General statement of the problem

If some quantity of interest z is calculated from other quantities x, y, q, w, etc.,

$$z = f(x, y, q, w, \cdots)$$

what is the uncertainty in z, and how is it related to the uncertainties in x, y, q, w, etc? In other words, given the function f above, what is the corresponding function g

$$s_{\bar{z}} = g(s_{\bar{x}}, s_{\bar{y}}, s_{\bar{q}}, s_{\bar{w}}, \cdots)$$

that permits us to compute the uncertainty in z from the uncertainties in its component parts? Note that we could estimate the uncertainty in the average value of z (the *standard error*) from the standard errors of the component means (as in the formula above). We could also calculate the uncertainty (or variability) in individual estimates of z (the *standard deviation*) using the same function g, but in this case the inputs are the standard deviations of the input variables x, y, q, and w, rather than the standard errors of their means:

$$s_z = g(s_x, s_y, s_q, s_w, \cdots)$$

Sources of uncertainty

1. statistical error/random variation of replicate measurements
2. spatial and temporal variability
3. systematic error (bias)
4. imprecise definitions or unrepresentativeness of samples
5. uncertainty in the form of the function relating z to x, y, q, w, etc.

This toolkit explains methods for quantifying uncertainty that arises from random measurement error and from spatial or temporal variability (where one wants to average over that variability). Uncertainty arising from sources (3)-(5) *is not adequately addressed by these methods* (or by any other general techniques either).

Ways to express uncertainty or variability

Variance

Advantages: -easy to manipulate mathematically (see below)
Disadvantages: -measured in units of x squared, thus cannot be compared directly to values of x.

$$Var(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation and standard error

Advantages: -intelligible directly in units of x
Disadvantages: -harder than variance to manipulate mathematically

$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ is the standard deviation of the individual x's,

$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$ is the standard error of the mean ($\bar{x}$).

The term *standard deviation* is usually reserved for expressing the spread of values of individual observations x_i of a variable x. The term *standard error* is most often used to express the uncertainty in the mean of x, but it is also more generally applied to express the uncertainty associated with any form of a central estimate. Thus one can speak of the standard error of a quantity q, whether or not q is the mean of a set of measurements. The standard error is denoted by several different symbols, including $s_{\bar{x}}$, $\sigma_{\bar{x}}$, SE(x), and s.e.(x).

Confidence intervals (for mean of x) and prediction intervals (for individual values of x)

Advantages: -associates an explicit degree of confidence with a specified interval of variability.
Disadvantages: -more difficult to 'propagate' through a calculation than standard error is.

Probability distribution p(x)

Advantages: -conveys much more information about variability in x (skew, shape of distribution, characteristics of tails, etc.) than a single parameter of spread, such as variance or standard deviation or standard error.
Disadvantages: -without huge numbers of measurements, or without good *a priori* reasons to believe that p(x) conforms to one of the well-known distributions (normal, Poisson, beta, gamma, etc.), accurately estimating the shape of p(x) is difficult (the tails are particularly unstable).

How to report uncertainty or variability

Most commonly, quantities are reported as the mean (or central estimate), plus-or-minus the standard error, like this: q=1.23±0.45. That means the central estimate of q is 1.23, and the standard error of the estimate is 0.45. *You should generally use this convention for reporting results in this course.* The central estimate of q may be a mean of sample measurements, or it may be derived in other ways. Sometimes the standard error is given in parentheses, e.g. 1.23(0.45).

Warning: quantities are sometimes reported as means±*standard deviations* or ±*confidence intervals* rather than standard *errors*. Where the number of measurements is known, one can convert from standard deviations to standard errors (and back again) by scaling by $\sqrt{n}$. Sometimes the number of measurements is reported in parentheses, e.g. q=1.23±1.10 (6). There are many diverse conventions; to prevent confusion, you should always make clear what your measure of variability is, e.g.: "the average concentration was 1.23±0.46 ppb_{wt} (mean±standard error)".

Important preliminary notes

1: The formulas in this toolkit will be given in terms of the sample standard error $s_{\bar{x}}$, rather than the population standard error $\sigma_{\bar{x}}$, because uncertainty is commonly estimated from the standard deviations of samples, rather than from *a priori* knowledge of the population standard deviation. The formulas are identical in either case; merely substitute the greek symbols for the equivalent formulas in terms of μ and σ.

2: The "uncertainty in z" can refer to two different, but related, things: the variability of individual measurements of z (usually termed a standard deviation), and the uncertainty in a central estimate of z (usually termed a standard error). Likewise the variable z can be used to refer to an individual measurement, or to the central value for which the sample mean is an approximation. *The formulas below will work fine in either case, as long as you don't mix them up*: you can apply them to the variability in individual measurements (s_x), or to the uncertainty in central estimates ($s_{\bar{x}}$), but you can't mix both kinds of uncertainties in the same equation. Whether you are propagating the uncertainty in individual values of z, or the uncertainty in the central estimate of z, the *equations* are exactly the same, but the numerical *values* you use for s_x will be different (smaller, by $1/\sqrt{n}$ when x represents the mean of n independent measurements).

3: None of the calculations below assume that variables are normally distributed. You can always use $s_{\bar{z}}$ as an estimate of the uncertainty in z, and the equations below will be equally accurate in estimating $s_{\bar{z}}$, whether or not z is normally distributed. Of course, if z is normal, the $s_{\bar{z}}$ will be a *complete* description of the uncertainty in z, whereas if z is not normal, $s_{\bar{z}}$ will be be equally accurate, but you will also need to know the particular *shape* of the distribution in order to describe the variability in z.

4: At the risk of stating the obvious: the symbol z used here is not the "Z statistic" (a.k.a. the standard normal deviate). It is simply some variable that is a function of other variables x, y, q, etc.

5. Only use the "simple rules" given below for the simple functions to which they apply. For example, *don't* use the Simple Rule for Products and Ratios for a power function (such as $z=x^2$), since the two x's in the formula would be correlated with each other. When in doubt, use the method of moments (which incorporates these simple rules as a special case).

Simple rule for sums and differences

<u>if</u> $z=x\pm y\pm q\pm w\pm...$
<u>and</u> x, y, q, w, etc. are uncorrelated with one another
then the variance of the sum or difference is the sum of the variances:

$$Var(z) = Var(x) + Var(y) + Var(q) + Var(w) + \cdots$$

or, equivalently, the standard errors add *in quadrature* (that is, squared, added, and then square rooted).

$$s_{\bar{z}} = \sqrt{(s_{\bar{x}})^2 + (s_{\bar{y}})^2 + (s_{\bar{q}})^2 + (s_{\bar{w}})^2 + \cdots}$$

Important note: The uncertainty in each variable *increases* the uncertainty in z, whether the variable was added or subtracted in calculating z. As more variables are included in z, z may either grow or shrink (depending on whether the new variables add or subtract), but the uncertainty in z *always increases*.

Important note 2: Therefore beware the small difference between two or more large numbers. The percentage uncertainty in $z = x - y$ can be *very large*, even if the percentage uncertainties in x and y are very small.

Simple rule for weighted sums and differences

if $z = ax \pm by \pm cq \pm dw \pm \ldots$
and $a, b, c, d,$ etc. are constants (with no uncertainty)
and $x, y, q, w,$ etc. are uncorrelated with one another
then the variance of the weighted sum or difference is the weighted sum of the variances:

$$Var(z) = a^2 Var(x) + b^2 Var(y) + c^2 Var(q) + d^2 Var(w) + \cdots$$

or, equivalently, the standard errors add as a weighted sum *in quadrature* (that is, squared, weighted, added, and then square rooted).

$$s_{\bar{z}} = \sqrt{(a \cdot s_{\bar{x}})^2 + (b \cdot s_{\bar{y}})^2 + (c \cdot s_{\bar{q}})^2 + (d \cdot s_{\bar{w}})^2 + \cdots}$$

Important note: the uncertainty in each variable *increases* the uncertainty in z, whether the variable was added or subtracted to make z.

Simple rule for products and ratios

if $z = x$ times/divided by y times/divided by q times/divided by w times/divided by ...
and $x, y, q, w,$ etc. are uncorrelated with one another
then the percent (or fraction) standard error of z can be found by adding the percent (or fraction) standard error in each of its components, in quadrature:

$$\frac{s_{\bar{z}}}{\bar{z}} = \sqrt{\left(\frac{s_{\bar{x}}}{\bar{x}}\right)^2 + \left(\frac{s_{\bar{y}}}{\bar{y}}\right)^2 + \left(\frac{s_{\bar{q}}}{\bar{q}}\right)^2 + \left(\frac{s_{\bar{w}}}{\bar{w}}\right)^2 + \cdots}$$

Gaussian error propagation

if $z = f(x, y, q, w, \cdots)$, where f can be a nonlinear function,
and $x, y, q, w,$ etc. are uncorrelated with one another
then the standard error of z can be approximated by the *Gaussian error propagation rule*:

$$s_{\bar{z}} \approx \sqrt{\left(\frac{\partial z}{\partial x} s_{\bar{x}}\right)^2 + \left(\frac{\partial z}{\partial y} s_{\bar{y}}\right)^2 + \left(\frac{\partial z}{\partial q} s_{\bar{q}}\right)^2 + \left(\frac{\partial z}{\partial w} s_{\bar{w}}\right)^2 + \cdots}$$

where $\partial z / \partial x$, etc. are the partial derivatives of z with respect to its component variables. The Gaussian error propagation rule is a special case of the more

complete *method of moments*, detailed below; in turn, the "simple rules" outlined above are all special cases of Gaussian error propagation.

Method of moments

Advantages: -more general than the simple rules given above
-computationally simpler than exact analytic methods and Monte Carlo methods
Disadvantages: -less accurate than exact methods and Monte Carlo methods, particularly for functions with substantial nonlinearities within the range of uncertainty in the inputs

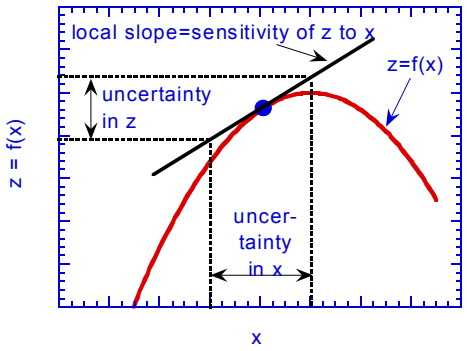
The method of moments is a very general technique for estimating the moments of z--its first moment (mean), its second moment (variance or standard deviation), its third moment (skewness)--based on various approximations to the function f. The most commonly used form, called *first-order second-moment* uncertainty analysis, estimates the *second moment* of z (its spread), based on a *first-order* approximation to f (that is, approximating the function f by a flat plane tangent to the curved surface of f at the mean x, y, etc.). Higher-order approximations are possible--Ang and Tang (reference below) give an example--but they are rarely used because (a) they are much more complex to calculate, and (b) they require higher moments (skewness, kurtosis, etc.) of the input variables, which are difficult to estimate reliably unless sample sizes are large. The mathematical derivations underlying first-order, second-moment uncertainty analysis are given in an appendix to this toolkit.

Simplest case: single-variable function $z=f(x)$

Approximate f by the tangent line (which has a slope of dz/dx) at the mean of x. Then the standard error of z is approximately

$$s_{\bar{z}} \approx \left|\frac{dz}{dx}\right| s_{\bar{x}}$$

Note that the uncertainty in z depends on two things: how uncertain x is (that is, $s_{\bar{x}}$), and how sensitive z is to x (that is, dz/dx). See diagram below:



Function of two variables $z=f(x,y)$

Approximate f by the tangent plane whose slope in the x and y dimensions is described by the partial derivatives $\partial z / \partial x$ and $\partial z / \partial y$ (again, these are evaluated at the mean x and mean y). The variance of z is,

$$Var(z) \approx \left(\frac{\partial z}{\partial x}\right)^2 Var(x) + \left(\frac{\partial z}{\partial y}\right)^2 Var(y) + 2\frac{\partial z}{\partial x}\frac{\partial z}{\partial y} Cov(x,y)$$

where Cov(x,y) is the covariance of x and y, defined as:

$$Cov(x,y) = s_{xy}^2 = r_{xy} s_x s_y$$

where r_{xy} is the correlation coefficient of the relationship between x and y. The covariance of x and y is positive if they tend to vary together (that is, if x is usually high when y is high, and vice versa). The covariance is negative if they vary in opposite directions, and it is zero if they are uncorrelated with one another. [Note that because the correlation of any variable with itself is perfect (r=1), the covariance of any variable with itself is simply its variance: $Cov(x,x)=r_{xx}\, s_x\, s_x = s_x^2 = Var(x)$]. From the formula above, one can see directly that the standard error of z is:

$$s_{\bar{z}} \approx \sqrt{\left(\frac{\partial z}{\partial x} s_{\bar{x}}\right)^2 + \left(\frac{\partial z}{\partial y} s_{\bar{y}}\right)^2 + 2 r_{xy} \left(\frac{\partial z}{\partial x} s_{\bar{x}}\right)\left(\frac{\partial z}{\partial y} s_{\bar{y}}\right)}$$

Note: The uncertainty in z will depend on three things: (1) how much z changes for a given change in x and y ($\partial z / \partial x$ and $\partial z / \partial y$), (2) how uncertain x and y are ($s_{\bar{x}}$ and $s_{\bar{y}}$), and (3) how closely x and y are correlated (r_{xy}).

Note: If x and y are uncorrelated, the third term vanishes and the formula becomes Gaussian error propagation (which in turn is analogous to the simple rule for weighted sums, where the weighting constants are the partial derivatives).

Note: Correlations between x and y can either raise or lower the uncertainty in z, depending on whether the product $r_{xy}\, \partial z / \partial x\, \partial z / \partial y$ is positive or negative. If x and y are positively correlated and have similar effects on z (i.e., $\partial z / \partial x$ and $\partial z / \partial y$ have the same sign), or if they are negatively correlated and have offsetting effects on z ($\partial z / \partial x$ and $\partial z / \partial y$ have opposite sign), then the overall uncertainty in z will be greater than if they were not correlated. Conversely, if x and y are negatively correlated and have similar effects on z or are positively correlated but have offsetting effects on z , their uncertainties will tend to cancel one another out, resulting in a lower overall uncertainty in z than if x and y were uncorrelated.

Note: *What do we mean by correlations here?* We mean correlations in the *uncertainties* in the variables. These can arise from two different sources.

First, the *measurement errors* may be correlated even if the underlying true values are not. For example, evaporation from a water sample will increase the concentrations of all the solutes (and thus create a correlation in the measured concentrations), even if the concentrations in the stream are not correlated.

Second, when the *underlying values* are correlated across a *population*, then the *uncertainties in the group averages* will be correlated, even if the uncertainties in the individual values are not. For example, heart attack risk is a function of height, weight, and cholesterol levels. Height and weight are correlated among individuals, as are weight and cholesterol levels. Thus the uncertainties in the *average* height, weight, and cholesterol levels of a *group* of people will be correlated because of sampling error. That is, if the mean weight of the sampled group is greater than the true mean for the population, the mean heights and cholesterol levels will also probably be higher than the true means, and the correlation would need to be taken into account in assessing uncertainty in the *average* heart attack risk for a group. Thus error propagations for uncertainties in group averages need

to account for both kinds of correlations. Fortunately, a scatterplot of the individual measured values will show the combined effects of both correlations.

But if we are trying to estimate the uncertainty in the heart attack risk for a *single individual*, then only the first kind of correlation matters. An individual's height, weight, and cholesterol are correlated with one another, but the uncertainties are not. Since these measurements are made independently, the measurement errors per se are uncorrelated.

Function of many variables $z=f(x_1, x_2, x_3...)$

Caution: different notation. Here $x_1, x_2, x_3...x_m$ to refer to m different variables (e.g., $x_2=y$, $x_3=q$, etc.) rather than different measurements of a single variable.

Approximate f by the tangent m-dimensional plane whose slope in each of the j=1..m dimensions is described by the partial derivatives $\partial z / \partial x_j$, evaluated at the mean x_j's. The standard error of z is a direct extension of the two-variable case considered above:

$$s_{\bar{z}} \approx \sqrt{\sum_{j=1}^{m}\left(\frac{\partial z}{\partial x_j} s_{\bar{x}_j}\right)^2 + 2\sum_{j=1}^{m}\sum_{k=j+1}^{m} r_{x_j x_k}\left(\frac{\partial z}{\partial x_j} s_{\bar{x}_j}\right)\left(\frac{\partial z}{\partial x_k} s_{\bar{x}_k}\right)}$$

As above, the uncertainty in z will depend on the uncertainty in each of the x_j, the sensitivity of z to changes in the x_j, and the interrelationships among each pair of x_j.

Where the x variables are uncorrelated with one another, this reduces to the familiar *Gaussian error propagation* formula:

$$s_{\bar{z}} \approx \sqrt{\sum_{j=1}^{m}\left(\frac{\partial z}{\partial x_j} s_{\bar{x}_j}\right)^2}$$

Exact analytic methods

Advantages: -permit direct calculation of the full probability distribution, not just one or two moments

Disadvantages: -often analytically intractable

If $z=f(x)$, then the probability that x lies within an interval of width dx must equal the probability that z lies within the corresponding interval dz, such that,

$$p(z)|dz| = p(x)|dx| \quad \text{or} \quad p(z) = p(x)\left|\frac{dx}{dz}\right| = \frac{p(x)}{\left|\frac{dz}{dx}\right|}$$

thus if the probability density function p(x) is known, the corresponding p(z) can be derived directly. When z is a function of more than one variable, p(z) depends on the joint probability distribution of all the input variables. See Ang and Tang (and references therein) for this and other complicated examples.

Monte Carlo methods

Advantages: -permit direct computation of uncertainty even when the function *f* is ill behaved (e.g., contains discontinuities or extreme nonlinearities) and when the input variables may not be readily described by the usual moments (e.g., a binary variable that has only values of 0 and 1, and therefore lacks a central tendency)

Disadvantages: -time-consuming to program
-difficult to infer role of each variable in contributing to total uncertainty
-difficult to document and explain to others

Monte Carlo methods are computer-based techniques for brute force numerical simulation of probabilistic processses, and can be summarized in the following steps.

1: Generate a large number of sets of random numbers that have statistical properties similar to those of the real-world variables x, y, q, w, etc. Relevant properties of these random numbers include: central tendency, spread, shape of distribution, and (importantly) correlation between variables. Algorithms for generating random variables from many different distributions are given in *Numerical Recipes* (reference below). Morgan and Henrion (section 8.5.7) describe approaches to generating correlated random variables.

1a: If there are actual measurements describing x, y, etc., one can randomly pick among the measured values, as an approximation to the (usually unknown) distributions underlying the particular values that were measured. Monte Carlo exercises that use measurements in this way are often called "bootstrap" or "resampling" methods.

2: From each set of values for the input variables x, y, q, w, etc., calculate the corresponding value of $z = f(x, y, q, w, \cdots)$.

3: Examine the distribution of the simulated values of z, to determine the moments and any other features of interest (e.g., range of outlying values). The contribution of each variable to the total uncertainty in z can be estimated from the correlation between that variable and z.

4: Repeat steps 1-3 several times to test whether the conclusions that were drawn are sensitive to the particular random values that have been generated. If so, increase the number of random input sets generated in step 1. Morgan and Henrion (Chapter 8) use statistical sampling theory to estimate how many Monte Carlo samples should be sufficient to achieve the desired level of precision.

References: Morgan, M. G. and M. Henrion, *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*, 332 pp., Cambridge University Press, 1990 (Chapter 8).

Taylor, J. R., *An Introduction to Error Analysis*, 270 pp., University Science Books, 1982.

Bevington, P. R., Data Reduction and Error Analysis for the Physical Sciences, 336 pp., McGraw-Hill, 1969 (Chapters 4 and 5).

Ang, A. H.-S. and W. H. Tang, *Probability Concepts in Engineering Planning and Design*, 409 pp., John Wiley and Sons, 1975 (Chapter 4).

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes: the Art of Scientific Computing*, 818 pp., Cambridge University Press, 1986.