinite, $k_{2,p}^2(\infty) = \chi_{2,p}^2$

Complex coverage factors $k_{2,p}$ are <u>not</u> the same as the k_p used in real uncertainty statements:

ν	$k_{2,0.95}$	$k_{0.95}$	ν	$k_{2,0.95}$	$k_{0.95}$
2	28.3	4.3	7	3.5	2.4
3	7.6	3.2	8	3.3	2.3
4	5.1	2.8	9	3.2	2.3
5	4.2	2.6	10	3.1	2.2
6	3.7	2.5	50	2.6	2.0
			∞	2.45	1.96

Notes

A useful relation in the infinite dof case is

$$\chi_{2,p}^2 = -2\log_{\mathrm{e}}(1 - p)$$

It is worth emphasizing the fact that the area of the uncertainty region is scaled by $k_{2,p}$ <u>squared</u>. So the coverage factor is important when ν is small.

What is the effect of ignoring the degrees of freedom completely? If an uncertainty statement is evaluated using the coverage factor for infinite degrees of freedom, when in fact there are finite degrees of freedom, then the actual coverage probability will be less then nominal.

For example, this table shows how actual coverage probability falls off if the coverage factors $k_p = 1.96$ and $k_{2,p} = 2.45$ are used to evaluate 95% expanded uncertainty instead of correct values for three sample sizes ($\nu = 3, 5, 10$).

ν	coverage (real)	coverage (complex)
3	85%	66%
5	89%	78%
10	92%	88%

Clearly, it is important to take the degrees of freedom into account when calculating a coverage factor. This is <u>even more</u> important for uncertainty statements about complex quantities.

Effective degrees of freedom

There is a method for calculating ν_{eff} in the complex case [1].

As in the GUM (WS) method, there are some restrictions:

- Correlation is allowed between real and imaginary components of individual influence quantities

- Correlation is <u>not</u> allowed between real and imaginary components of <u>different</u> influence quantities

- Tests show that the method provides good overall performance in terms of coverage

[1] R Willink and B D Hall, *A classical method for uncertainty analysis with multidimensional data*, Metrologia **39** (2002) pp 361–369.

Notes

Reference 1 describes three methods for evaluating an effective degrees of freedom. The 'Total Variance' method is preferred for propagation of uncertainty. An appendix to the paper shows how to apply the total variance method to the complex (bivariate) problem.

There are some cases for which the Total Variance method did not perform well, but this is not unexpected. The Welch-Satterthwaite formula is also known to have problems in some cases.

Coverage can fall well below nominal when there is a very small number of samples for one quantity and a very large number of samples for the other (e.g., $n_1 = 3$ and $n_2 = \infty$). In such cases, the mean coverage, over a wide range of population parameters, drops slightly below 95% (e.g., 93% for $n_1 = 3$ and $n_2 = \infty$, which was the worst case in the study). However, the range of coverage observed becomes wide (for $n_1 = 3$ and $n_2 = \infty$, the minimum coverage was 81% and the maximum 99%). In other words, there are a few populations with rather extreme covariance structures that are not handled well by the total variance calculation.

The PoD alternative

The PoD method uses a form of bivariate t-distribution to model a quantity estimated from a small sample of measurements.

There are problems with this approach:

- There is not a unique choice of t-distribution in the bivariate case

- One t-distribution does not provide 95% coverage in the simplest case [1] (one quantity). E.g, for $n = 3$, the mean coverage is approximately 77%, rising to approximately 92% for $n = 8$

- Another type is correct for single quantities, but becomes very conservative when applied to the sum of two quantities [1]

- The analytic method gives significantly better performance in all cases [1]

[1] B D Hall, *Monte Carlo uncertainty calculations with small-sample estimates of complex quantities*, Metrologia **43** (2006) pp 220–226; B D Hall, *Complex-value Monte Carlo uncertainty calculations with few observations*, CPEM Digest (9-14 July 2006, Torino, Italy).

Notes

Wikipedia may be consulted for references to multivariate t-distribution literature:

`http://en.wikipedia.org/wiki/Multivariate_Student_distribution`

Any discussion?

Is there a need to evaluate uncertainty for small numbers of repeat measurements?

■ Effective degrees of freedom will be required if calculations of uncertainty for complex quantities use the extended GUM (LPU) method.

◆ Is there any foreseeable need for this?

◆ If so, how will degrees of freedom be handled?

◆ Is it well-known that the 1D (WS) method does not apply (e.g., phase and magnitude uncertainies)?

■ The GUM Supplements favour the PoD method, which does not always generate results compatible with the conventional notion of 'confidence level'.

◆ Is this acceptable?

◆ Some advocates of PoD suggest that the actual coverage is irrelevant. Is that OK with you?

◆ Some colleagues feel it best to follow a common methodology, regardless of 'correctness' (i.e., for consistency)

Notes

As far as I know, few laboratories report the uncertainty of a complex quantity as a region of the complex plane (NPL, being an exception).

For real quantities, many national labs report report uncertainties with a coverage factor of $k = 2$, suggesting that degrees of freedom is not considered to be a concern. Perhaps the situation for complex quantities will be similar.

Is it understood that separate, independent, univariate statements of uncertainty, for instance in the magnitude and phase of a complex quantity, do not provide the same coverage as a bivariate uncertainty region?

RF uncertainty calculations can become pretty complicated (e.g., VNA). So software will be used in future. If that software is designed according to current guidelines, the small-sample problem will not be addressed.

Both software development and producing BIPM uncertainty guides are slow processes. If this issue is not addressed within current activities, it is likely to remain a problem for a long time.