COMMENTS ON THE "PROPAGATION OF ERRORS"

William C. Evans *September 28, 2005*

Most students of the physical sciences and engineering learn a method for the so-called "propagation of error" which is based on a first-order Taylor series expansion. For years, and perhaps still today, the standard text for this material was Bevington [1], so for convenience we will refer to this calculation by this name. The limitations of this *numerical*, not *statistical*[1], approach might be mentioned in passing, but are rarely demonstrated.

The estimation of a confidence interval for a ratio of two Normally-distributed random variables is a relatively commonly-occurring issue in scientific data analysis. We will consider the *proper* calculation of confidence intervals on a ratio, to provide a useful comparison with this approximate "propagation of error" method.

The Bevington calculation begins with a function

$$z = f\left(x_1 \;\; x_2 \;\; x_3 \; \ldots \; x_n\right)$$

where *f* need not be, and often is not, linear, and the *x* are random variables which in general need not be Normally-distributed, and which in general may be mutually correlated. What we seek is the variance of *z*, which is defined as

$$\sigma_z^2 = \mathrm{E}\left[\left(z - \mu_z\right)^2\right]$$

i.e., the expected value E[ ] of the second moment about the mean. The expected value is found using an integral, for the continuous variables we are considering here. To evaluate this integral we need a functional form for the probability density function (PDF) of *z*. It has been noted that

> *The exact calculation of [variances] of nonlinear functions of variables that are subject to error is generally a problem of great mathematical complexity. In fact, a substantial portion of mathematical statistics is concerned with the general problem of deriving the complete frequency distribution [PDF] of such functions, from which the [variance] can then be derived.* [2]

As is usual in applied mathematics, one approach for avoiding complexity is to approximate a function with another, simpler, function, and often this is done using a low-order Taylor series expansion. It can be shown [3] that, if we replace the function *z* with a first-order expansion about a point defined by the mean values of each of the variables *x*, we can then write for the variance of the linearized function

$$\sigma_z^2 \approx \sum_i^n \sum_j^n \left(\frac{\partial z}{\partial x_i}\right)_{\mu_{i,j}} \left(\frac{\partial z}{\partial x_j}\right)_{\mu_{i,j}} \mathrm{cov}\left(x_i \;\; x_j\right) \tag{1}$$

where the summations are taken over all combinations of *i, j* and where it is understood that the covariance of a variable with itself is its variance. The partials are functions of the several variables, each of which is to be evaluated at its mean μ. Of course, to use (1) we must have values for the means and variances of the component variables *x*. Note that if *f* is linear then, *and only then*, (1) is exact. To take a specific example of a function *z*, let us consider the relatively simple, commonly-occurring case of

$$z = \frac{x}{y}$$

with *x* and *y* possibly correlated. Then (1) becomes

[1] It is very rare for a text in applied statistics, at any level, to even mention this "propagation of error."

$$\sigma_z^2 \approx \left(\frac{\partial z}{\partial x}\right)^2_{\mu_{x,y}} \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2_{\mu_{x,y}} \sigma_y^2 + 2\left(\frac{\partial z}{\partial x}\right)_{\mu_{x,y}} \left(\frac{\partial z}{\partial y}\right)_{\mu_{x,y}} \operatorname{cov}(x\ y) \tag{2}$$

and we also note for convenience that the correlation coefficient ρ is defined as

$$\rho = \frac{\operatorname{cov}(x\ y)}{\sigma_x\ \sigma_y}$$

It is often the case that the covariance term in (2) is ignored. *This is unwise as a general policy*, although there are situations where there is a lack of information about the correlation, and so there may be no choice but to set $\rho = 0$.

We can make these observations about the application of (1) or (2):

> Since the derivation used an approximation based on a derivative (Taylor series), the variation in the variables x must be "small," such that they can be considered differentials in x. As a practical matter this means that σ_x / x ("coefficient of variation") should be around 0.10 or less.
>
> When this condition is not met, the linearization can fail, depending on where the evaluation point is on the function. (The evaluation point is a particular set of values of the variables x.) We are attempting to replace a (hyper-)surface in $n+1$ dimensions with a tangent (hyper-)plane, and this clearly will not work as we move farther away from the evaluation point, unless the function happens to be nearly linear in this region. Of course, how well the plane represents the surface also depends on the nonlinearity of the function, and where the evaluation point is, with regard to any discontinuities, etc.

As a practical example to illustrate these issues, we now consider the problem of finding a confidence interval for a ratio of two bivariate-Normal (BVN) random variables. That is, we will be interested not only in the variance, or, as is more commonly quoted, the standard deviation ("sigma") of the function $z = x / y$, but we also consider the issue of what we are supposed to infer when we see a result like, e.g., $z = 10.5 \pm 3.2$.

What does this statement tell us about z? Presumably we would like to be able to say that there is some given probability (often, 95%) that the true but unknown value of z lies in, say, [4 16]. While this is not what the confidence level of an interval really means, it is how these ranges are usually interpreted. In fact, the probability does not apply to the *parameter* but to the *interval*. The true value of z is not random and it either is in [4 16] or it is not. But if we repeated the experiment many times, and processed the data the same way, then, e.g., 95% of the *intervals* we constructed would contain the true parameter value.

To illustrate this, consider Fig. 1. Here we have plotted 100 of 10000 replicates of a simulation experiment where confidence intervals are constructed on the mean, with each replicate having a sample size of 100. Each replicate is simulating an experiment where we observed data and calculated a confidence interval on a parameter (here, the mean). The several intervals with a circle do *not* include the true population mean, and we would conclude, based on those experiments, that the parameter value is not 100, when in fact it is 100. Out of the 10000 replicates, 9528 did include the true population mean, so the "confidence level" was 0.953, which is in excellent agreement with the desired level of 95% (two-sided). That level was used to find the t-statistic, which in turn is used to create the confidence intervals.

To find a confidence interval we must know the PDF of the variable in question, because we need to know the probability content above or below the points we will be calculating. That is, the upper bound of the interval will be set at that point on the abscissa where the integral of the PDF from that point to positive infinity yields a probability content of, e.g., 0.025. A similar calculation is done for the lower bound.

> *Note that Bevington will not lead to a confidence interval unless Normality is assumed. Bevington only produces an* ***approximation*** *for the variance, which may or may not be particularly close to correct, and it says nothing about the PDF of the variable. These points, along with possible correlations among the x-variables, are usually ignored.*

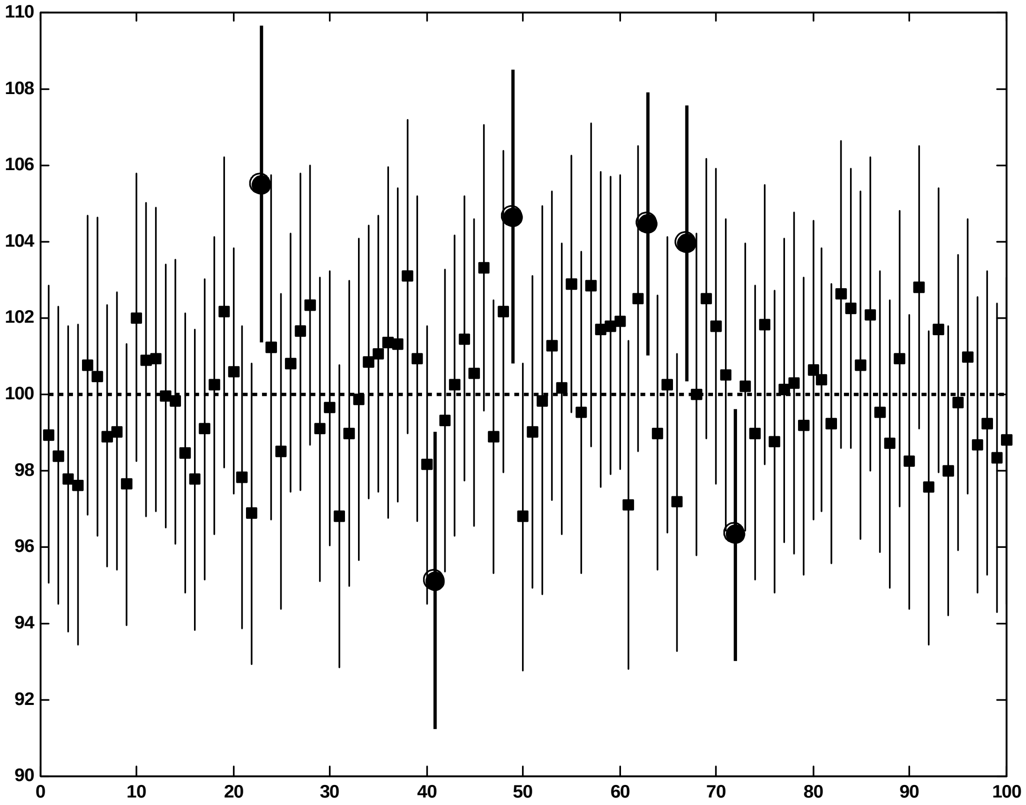

Figure 1. Simulated confidence intervals, N=100 per replicate, 100 replicates of 10000 are shown here; 9528 of the 10000 included the true mean. The intervals with the circles do not include the correct mean.

To explore the relation between the correct calculation and the Bevington approximation, we now consider our example function $z = x / y$, with x, y distributed BVN. It has been shown recently (2002), with some relatively advanced mathematics, that the PDF of z is given by [4]

$$PDF(z) = \frac{k}{\sigma_y^2 z^2 - 2\rho\sigma_x\sigma_y z + \sigma_x^2} \; {}_1\mathrm{F}_1\left[1,\, 0.5,\, \theta(z)\right] \tag{3a}$$

where

$$k = \frac{2\left(1-\rho^2\right)\sigma_x^2\sigma_y^2}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\sigma_y^2\mu_x^2 - 2\rho\sigma_x\sigma_y\mu_x\mu_y + \mu_y^2\sigma_x^2}{2\left(1-\rho^2\right)\sigma_x^2\sigma_y^2}\right] \tag{3b}$$

and[2]

$$\theta(z) = \frac{\left[\sigma_y^2\mu_x z - \rho\sigma_x\sigma_y\left(\mu_y z + \mu_x\right) + \mu_y\sigma_x^2\right]^2}{\sigma_x^2\sigma_y^2\,2\left(1-\rho^2\right)\left(\sigma_y^2 z^2 - 2\rho\sigma_x\sigma_y z + \sigma_x^2\right)} \tag{3c}$$

[2] There is an error in the expression for *θ(z)* in the online document. It is corrected here. Also, the expression for *k* can of course be simplified but it is left in this form to facilitate comparison with the online document.

It is assumed that we have the means and standard deviations for the numerator x and denominator y, and a value for ρ. This PDF calculation was implemented in MATLAB. Note that the function F in (3a) is a confluent hypergeometric function.[3]

To verify that the PDF (3) is correct, correlated data was generated. That is, many thousands of values for x and y were sampled randomly from a BVN distribution, with an input correlation coefficient. This Monte Carlo technique involves the use of the Cholesky decomposition of the BVN covariance matrix; an example of the correlated data is shown in Fig. 2, for a correlation coefficient of 0.8.

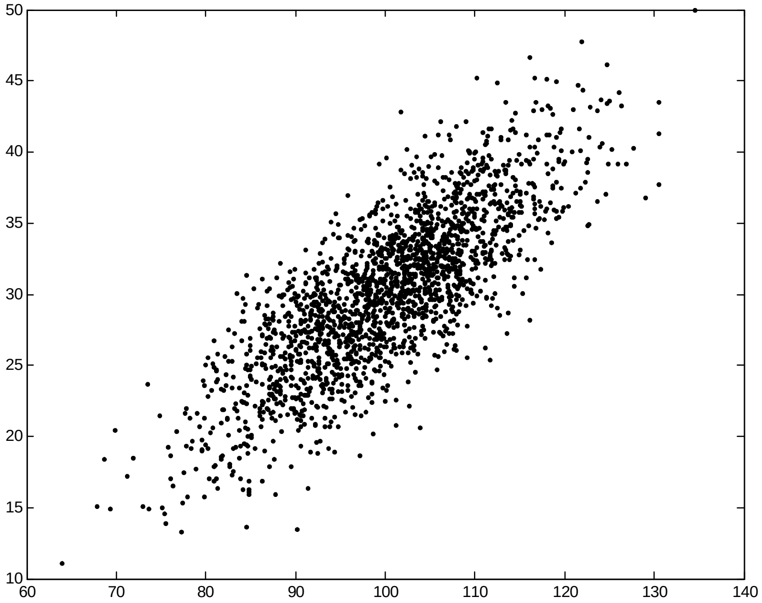

Figure 2. Scatter diagram of BVN data, 1000 points, correlation coefficient = 0.8

With these correlated x, y pairs we can find thousands of z values, which are then histogrammed to produce an empirical PDF. The PDF from (3) is overlaid, and we can compare the results. To this plot is added a Normal PDF using Bevington, (2), both with and without a term for the covariance.

In Fig 3 is a simulation case with 10000 trials where the numerator had a mean of 100 and a “sigma” (standard deviation) of 10, while the denominator mean was 30 with a sigma of 5.48 (the square root of 30). The correlation coefficient was 0.2. The solid line is (3), the dotted line is (2) with the covariance term, and the dashed line is (2) without the covariance term.

We observe that the empirical PDF (histogram[4]) and (3) agree well, and that the PDF is skewed. This would lead to an asymmetric confidence interval. The Bevington approximations are in the general area of correctness, but of course do not show the skewness of the actual data. Next, in Fig. 4, we have a case with the same x and y specifications, but a correlation of 0.8. We see that the PDF (3) and the histogram agree, and that the Bevington with the covariance term is not too far off. However, without that term, (2) is a poor representation of this data.

To complete the analysis we would find a confidence interval for the ratio z. For the upper bound we would need to integrate (3) and use a root-finder to determine the value of z at which the integral of (3) from z to infinity was equal to half the Type I error rate (commonly denoted by α), and similarly for the lower bound.

[3] MATLAB did not have this function, and one was found online, in a library of advanced mathematical functions translated from FORTRAN.

[4] To compare with a true PDF a histogram must be properly normalized. This requires a correction to each bin. MATLAB does not do this, and a corrected-histogram function was found online. (The area under a frequency histogram should be unity.)

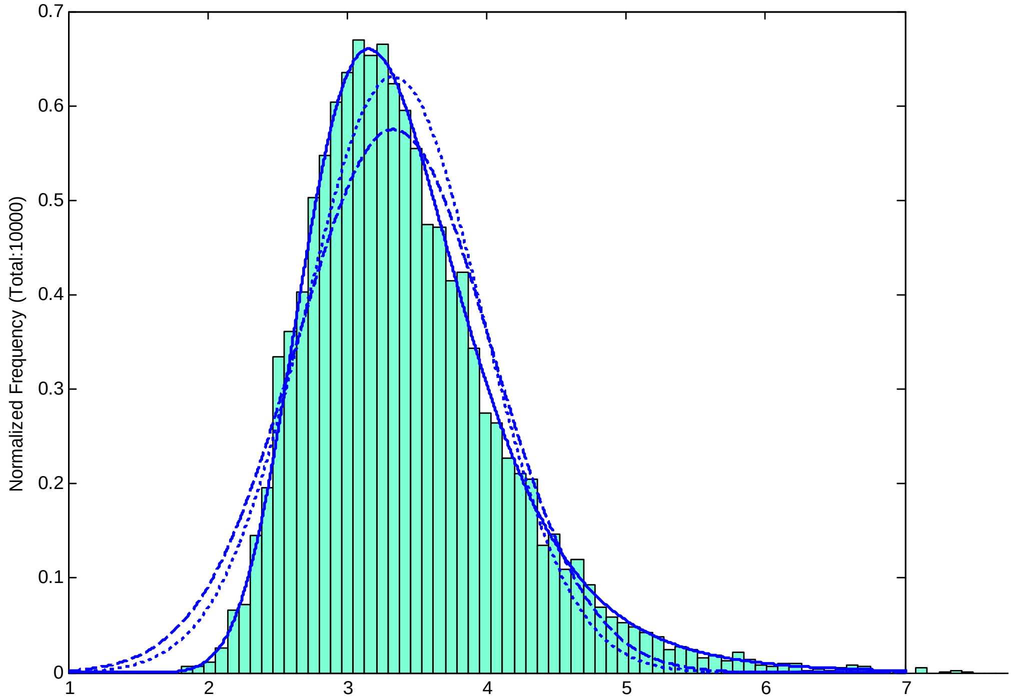

Figure 3. Histogram of simulated BVN ratio data, with correct and approximate PDF's.

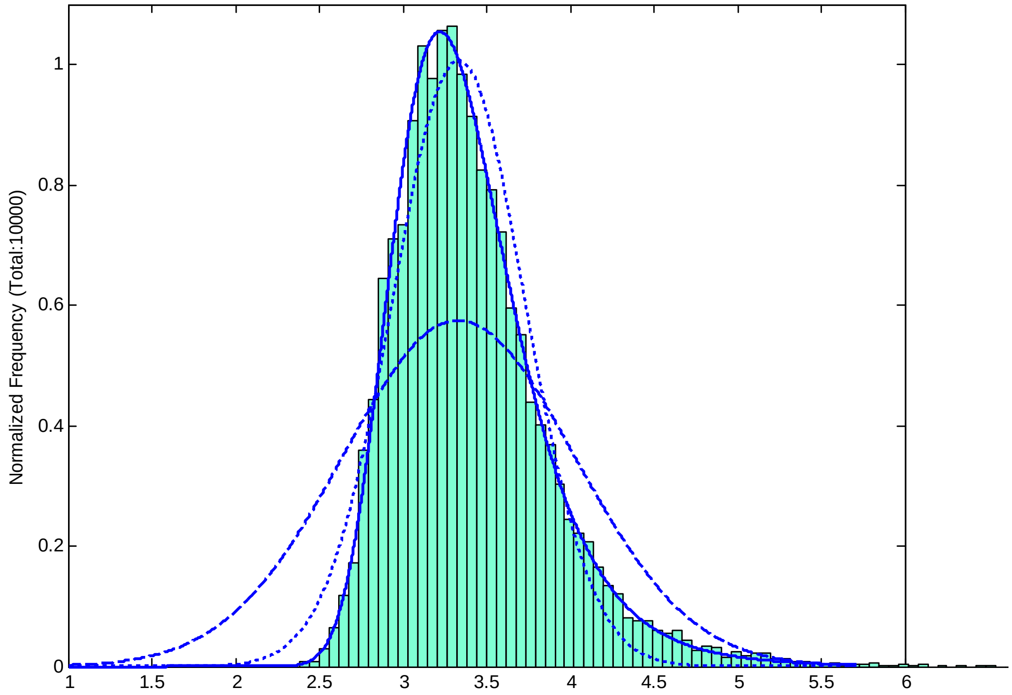

Figure 4. Simulated data and PDF's, correlation coefficient = 0.8

Figure 5 shows the cumulative probability distribution function, obtained by numerically integrating each of the three PDFs in Fig. 4. The abscissas where these CDFs intersect the lines at probabilities of 0.025 and 0.975 provide the lower and upper bounds for a confidence interval on the ratio *z*. We observe that these intersections are quite different, although in this particular example the measurement scale is such that the differences are not numerically large. For the correct PDF (3) we see that a 95% interval on *z* is about [2.7 4.5], which is asymmetric about the mean value of 3.333.

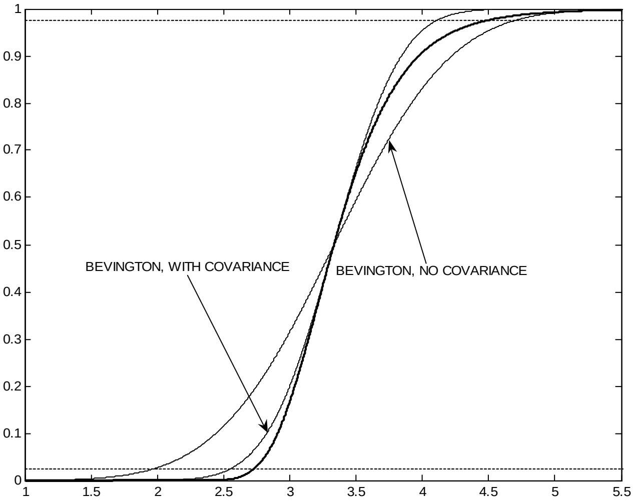

Figure 5. Cumulative probability distribution functions for the PDFs in Fig. 4.

It should also be noted that when the denominator *y* of the ratio *z* becomes close to zero, either in the mean or by virtue of its variation about that mean (i.e., a large "sigma"), the PDF becomes pathological, and, while (3) can still describe it correctly, (2) cannot. In short, the uncertainty for a ratio needs very special attention, and the blind use of (2), as so often happens, is simply not acceptable scientific practice.

Also, more generally, when there are complicated functions of several (possibly correlated) random variables, we sometimes see what might be called "Bevington gone mad" with page after page of algebra purporting to find the uncertainty in a derived quantity. These exercises almost always ignore the covariance. Further, even if this "sigma" could be shown to be correct (simulation studies should always be done to verify this), *we still do not know the PDF of the derived quantity.*

So, how do we set a confidence interval? Apparently, we are just to assume that the derived quantity is Normally-distributed, and take, e.g., twice the "sigma" to construct a 95% interval. This is poor science. *The only statistically-defensible, and practical, way to find the PDF and the corresponding confidence intervals on a derived quantity for nontrivial, nonlinear functions of several random variables is via Monte Carlo simulation*. This is especially important for "limits of detection" types of analysis, where we are interested in the percentage points at the upper and lower tails of possibly highly-skewed PDFs.

REFERENCES

[1] Bevington, P. R. and Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences*, 2nd Ed. (1992) McGraw-Hill, p. 43

[2] Mandel J., *The Statistical Analysis of Experimental Data*, Dover (1984), p. 73

[3] Meyer, S. L. *Data Analysis for Scientists and Engineers*, Wiley (1975), p.39

[4] Pham-Gia, Turkkan, Marchand; "Density of the Ratio of Two Normal Random Variables"
www.usherbrooke.ca/mathematiques/telechargement