Errors [*]

Oscar García

All measurements are subject to error and uncertainty. Error sources are varied, and could be classified in many ways. For instance, there are what we might call "mistakes", due to wrong readings on an instrument scale, transcription errors, etc. There are instrumental errors, due to defects or bad use of an instrument Personal errors, caused by deficiencies in the observer senses, or by subconscious influence of his interests or preferences. Very important and often ignored are errors due to the model; for instance, in most calculations with tree diameters and cross-sections it is assumed that the cross-section is circular. Systematic errors are those that always act in the same direction.

In relation to an instrument or method that generates a (real or hypothetical) series of measurements, it is useful to distinguish between *accuracy* and *precision*. Accuracy refers to the closeness between measurements and the true value. Precision has to do with consistency, closeness of the measurements among themselves. Measurements can be precise but inaccurate. Some authors understand accuracy as the absence of systematic errors ("bias"), closeness of the measurements mean to the true value.

1 Error bounds

In engineering calculations it is common to work with uncertainties or estimated errors assumed to span the true value. That is, a value is given as $x \pm \Delta x$, where x is the estimated value and Δx is a maximum error bounding the true value (it is taken as a positive number, the error absolute value). In other words, by *error* here we understand an error bound.

In particular instances the error in the result of calculations with quantities subject to error can be determined by substituting all possible combinations of negative and positive errors, and taking the extreme results (the combinations to be tried can be reduced if it is clear which are the most

[*] Translated from Appendix A in *Apuntes de Mensura Forestal – Estática*, Universidad Austral de Chile, Facultad de Ciencias Forestales, 1995

unfavorable situations). It is a good idea to do this in important instances. The methods described below are more convenient, and can provide useful relationships between errors and variables.

It is clear that in a sum or difference errors add up, because they are assumed independent and the direction of their action is unknown (for a bound, the most unfavorable situation must be taken):

$$(x \pm \Delta x) + (y \pm \Delta y) = (x + y) \pm (\Delta x + \Delta y)$$

$$(x \pm \Delta x) - (y \pm \Delta y) = (x - y) \pm (\Delta x + \Delta y) \ .$$

Multiplication and division is somewhat more complicated:

$$(x \pm \Delta x)(y \pm \Delta y) = xy \pm x\Delta y \pm y\Delta x \pm \Delta x \Delta y \ .$$

The last term is small relative to the others, and omitting it we can write (assuming that x and y are positive)

$$(x \pm \Delta x)(y \pm \Delta y) = xy \pm xy(\Delta x/x + \Delta y/y) \ .$$

$\Delta x/x$ is the *relative error* for x (while Δx is the *absolute error*). It is seen, then, that the relative error for the product is approximately the sum of the relative errors for the factors. The same happens with division.

More generally, the error for a function of x and y may be approximated by the initial terms of its Taylor series:

$$g(x + \Delta x, y + \Delta y) = g(x, y) + \frac{\partial g(x, y)}{\partial x}\Delta x + \frac{\partial g(x, y)}{\partial y}\Delta y + \dots \ .$$

The omitted terms contain products of errors and, as in the multiplication, can be neglected if the errors are not too large. Considering the uncertainty in the error signs, we find then that in the worst case the error in g is approximately

$$\Delta g = \left|\frac{\partial g(x, y)}{\partial x}\right| \Delta x + \left|\frac{\partial g(x, y)}{\partial y}\right| \Delta y \ .$$

The generalization to any number of variables is obvious.

Let us see two simple examples.

(i) Let $z = g(x, y) = xy$. Then

$$\Delta z = |y|\Delta x + |x|\Delta y \ ,$$

which agrees with the results above.

(ii) The error in the one-variable function $g(x) = \ln x$ is

$$\Delta \ln x = \left|\frac{1}{x}\right| \Delta x = \frac{\Delta x}{x}$$

(x must be positive), so that *the relative error in x is approximately equal to the absolute error in* $\ln x$.

Questions, exercises

1. Use the relationship $\ln xy = \ln x + \ln y$ and the result from example (ii) to obtain the relationship between the relative errors of xy, x and y. Obtain also the relative error of x/y.
2. Calculate the error (bound) for a tree height given the errors in the distance measurement and in the top and base angle measurements.
3. Assume that the height error is dominated by the error in the angle α between the top and the horizontal, and that this error is independent of α (other errors are negligible). Show that the error is a minimum when $\alpha = 45°$.

2 Significant figures

Using significant figures is an alternative to expressing an error as $x \pm \Delta x$. Significant figures are the digits, excluding zeros used only for establishing the position of the decimal point. For instance, the numbers 1302, 0.8206, 0.0002135, 60.60 and 1.490×10^3, all have 4 significant figures. Without further information, it is nor known if 1490 has 3 or 4 significant figures.

The indication of errors through significant figures is not fully standardized. Usually, uncertainty in the last given figure is assumed, with that digit giving an idea of the most likely value (the figures "signify something"). Some authors (e.g. Husch) use an stricter criterion, that the error must not exceed one unit in the last figure. Others accept some uncertainty in the before-last figure. In general, it is considered that it does not make sense to specify more than one or two figures in Δx, and that x should be given up to the digit corresponding to the last figure in the error. 15.04 ± 0.15 is correct, not 15.036 ± 0.15. More figures would suggest false accuracy, less would result in unnecessary accuracy loss.

Anyhow, the number of significant figures reflects the relative error, while a number of decimal places reflects absolute error. The precision indicated by significant figures, or the relative error, are independent of the measurement units: 3.24 m and 324 mm carry the same precision.

These relationships between figures and errors allow us to establish certain rules about the significant figures to be used in results from arithmetical operations. The error in a sum or difference is dominated by the largest absolute error in their components (as seen above, maximum errors add up; other error measures combine with less weight on the smaller errors, as will be seen below). Therefore, a rule is adopted to give the result with a *number of decimal places* equal to the least number of decimal places among the terms added or subtracted:

```
  123
   32.3
+   0.276
  -------
  156
```

In multiplication and division the same happens with the relative errors, so that in the result the least *number of significant figures* among the factors is used:

```
754.1 x 0.052 = 39
```

In the intermediate steps of a calculation sequence it is advisable to retain additional figures, and round the final result.

It is important to take into account that in some operations important losses of significant figures (precision) can occur. This is the case of "catastrophic cancellation" when subtracting large nearly equal numbers.

Questions, exercises

1. Indicate the number of significant figures in: (a) 1.00025 (b) 0.002710 (c) 10.003 (d) 100000

2. In the examples of sum and multiplication just given, assume errors of ±2 units in the last significant figure. Compute the error limits by extreme value substitution. Compare to the significant figures.

3. In an evaluation of silvicultural regimes, incomes of \$3,274,531 and costs of \$3,256,890 are obtained. Compute the expected profit.

(a) Assume now an error of about 1%. Repeat the profit calculation using the appropriate number of significant figures. What can you say about the profitability?

(b) With the 1% errors, obtain error limits by substituting the most optimistic and most pessimistic values.

4. A sample variance can be computed as $\frac{1}{n}\sum(x_i - \bar{x})^2$, where $\bar{x}$ is the mean $\frac{1}{n}\sum x_i$. It is often suggested to simplify calculations by using the formula $\frac{1}{n}\sum x_i^2 - \bar{x}^2$.

   (a) Show that both formulas are mathematically equivalent.

   (b) Compute with both formulas the variance for the three numbers $x_1 = 100001$, $x_2 = 100002$ and $x_3 = 100003$. What happens?

♡ The statistical approach

In calculating error limits we took the most unfavorable situation, with signs for the various errors such that the error in the result is the largest possible. For instance, when adding x to y it is assumed that Δx and Δy act in the same direction, positive or negative, compounding their effects. This is useful because it provides an upper error bound. However, specially with several variables, these limits may be too wide to be useful, and it may seem unrealistic for all errors conspiring to produce the worst possible result. Instead of error limits, it is therefore possible to work with a statistical or probabilistic model of measurement uncertainty.

Statistics deals with the use of information in situations of uncertainty. It uses *Probability Theory*, which deals with the mathematical properties of some uncertainty models.

An uncertain quantity can take any value within a set of possible values. Some values are more plausible than others, so that we give them different weights. These weights might represent relative frequency under repeated observation, a subjective degree of credibility for the various values, etc. In the model we represent the uncertain quantity by a *random variable*, and the weights by a probability. As always, the theory and mathematical manipulation of the model are independent of its interpretation, but obviously this is important when assessing the applicability of the results.

For now, we consider quantities that take on numerical values, so that the weights can be represented by a probability density function defined on the real numbers. The probability for the random variable X to be between a and b is $\int_a^b f(x)\,dx$. Obviously, $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Sometimes it is convenient to distinguish between the random variable X and the observed values x.

PRACTICAL SITUATION		PROBABILISTIC MODEL
Uncertainty in x	$\rightsquigarrow$	X is a random variable
Weighting of possible values	$\rightsquigarrow$	density $f(x)$
Weighted mean of $g(x)$	$\rightsquigarrow$	expected value $E[g(X)]$

The expected value or expectation of a function $g(X)$ is the weighted mean

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \ \ .$$

Important special cases:

$$\text{Mean: } E[X] = \mu$$

$$\text{Variance: } E[(X-\mu)^2] = E[X^2] - (E[X])^2 = \sigma^2 = V[X] \ .$$

The meal is a measure of location, the center of gravity around which uncertainty is distributed. The standard deviation $\sigma = \sqrt{\sigma^2}$ is an important measure of spread.

Back to errors, let us represent a measurement or observed or calculated value by a random variable X, and denote the true value as x_0. The error (another random variable) is $\varepsilon = X - x_0$, that is, $X = x_0 + \varepsilon$. Then, $E[\varepsilon] = \mu_\varepsilon$ is the *bias*. A measure of precision is the *standard error* $\sqrt{V[\varepsilon]} = \sigma_\varepsilon$ (it is common to call standard error to the standard deviation of an estimator). Another measure that combines accuracy and precision is the *mean squared error*: $MSE = \sqrt{E[\varepsilon^2]}$. Note that

$$MSE^2 = \sigma_\varepsilon^2 + \mu_\varepsilon^2 = \text{variance} + \text{bias}^2 \ .$$

The error bound or absolute maximum that we used previously would be (if it exists): $\Delta x = \max|\varepsilon|$, and the relative one, $\Delta x / x_0$ (or $\Delta x/(x_0 + \varepsilon)$ which is almost the same if the error is small).

To study the propagation of errors when computing with variables subject to error (random variables), we need some simple properties of expectations and variances. From its definition as integral it is easily found that expectation is a linear operator:

$$E[aX + bY] = aE[X] + bE[Y] \ .$$

Let us find the variance of a linear combination.

$$\begin{aligned} V[aX+bY] &= E[(aX+bY-E[aX+bY])^2] = E[\{a(X-E[X]) + b(Y-E[Y])\}^2] \\ &= E[a^2(X-E[X])^2 + 2ab(X-E[X])(Y-E[Y]) + b^2(Y-E[Y])^2] \\ &= a^2V[X] + b^2V[Y] + 2abE[(X-E[X])(Y-E[Y])] \ . \end{aligned}$$

The expectation in the last term is the *covariance* between X and Y, $Cov[X,Y]$. Therefore we have

$$V[aX+bY] = a^2V[X] + b^2V[Y] + 2abCov[X,Y] \ .$$

The covariance is related to the *correlation coefficient*

$$\rho = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}} \ ,$$

which is zero if X and Y are independent (more precisely, uncorrelated), and can reach 1 if X and Y tend to vary jointly or -1 if the vary in opposite ways. Finally, note that if a is not random,

$$V[X + a] = V[X] \ .$$

♡♡ The density $f(x)$ that defined the probability for intervals on the x line generalizes to higher-dimensional spaces. For instance, the *joint density* $f(x, y)$ applied to the plane of points specified by coordinate pairs (x, y). (These pairs and their analogs in more dimensions can be seen as lists of numbers, or *vectors*). It is said that the random variables X and Y are independent if their joint density is of the form $f(x, y) = f_1(x)f_2(y)$. A consequence that derives from the definition of expectation as a multiple integral is that if the variables are independent, then $E[XY] = E[X]E[Y]$. It is easily verified that this implies $Cov[X, Y] = 0$. It may be mentioned that zero covariance (uncorrelated variables) does not necessarily imply independence, except in the important case of the Normal distribution.

We are ready now to examine error propagation. Let us see first the addition case.

$$E[\varepsilon_{x+y}] = E[(X + Y) - (x_0 + y_0)] = E[\varepsilon_x + \varepsilon_y] = E[\varepsilon_x] + E[\varepsilon_y] \ ,$$

so that biases add up.

$$V[\varepsilon_{x+y}] = V[\varepsilon_x + \varepsilon_y] = V[\varepsilon_x] + V[\varepsilon_y] + 2Cov[\varepsilon_x, \varepsilon_y] \ .$$

If errors act independently, it is seen that the standard error for the sum is

$$\sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2} \ .$$

Measured this way, the error grows more slowly than the maximum error Δ.

For the general case we use, as before, the Taylor series:

$$\begin{aligned} \varepsilon_g &= g(X, Y) - g(x_0, y_0) = g(x_0 + \varepsilon_x, y_0 + \varepsilon_y) - g(x_0, y_0) \\ &\approx \frac{\partial g(x_0, y_0)}{\partial x_0}\varepsilon_0 + \frac{\partial g(x_0, y_0)}{\partial y_0}\varepsilon_y \ , \end{aligned}$$

Assuming independent errors, we have then approximately

$$\sigma_g^2 = \left(\frac{\partial g(x_0, y_0)}{\partial x_0}\right)^2 \sigma_x^2 + \left(\frac{\partial g(x_0, y_0)}{\partial y_0}\right)^2 \sigma_y^2 \ .$$

In the derivatives we could have used the means or the observed values, instead of the actual values x_0 e y_0. The approximations would still be valid, provided that the errors are not too large.

Let us use this to calculate the standard error for a logarithm:

$$\sigma^2_{\ln x} = (1/x_0)^2 \sigma^2_x \; .$$

Using the mean instead of x_0,

$$\sigma_{\ln x} = \sigma_x / \mu_x \; .$$

The expression in the right-hand-side is the coefficient of variation (CV) for x.

Questions, exercises

1. Obtain an expression for the coefficient of variation of the product of two independent variables X and Y as a function of the coefficients of variation of the factors.
2. For the previous problem, graph $\mathrm{CV}(XY)/\mathrm{CV}(X)$ over $\mathrm{CV}(Y)/\mathrm{CV}(X)$ for $\mathrm{CV}(X) > \mathrm{CV}(Y)$. What effect have the smaller and larger errors on the error of the result? Implications for model building?