RADIOCARBON, Vol 49, Nr 2, 2007, p 427–440

ERROR AND UNCERTAINTY IN RADIOCARBON MEASUREMENTS

E Marian Scott[1] • Gordon T Cook[2] • Philip Naysmith[2]

ABSTRACT. All measurement is subject to error, which creates uncertainty. Every time that an analytical radiocarbon measurement is repeated under identical conditions on an identical sample (even if this were possible), a different result is obtained. However, laboratories typically make only 1 measurement on a sample, but they are still able to provide an estimate of the analytical uncertainty that reflects the range of values (or the spread) in results that would have been obtained were the measurement to be repeated many times under *identical* conditions. For a single measured ^{14}C age, the commonly quoted error is based on counting statistics and is used to determine the uncertainty associated with the ^{14}C age. The quoted error will include components due to other laboratory corrections and is assumed to represent the spread we would see were we able to repeat the measurement many times.

Accuracy and precision in ^{14}C dating are much desired properties. Accuracy of the measurement refers to the deviation (difference) of the measured value from the true value (or sometimes expected or consensus value), while precision refers to the variation (expected or observed) in a series of replicate measurements. Quality assurance and experimental assessment of these properties occupy much laboratory time through measurement of standards (primary and secondary), reference materials, and participation in interlaboratory trials. This paper introduces some of the most important terms commonly used in ^{14}C dating and explains, through some simple examples, their interpretation.

INTRODUCTION

> *"When you can measure what you are speaking about and express it in a number, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science ..."* Lord Kelvin, 1893.

When a radiocarbon measurement is made, a complex process involving chemistry and physics and the use of complex instrumentation must be gone through. The laboratory provides the user with an estimate (the measured value) of the sample's true ^{14}C age. However, every measurement is subject to error, where the error is defined simply to be the difference between the measured value and the true value. The true value is often unknown, so the error itself is also unknown. How does the laboratory calculate an estimated or quoted error and what does the quoted error published in the age report represent? How is a user to identify "good" measurements?

Users of ^{14}C ages should be aware of the practice in calculating routinely quoted errors; although some details may be specific to the laboratory, the principles are constant. A better understanding of errors will aid in the interpretation of ^{14}C ages. Users need also to be aware of the laboratory efforts to achieve both accurate and precise measurements. Taken together, users need to understand the language of determinations so that they make valid interpretations (and have realizable expectations) of ^{14}C results.

The measurement issues that concern dating are not unique, and indeed there is a scientific discipline of metrology, which is the science of measurement in any application field. International and national guidance documents including those prepared by the Royal Society of Chemistry (RSC), Analytical Methods Committee (UK) (RSC 2003a,b) and the National Institute of Standards and Technology (USA) (Taylor and Kuyatt 1994), are useful background reference material.

[1]Department of Statistics, University of Glasgow, Glasgow G12 8QW, United Kingdom. Corresponding author. Email: marian@stats.gla.ac.uk.
[2]SUERC, Scottish Enterprise Technology Park, East Kilbride, United Kingdom.

Proceedings of the 19th International ^{14}C Conference, edited by C Bronk Ramsey and TFG Higham
RADIOCARBON, Vol 49, Nr 2, 2007, p 427–440

In the following text, the word "sample" will be used in several senses. In a statistical sense, a sample is a set of objects "representative of a population." A sample may also be in the archaeological sense, a single bone, a piece of charcoal, etc. The context of the discussion should make clear the usage. Errors and uncertainties will also be described.

THE ^{14}C AGE MODEL AND ITS UNCERTAINTY

The conventional age is reported as $t \pm s_t$ years BP (yr BP hereafter). s_t represents the analytical uncertainty on the measurement, as estimated by the laboratory (called the quoted error). How is it calculated? A number of authors have written about the calculation of ^{14}C dates and their associated errors (Stuiver and Polach 1977; Donahue et al. 1990; Mook and van der Plicht 1999; Cook and van der Plicht 2007).

The analytical estimate s_t is provided by the laboratory and will typically take into account the analytical uncertainties on the individual components that make up the conventional age. In its simplest form, the quoted error is based on Equation 1 (below) for ^{14}C age calculation. The quoted error on the calculated age, t, is calculated through propagation of the uncertainties on the measurements of A_0 and A_t.

The most basic form of the radiometric ^{14}C age equation is:

$$t = \frac{1}{\lambda}\ln\left(\frac{A_0}{A_t}\right) \tag{1}$$

where t = sample age, which is the time that has elapsed since removal of the sample material from the carbon cycle (e.g. death of an organism); λ = decay constant ($\ln 2/t_{½}$) where $t_{½}$ = 5568 yr (Libby half-life); A_t = the activity of the sample material t years after death (measured in the laboratory); and A_0 = "modern equilibrium living activity" of the sample, estimated in the laboratory using the primary standards of NIST OxI and/or OxII.

The quoted error on t will therefore include terms based on the counting of primary standards, background samples, the unknown-age sample itself, fractionation, etc. The actual quoted error on t (as a function of A_0 and A_t) is derived through an error propagation formula (Bevington and Robinson 2003). If we denote $\sigma(A_t)$ and $\sigma(A_0)$ as the errors on A_t and A_0, then:

$$\sigma(t) = \frac{1}{\lambda}\sqrt{\frac{\sigma(A_t)^2}{A_t^2} + \frac{\sigma(A_0)^2}{A_0^2}} \tag{2}$$

$\sigma(A_t)$ and $\sigma(A_0)$ are calculated from the counts recorded for the unknown-age sample, the primary standard, and the background. (Often in radioactive decay, the assumed model is a Poisson stochastic model, and the uncertainty on the counts is the square root of the counts.)

MEASUREMENT UNCERTAINTY AND ERROR

The error of a measurement, e, is the result of a measurement minus the true value. It is a single value. An error can be positive or negative, and more importantly is classified as being either random or systematic. It should always be reported with the measurement. Therefore, more formally, the routinely quoted error is in fact a measurement uncertainty.

Every time that an analytical ^{14}C measurement is repeated under identical conditions on an identical sample (even if this were possible), a slightly different result is obtained. This scatter of results illus-

trates the effects of small errors. If we repeat a measurement (e.g. on 10 subsamples taken from a bulk cellulose sample) under as near-identical conditions as are practicable, then we will most likely obtain 10 different values (any results that are identical will have occurred by chance). Each measurement is affected by small but uncontrollable changes in the measurement conditions or in the source material itself (and, of course, radioactive decay is a random process). Such variation in values is interpreted as the effect of small but random errors, which themselves are varying. The variation in the set of replicate measurements provides the means to calculate the measurement uncertainty. In contrast, a systematic error is one that remains constant over the repeated measurements. Systematic errors cannot be reduced by making multiple measurements. Occasionally, if the source of the problem can be identified, the systematic error can be corrected.

Table 1 shows 10 replicate measurements made in a single laboratory of a bulk cellulose sample derived from dendrochronologically dated wood provided by Queen's University, Belfast (FIRI, Scott 2003). Each measurement is subject to small, random errors. The replicate values range from 4442 to 4542 (a range of 100 yr), while quoted errors on an individual measurement range from 16 to 30 yr.

Table 1 A series of 10 replicate measurements made on a bulk dendrochronologically dated wood sample.

^{14}C age (BP)	4483	4442	4509	4511	4519	4482	4542	4540	4494	4522
Error (yr)	22	17	20	16	16	16	30	30	21	24

The error of a measurement may include a random component and a systematic component, but the combined effect of errors is to produce an uncertainty, and we can apply established statistical methods to calculate the uncertainty. A measurement uncertainty[3] defines a range of values, often expressed as the interval "measurement ± e" (sometimes described as the 1-σ uncertainty) or "measurement ±2 e" (sometimes described as the 2-σ uncertainty). The actual meaning of such ranges (in terms of confidence or "plausibility") depends on the stochastic model (or probability distribution) chosen to model the random error components. Thus, for 1 measurement from Table 1, namely 4509 BP with a measurement error of 20 yr, the measurement uncertainty or the range of values at 2 σ would be 4509 ± 40 yr or 4469–4549 yr BP. We would say that the true age is highly likely to lie within the measurement uncertainty or within the range.

Replicate Measurements

Making replicate measurements and averaging them will cancel out some of the random errors; therefore, the average will provide a better (less uncertain) estimate of the true age.

The dispersion of experimental data is characterized numerically by the standard deviation, s, and from this standard deviation we can calculate a range of values within which we are confident that the true ^{14}C age value lies. We can estimate the true age as the mean or average of the data and its uncertainty (based on the standard deviation). Commonly, the true age is denoted by μ.

Thus, from the n ^{14}C measurements, labeled X_1 to X_n, the sample mean, $\bar{X}$, is given by

[3]In standard metrology terminology, there are 2 classes of uncertainty, Type A and Type B. Type A can be handled statistically and the scatter is the measure of uncertainty in a sequence of repeated measurements. Type B uncertainties are often pre-specified (e.g. for specific physical properties or the range of values associated with a consensus value for a reference material). When combined with other uncertainties, they are combined statistically.

$$\bar{X} = \frac{1}{n}\sum X_i \tag{3}$$

and the standard deviation, s, is given by

$$s = \sqrt{\frac{\sum_1^n (X_i - \bar{X})^2}{n-1}} \tag{4}$$

For the 10 measurements in Table 1, the average age is 4504 yr BP and the standard deviation is 30 yr. The standard deviation is larger in some cases than the quoted error for individual measurements; thus, a laboratory might in this case choose to quote a minimum routine error of 30 yr on samples of about 1 half-life.

It is worth introducing another commonly used term, namely the standard error of the mean (SEM). It is a measure of precision associated with the estimate of $\hat{\mu}$ and is calculated as

$$\text{SEM}\ (\hat{\mu}) = \frac{s}{\sqrt{n}} \tag{5}$$

distinguishing it from the standard deviation.

So, in the example above, the best estimate of the true age is 4504 yr BP and the standard error on the mean is $(30/\sqrt{10})$ or about 10 yr, so the uncertainty on the true age at the 2-σ level based on the mean is 4504 ± 20 yr or 4484–4524 yr BP.

Figure 1 shows a histogram of a series of 27 measurements on the same Belfast cellulose made in a single laboratory over a period of several months. The mean or average of the series is 4497 yr BP. The standard deviation of the series is 30.2 yr, so the mean on this series is slightly different (4497 vs. 4504) to that from Table 1 and the error on the mean is also smaller at $(30.2/\sqrt{27})$, or 6 yr. Thus, the uncertainty (at 2 σ) on the true ^{14}C age is 4497 ± 12 yr or 4485–4509 yr BP from this experiment. It is clear that as n increases, the SEM decreases.

Effect of Extreme (Outlier) Values

What is an outlier? An outlier is an unusual observation, either extremely small or extremely large. The formal definition of what constitutes an outlier depends on the probability model used to describe the distribution of result. One statistical (and relatively simple numerical) definition is based on the quartiles and the interquartile range. From a distribution of results, it is possible to define several numerical quantities including the median and quartiles. The median is the "middle" value (e.g. 50% of the data lie below the median value). Two quartiles are commonly calculated: Q_1 is the lower quartile; it is the numerical value below which 25% of the data lie. Q_3 is the upper quartile; it is the numerical value above which 25% of the data lie. Outlying results can be defined as those values that are greater than 3 interquartile ranges from the nearest of either the lower or upper quartiles—that is, when a ^{14}C age is either greater than $Q_3 + 3(Q_3–Q_1)$ or less than $Q_1 – 3(Q_3–Q_1)$. Using the data in Table 1, the median is 4510 (half-way between 4509 and 4511), Q_1 is 4483, and Q_3 is 4522.

If there is an outlier in a set of replicate results, then it will affect the numerical value of both the mean and the standard deviation. Thus, both summary statistics are described as non-robust,

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

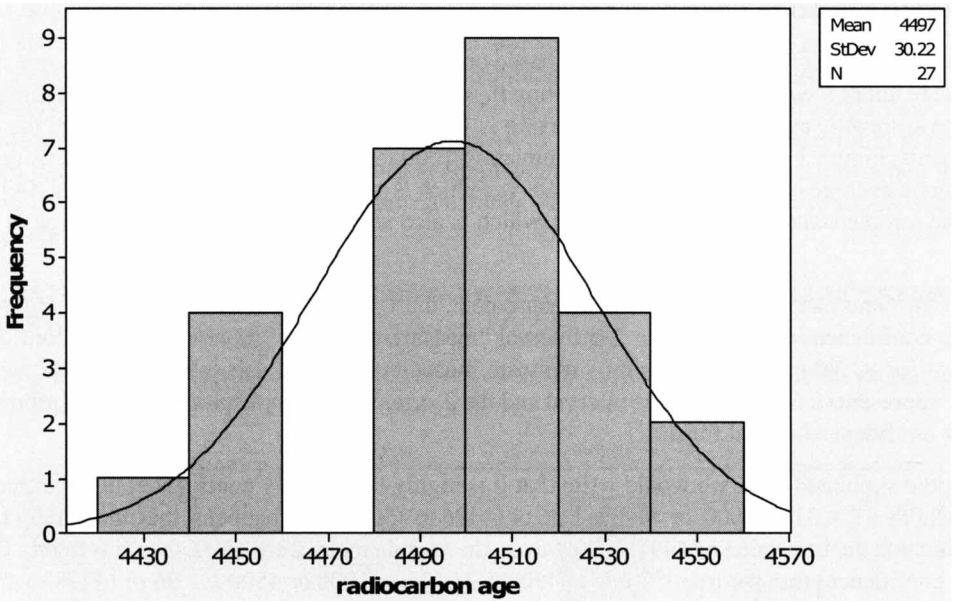

Figure 1 Histogram of a series of 27 Belfast cellulose measurements

although the standard deviation will be more severely affected. The median and quartiles will remain relatively unaffected.

The example below in Table 2 shows the effect of a single outlying measured ^{14}C age on the mean ^{14}C age and the standard deviation. The median and the 2 quartiles remain unchanged. In this example, the largest of the original 10 measurements in Table 1 is replaced by the value in Table 2 below; all other values remain the same. It is clear that the standard deviation increases, demonstrating its sensitivity to any extreme values, but that the mean is affected much less (i.e. is more robust).

Table 2 Effect of a single outlier on the mean and standard deviation of 10 replicate measurements.

Outlier	4590	4640	4840
Mean	4509	4514	4534
Standard deviation	39	52	111

THE NORMAL OR GAUSSIAN MODEL

We must consider the statistical basis for the calculations described in the previous section and ask under what circumstances are these calculations and interpretations valid? Since an estimated ^{14}C age, X, can be considered as a realization of a random process (due to the nature of radioactive decay), we can consider a conceptual model that states that there is a true but unknown ^{14}C age (often in statistical terminology called a parameter, written as μ), and that our estimated ^{14}C age can be modeled as a realization of that random variable with an expected value equal to the true age and standard deviation (or uncertainty) represented by a further unknown parameter, σ. The commonly used probability model for the ^{14}C dating process is the Gaussian or Normal model, frequently written as $X \sim N(\mu,\sigma^2)$. μ equals the true but unknown ^{14}C age and the standard deviation, σ, is the measure of uncertainty. In practice, the uncertainty or σ is quantified in the error (standard deviation or sigma) s that is reported with a measured (estimated) ^{14}C age.

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

Estimating μ and σ

The usual method of estimating μ and σ is based on a series of ^{14}C measurements on the same material made under identical conditions. Denoting these measurements by $x_1,..., x_n$, then the estimate of μ is the arithmetic average and the estimate of σ is s, the standard deviation calculated from the measurements. Figure 2 shows the results of simulating 1000 results from a $N(4497, 30.2^2)$ density. The arithmetic average of all 1000 results is 4498 yr (which is very close to the true ^{14}C age of 4497 yr) and the sample standard deviation is 31 yr, which is also very close to the theoretical value of 30.2 yr.

Under the Gaussian model for a single measurement, the 1-σ interval "measurement ± s" represents a 68% confidence interval and the 2-σ interval "measurement ±2 s" represents a 95% confidence interval for μ. When we have a set of replicate measurements, the 1-σ interval "sample mean ± SEM" represents a 68% confidence interval and the 2-σ interval "sample mean ± 2 SEM" represents a 95% confidence interval for μ.

Using the simulated data, we would write that it is highly likely (95% confidence) that the true ^{14}C age is $4498 \pm 2 \times 31/\sqrt{1000}$ or 4498 ± 1.96 or (4496 to 4500). As it happens, this uncertainty range does include the true value of 4497. If, however, the sample mean were 4500, then it is highly likely (95% confidence) that the true ^{14}C age is $4500 \pm 2 \times 31/\sqrt{1000}$ or 4500 ± 1.96 or (4498 to 4502). This uncertainty range does not include the true value of 4497, which can happen. Indeed, if we were to repeat this experiment 100 times, we would expect 5 of the 100 such intervals not to include the true value. A statistical confidence interval is, by its nature, random and the confidence levels are based on the long-run properties of such intervals.

Both Figures 1 and 2 also show the probability density function for a Normal (or Gaussian) variable with mean 4497 yr BP and standard deviation 30.2 yr. It is clear that for the relatively small numbers of measurements in Figure 1, it is difficult to assess whether the Normal model is in fact appropriate. In comparison, in Figure 2 with a much larger data set, it is clear that the theoretical shape of the Normal density function is well supported by the data. One of the key properties of the Normal distribution is the fact that it is symmetrical about the mean and this property is important in the calculation and manipulation of ^{14}C ages.

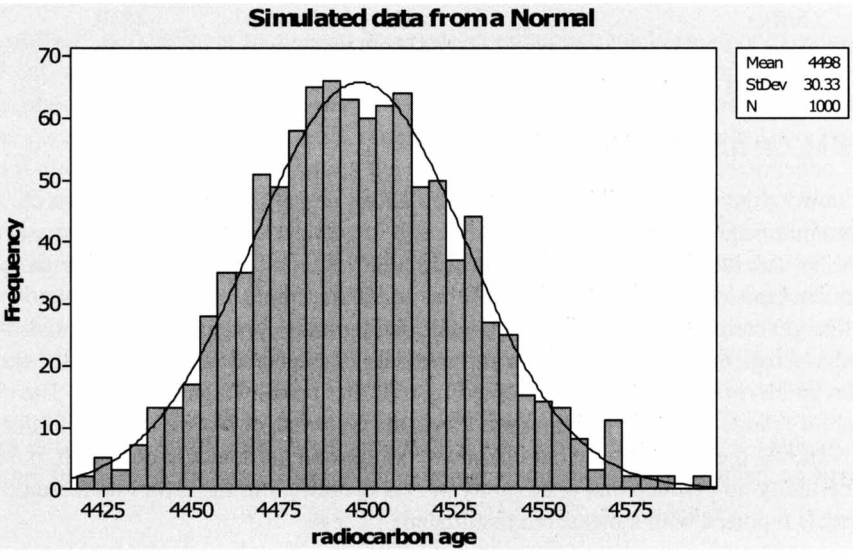

Figure 2 Histogram of simulated Normal data

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

Assuming of course that the Normal or Gaussian model is appropriate for ^{14}C ages means that we can interpret the usual summary statistics (mean and standard deviation) in a specific way. If the stochastic model is not Gaussian, e.g. one which is bimodal, then such simple summary statistics are not easily interpretable and the uncertainty ranges no longer have the desired properties. This is sometimes the case when a ^{14}C age measurement is very old (when the quoted error is given as 2 values, so the uncertainty is not symmetrical about a central value [the estimate]) or when the ^{14}C age is calibrated, the resulting calibrated age range is not symmetrical and must be evaluated numerically.

ESTIMATING THE ERROR AND ASSESSING WHETHER THE RESULTING UNCERTAINTY IS REALISTIC

Figure 3 shows the histogram of the quoted errors on the 27 individual measurements on Belfast cellulose. The largest error was 33 yr, the smallest 13 yr.

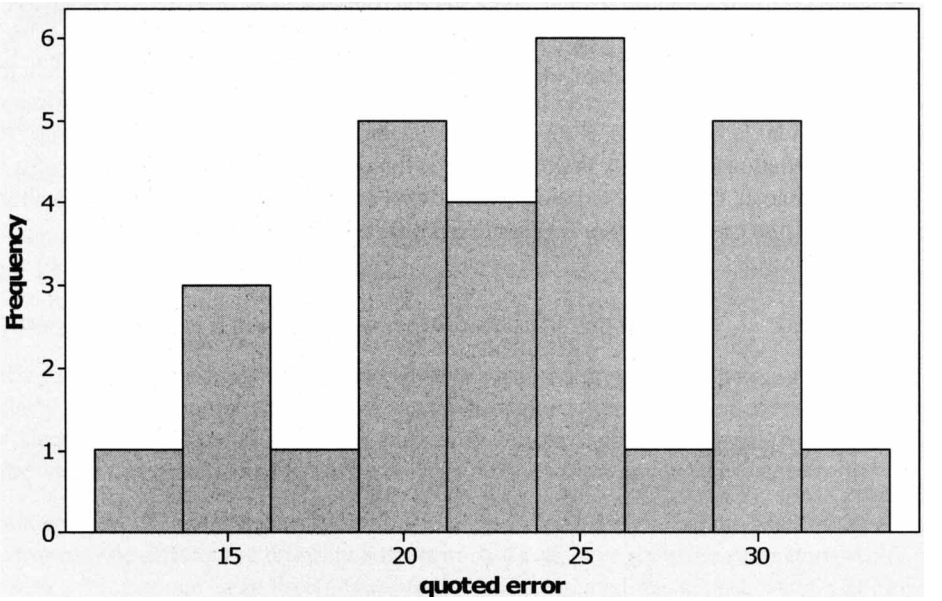

Figure 3 Histogram of quoted errors on Belfast cellulose results

Is My Uncertainty Estimate Realistic?

For a single estimated ^{14}C age, the commonly quoted error is based on counting statistics, but may also include components due to other laboratory corrections, and is assumed to represent the spread we would see were we able to repeat the measurement many times. The validity of the quoted error is often assessed by comparing the standard deviation on a replicate series of measurements to an individual quoted error. If a laboratory made 10 replicate measurements (i.e. under identical conditions, sometimes called repeatability conditions), each with the same quoted error (and if these were realistic, e.g. included all the components of variation), then the standard deviation of the series of measurements (often written as σ_r and called the **repeatability standard deviation**) should be comparable to the quoted error. More frequently, as in Figure 3, each measurement has a slightly different quoted error and commonly σ_r is larger than the individual quoted errors, suggesting that there are some sources of "random" and quite small variations not accounted for in the quoted errors.

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

For the Belfast cellulose example, we have a situation where the scatter (standard deviation) in the 27 measurements is 30.2 yr, the quoted errors are similar to or slightly smaller than this value, and so a laboratory when reporting an individual result might choose to use the repeatability standard deviation as the minimum error that they would quote. This type of calculation is one of the reasons why many laboratories adopt a secondary or in-house reference material, which they will routinely measure. They may have several such materials of different age, and each material will typically undergo all the laboratory procedures (including pretreatment). In this way, laboratories are able to check whether their quoted errors are realistic.

Error Multipliers

In recent years, there have been a number of publications where the use of a laboratory error multiplier has been introduced (Gulliksen and Scott 1995; Scott 2003). In such cases, the laboratory quoted error is increased (or decreased) by a multiple that is estimated, typically from a series of replicate measurements. The error multiplier captures sources of variation in the estimated ^{14}C age that are not accounted for in the quoted error. The theoretical (but still Gaussian) model for the ^{14}C measurement X is that $X \sim N(\mu, \theta^2\sigma^2)$, where μ, θ, and σ^2 are unknown. A series of replicate ^{14}C measurements are made, denoted by $x_1,\ldots,x_n$ with quoted errors $s_1,\ldots,s_n$ and from which these 3 unknown quantities will be estimated. As before, the quoted error s will be used as an estimate of σ. The theoretical model is interpreted as meaning that each measurement has the same true ^{14}C age, μ, but that the population uncertainty is $\theta\sigma$, where θ is the error multiplier. The estimate of μ is then the weighted average $\hat{\mu}$ (the weights being proportional to s_i) (Equation 6) of the measurements, and the estimate of the error multiplier θ (Equation 7), $\hat{\theta}$, is given by

$$\hat{\mu} = \frac{\sum_1^n \frac{X_i}{s_i^2}}{\sum_1^n \frac{1}{s_i^2}} \tag{6}$$

$$\hat{\theta}^2 = \frac{1}{n}\sum_1^n \left(\frac{X_i - \hat{\mu}}{s_i}\right)^2 \tag{7}$$

or

$$\hat{\theta} = \sqrt{\frac{1}{n}\sum_1^n \left(\frac{X_i - \hat{\mu}}{s_i}\right)^2}$$

It is increasingly common that laboratories will make a long series of replicate measurements on a reference material and assess the standard deviation of the set. If the standard deviation is greater than the quoted errors on the individual estimated ages, then this would suggest an unaccounted source of variation and a laboratory might chose to quote the larger of the standard deviation of the set and the individual quoted error, or they might use an error multiplier approach to provide a more realistic uncertainty. (Note this common error multiplier approach uses a multiplicative model; as an alternative an additive model could also be used, such that the overall uncertainty is $\sqrt{(s_i^2 + \tau^2)}$.)

ACCURACY AND PRECISION

Measurements are routinely described as being either accurate and/or precise.

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

- **Accuracy** is the closeness of agreement between a measurement and the true or reference value. If we imagine a series of measurements, each with the same true value, then if the average of the measurements does not equal (within error) the true value, then the measurement is said to be biased, where the bias is the difference between the **expected value** or average of a large series of measurements and the true value. Bias is usually considered to be a systematic error.
- **Precision** is the closeness of agreement between a series of independent measurements obtained under identical conditions. Precision depends on the **distribution of random errors**, and is commonly computed as the standard deviation of the results. As the standard deviation increases, the precision decreases (RSC 2003a,b).

The archery targets in Figure 4 below depict accuracy and precision graphically.

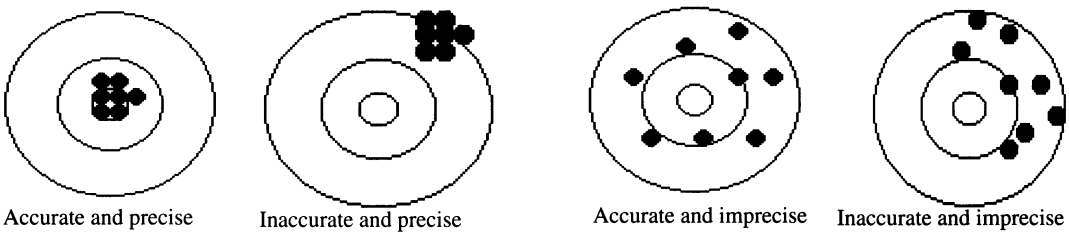

Figure 4 Accuracy and precision

These and many other definitions can be found in the Royal Society of Chemistry, AMC technical brief no. 13, September 2003 (RSC 2003a).

How Does a User Know if the Measurement is Accurate and Precise?

Both accuracy and precision in ^{14}C dating are much desired properties. The accuracy of ^{14}C dating is regularly assessed internally by the laboratory through routine measurement of primary and secondary standards or reference materials (which are either independently-known-age samples, or those for which a consensus age has been derived perhaps from an interlaboratory trial).

The precision of a ^{14}C age is quantified in the associated quoted error, but the basis of the calculation of the error may be different in different laboratories. As we have seen, the estimated precision associated with a ^{14}C age can be indirectly assessed through repeated measurements of a homogeneous material. In radiometric laboratories, replicate measurements of a single unknown-age sample are often impossible; thus, a radiometric laboratory will typically make a series of measurements of a secondary standard and use the variation in the results to provide a sample-independent estimate of precision, which can then be compared with (and used to adjust if necessary) the classical counting error statistic, which is derived for each unknown-age sample. A similar approach can be used in AMS laboratories; however, an alternative approach is also possible, since many AMS targets can be prepared from a single unknown-age sample and measured. Thus, in principle, the estimated precision of the AMS measurement for each unknown-age sample can be based on multiple measurements of replicated targets.

Another measure of assessing whether the laboratory uncertainty estimate is realistic and indeed whether a measurement is accurate can be obtained in a situation (known as **reproducibility** conditions) where independent measurements are obtained by the same method on identical samples in different laboratories. Such a set of conditions is commonly encountered in a **collaborative** or **proficiency trial**.

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

Repeatability (r) refers to measurements made under identical conditions in 1 laboratory, while reproducibility (R) refers to measurements made in different laboratories, under different conditions. Both repeatability and reproducibility are the closeness of agreement between the ^{14}C ages under these 2 different scenarios. The reproducibility standard deviation quantifies the maximum variability in results.

The ^{14}C dating community has been involved both in small-scale specialized intercomparisons and more general collaborative trials, and the results are regularly reported.

INTERLABORATORY COLLABORATIVE TRIALS

Proficiency testing is often used as a method for assessing the accuracy of laboratories in conducting particular measurements. It usually involves distributing "identical" portions of the test material to each laboratory and then analysis of the reported results, which helps each laboratory assess the accuracy and precision of their measurement.

The ^{14}C community has devoted considerable care and effort to establishing and maintaining primary standards and reference materials and in the routine organization of laboratory intercomparisons or collaborative trials to verify comparability of measurements. From such a series of collaborative trials, secondary standards or reference materials have been developed, including internationally recognized materials such as ANU sucrose (also known as IAEA-C6), Chinese sucrose, and the IAEA C1–C6 series (Rozanski et al. 1992), augmented by additional oxalic acid samples (now IAEA C7 and C8) (Le Clerq and van der Plicht 1998). The activity of these materials has been estimated from large numbers of measurements made by many laboratories. Recently, further natural materials from the Third and Fourth International Radiocarbon Intercomparisons, TIRI (Gulliksen and Scott 1995) and FIRI (Bryant et al. 2000; Scott 2003), have been added to this list. The activities of these standards and reference materials span both the applied ^{14}C age range and the chemical and biological composition range of typical samples. A summary of the intercomparisons organized within the ^{14}C dating community can be found in Scott (2003). The results from the most recent intercomparisons have indicated that laboratories are in general providing accurate results, but have pointed to variation in the results beyond that described by the quoted errors. This provides some evidence that the quoted errors are underestimates of the precision for some laboratories. VIRI (the Fifth International Radiocarbon Intercomparison) is currently under way (see Scott et al. 2007, these proceedings, for a report on the first phase of VIRI). The second phase (bone measurement) is completed and the analysis will be reported shortly. Typically, VIRI (and earlier intercomparison) reports have included assessments of accuracy and precision (including error multipliers).

Manipulating ^{14}C Dates

Having introduced some of the more fundamental properties of a ^{14}C measurement, it is now worth considering some of the inferential questions that ^{14}C measurements are used to answer and to consider how they can be addressed, given our understanding of the measurement processes. First, we will assume that the user selecting and collecting the sample and the laboratory making the measurement have used due care. Uncertainty due to sampling is not an issue we have discussed, but it is an important source of uncertainty—one which does not lend itself to statistical evaluation. In the ^{14}C context, we can imagine that the ^{14}C measurement is made on a small component of the whole sample (e.g. a few grains from a silo or a few shells from a midden), then the security of the sampled materials' connection with the event being dated represents an example of uncertainty due to sampling. These examples also illustrate how difficult it is to quantify such uncertainty.

Second, the most common manipulation is, of course, calibration of ^{14}C dates, but this will not be covered here, since it is a specialist subject in its own right, and through the availability of software such as OxCal, CALIB, and BCal (c14.arch.ox.ac.uk/oxcal.html; http://calib.qub.ac.uk/calib/; http://bcal.shef.ac.uk/, respectively), users now have access to sophisticated software and modeling tools. Conventional ^{14}C ages are frequently converted/transformed to a calibrated age range (on the historical timescale). The error in years BP (yr BP) on the estimated ^{14}C age must also undergo a transformation to give the corresponding error on the calendar year scale. However, complexities arise due to the complex pattern of ^{14}C variations. The simplest complication is that although we often model the estimated ^{14}C age using a Gaussian model, once transformed or calibrated the resulting stochastic model for the calibrated age is no longer Gaussian; it is often multimodal and no longer symmetric. The net effect is that often the calibrated result has a greater uncertainty and this uncertainty cannot be presented as a simple ± term; rather, the result must be quoted in the form of a range.

However, there are several commonly used, precalibration analyses and these are described below.

a) Are Two Radiocarbon Dates Identical Within Error?

Perhaps one of the simplest inferential questions is the comparison of 2 ^{14}C dates. Are they the same (at least within error)? With the interpretation of the quoted error as described above, we can apply some simple procedures to compare 2 ^{14}C dates (assuming that they are independent), such as might be the case if comparing measurements made on identical samples but in 2 different laboratories.

Again, assuming that the Normal or Gaussian model is appropriate, we would typically work with the 2-σ uncertainty range. The question is whether the 2 samples have the same true ^{14}C age. The answer depends on whether the 2 uncertainty ranges (calculated as estimated ^{14}C age $\pm 2\,\sigma$) overlap. If this is the case, then there is no evidence to suggest that the true ^{14}C ages are not the same, and so one might conclude that the 2 measurements relate to the same event. In fact, more formally, if we assume that x_1 and x_2 are the measured ^{14}C ages, and that each can be modeled as $X_i \sim N(\mu_i, s_i^2)$, then we can calculate the difference, $x_1 - x_2$, and the error on the difference as $\sqrt{(s_i^2 + s_2^2)}$. The uncertainty range on the difference is then $x_1 - x_2 \pm 2\sqrt{(s_i^2 + s_2^2)}$, and if this range includes the value 0, then we would conclude that it is highly likely (approximate 95% confidence) that the 2 samples have the same, true ^{14}C age.

Example: Two examples of measurements of a charred grain sample (VIRI sample B) from excavations at Beth Saida, which were provided by Elisabetta Boaretto of the Weizmann Institute, are considered. The measurements were made in 2 different laboratories and so are assumed statistically independent. The expected archaeological age is 2800 BP.

The Reported Ages are 2759 ± 39 BP and 2811 ± 20 BP

The difference is –52 yr and the error is 44 yr; therefore, the uncertainty range is –52 ± 88 yr and includes 0. There is no evidence that these 2 samples do not have the same true ^{14}C age. These 2 measurements could therefore be legitimately combined in a weighted average.

The Reported Ages are 2885 ± 37 BP and 2781 ± 30 BP

The difference is 104 yr and error is 48 yr; therefore, the uncertainty range is 104 ± 96 yr or 8–200 yr and does not include 0. We could conclude that within the individual uncertainties on the measurements, these 2 samples do not have the same true ^{14}C age. Therefore, these 2 measurements could not be legitimately combined.

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

Since we know the expected archaeological age, we can also compare each individual measurement with 2800 BP. Three of the results have uncertainty ranges at 2 σ that include 2800 BP; thus, these measurements provide support for the archaeological age. The fourth measurement, 2885 ± 37 BP, has an uncertainty range that does not include 2800 BP, so it is implausible based on this measurement that the true archaeological age is in fact 2800 BP.

A further complication arises if the 2 measurements are related in some way, e.g. if they are duplicates (the same material measured twice in the same laboratory, perhaps in the same batch or AMS wheel), then we might expect some of the small, random errors to be the same for each measurement, and so the measurements are not statistically independent (they are correlated). This should be taken into account in the calculation, although often it is ignored. One of the simplest ways of treating such paired measurements is to calculate the difference between each duplicate pair and to analyze the differences and the standard deviation of the differences. Alternatively, one must incorporate the covariance between the 2 measurements into the combined error on the difference, so that the error on x_1–x_2 is $\sqrt{(s_i^2 + s_2^2 - 2s_{12})}$, where s_{12} is the covariance term. The additional s_{12} term can be either positive or negative, but we might expect it to be positive, so that a covariance term will reduce the error on the difference. Estimation of the covariance (and hence correlation) is practically challenging and would require a series of related measurements.

b) Can I Combine a Series of Radiocarbon Dates? (Are They Homogeneous?)

If we have a series of measurements, a common question is to ask whether it is possible to combine the ^{14}C determinations to give a (weighted) average and much smaller uncertainty range, before calibration. This is often described as a test of homogeneity (Ward and Wilson 1978) and is implemented in many of the calibration packages as a pre-processing tool before calibration. The procedure is a test of the hypothesis that given a set of n ^{14}C determinations, they all have the same true ^{14}C age, which if the case, provides a justification for combining them in an average that is then calibrated. Recent developments in calibration software means that this approach is becoming less favored.

The basis for the test is again a statistical model where each ^{14}C measurement has a true ^{14}C age, μ_i. We also must assume that the quoted error s_i for each measurement is realistic. The hypothesis that each ^{14}C measurement has the same true ^{14}C age, μ, is then tested by calculating the weighted mean, $\bar{X}_P$, and a test statistic, T, which is effectively the standardized residual sum of squares given by

$$\bar{X}_P = \frac{\sum X_i / s_i^2}{\sum 1 / s_i^2} \tag{8}$$

$$T = \sum_1^n \left(\frac{X_i - \bar{X}_P}{s_i} \right)^2 \tag{9}$$

If the hypothesis is true, then T will have a specific distribution, namely it will be χ^2 with n–1 degrees of freedom. We would reject the hypothesis if in fact T is greater than a critical value read from statistical tables.

The error on the weighted average is given by

https://doi.org/10.1017/S0033822200042351 Published online by Cambridge University Press

$$\sigma(\bar{X}_P) = \sqrt{\sum_1^n \frac{1}{s_i^2}} \tag{10}$$

Example: In a study (Ascough et al. 2006) to estimate the marine reservoir age, several deposits were selected from 6 Scottish coastal archaeological sites where the resolution of the stratigraphy allowed the application of a rigorous selection protocol (cf. Ascough et al. 2005) to obtain marine and terrestrial material that was reliably of the sample calendar age, for ^{14}C measurement. The ^{14}C results for 6 terrestrial samples taken from Skara Brae on the Orkney Islands are used to illustrate this question. The samples consisted of single entities (i.e. individual organisms) that represented a relatively short growth interval. The terrestrial samples were either carbonized plant macrofossils (cereal grains or hazelnut shells) or terrestrial mammal bones (cattle or red deer).

Table 4 ^{14}C ages from 6 terrestrial samples from Skara Brae.

Age	4555	4605	4525	4530	4270	4735
Error	40	40	40	35	40	40

Using all 6 measurements, the weighted average is 4536.34 yr and T is 72.2789.

The formal inference, in this case, would require T to be compared with a χ^2 (5), for which the critical value is 11.07. Thus, we would reject the hypothesis that the samples all had the same true ^{14}C age.

Using a subset of 4 measurements (excluding 4270 and 4735), the weighted average is 4552 yr and T is 2.612. The formal inference, in this case, would require T to be compared with a χ^2 (3), for which the critical value is 7.8. Thus, we would not reject the hypothesis that the samples all had the same true ^{14}C age, and so the weighted average (with its error) could be used in subsequent calibration. This last example, for illustrative purposes only, omitted 2 measurements (the oldest and the youngest), but the response to a large T value and rejection of the hypothesis is context dependent. There may be several ways of generating a homogeneous subgroup from a set of measurements.

DISCUSSION

This paper has focused on some ^{14}C measurement issues, including general definitions and their applications to the ^{14}C measurement process. Of particular interest is the evaluation and interpretation of the measurement or analytical uncertainty. Other sources of uncertainty are not here discussed, including archaeological sampling uncertainty (such as selecting a few grains from a grain silo), which may be great and may well exceed any analytical uncertainty, but is difficult to quantify and the uncertainty arising from the calibration of the conventional ^{14}C ages where the measurement error in yr BP on the estimated ^{14}C age is transformed to give the corresponding error on the calendar year scale.

The fundamental premise the paper is built on is that all measurement is subject to error, which creates uncertainty. Every time that a ^{14}C measurement is repeated under identical conditions on an identical sample (even if this were possible), a different result is obtained. For a single measured ^{14}C age, the commonly quoted error, based on counting statistics, is used to determine the uncertainty associated with the ^{14}C age. The basis of the quoted error and the resulting analytical uncertainty are described in some detail in relation to interpretation and statistical manipulation. The use of replicate laboratory measurements, including the use of error multipliers to assess whether the measurement uncertainty is realistic, is discussed.

The 2 key measurement properties of accuracy and precision are described and linked to the preceding discussion of measurement uncertainty (involving both stochastic and systematic terms).

The final 2 sections consider several fundamental inferential questions requiring manipulation of ^{14}C ages. The answer to the questions will be based on a statistical model, which is defined by the method used to evaluate the measurement uncertainty. For the simple cases presented here, the probabilistic assumption is that a ^{14}C measurement can be modeled using a Gaussian framework. It is important to remember that all statistical models require assumptions, and where possible, such assumptions should be verified.

The level of mathematical detail in the exposition has been kept at a minimum, to emphasize the importance of the concepts, which, although not exhaustive, were chosen to stimulate discussion and ultimately to lead to the development of guidelines of good practice for evaluation of measurement uncertainty. The paper is intended both for the ^{14}C laboratory and the user of the ^{14}C dates.

Differences exist in the data manipulations used by individual ^{14}C laboratories, since not every laboratory operates in exactly the same way. AMS, gas proportional, and liquid scintillation facilities, while operating different measurement systems, however still adhere to the same underlying measurement principles. It is hoped that this paper will build on some of the seminal work of authors including Stuiver and Polach (1977), Donahue et al. (1990), and Mook and van der Plicht (1999).

REFERENCES

Ascough P, Cook G, Dugmore A. 2005. Methodological approaches to determining the marine radiocarbon reservoir effect. *Progress in Physical Geography* 29(4):532–47.

Ascough PL, Cook GT, Dugmore AJ, Scott EM. 2007. The North Atlantic marine reservoir effect in the Early Holocene: implications for defining and understanding MRE values. *Nuclear Instruments and Methods in Physics Research B* 259(1):438–47.

Bevington PR, Robinson DK. 2003. *Data Reduction and Error Analysis for the Physical Sciences*. 3rd edition. New York: McGraw-Hill. 352 p.

Bryant C, Carmi I, Cook G, Gulliksen S, Harkness D, Heinemeier J, McGee E, Naysmith P, Possnert G, Scott M, van der Plicht J, van Strydonck M. 2002. Sample requirements and design of a inter-laboratory trial for radiocarbon laboratories. *Nuclear Instruments and Methods in Physics Research B* 172(1–4):355–9.

Burr GS, Donahue DJ, Tang Y, Beck W, McHargue L, Biddulph D, Cruz R, Jull AJT. 2007. Error analysis at the NSF-Arizona AMS facility. *Nuclear Instruments and Methods in Physics Research B* 259(1):149–53.

Cook GT, van der Plicht J. 2007. Radiocarbon dating. In: Elias SA, editor. *Encyclopedia of Quaternary Science*. Amsterdam: Elsevier. p 2899–911.

Donahue DJ, Linick TW, Jull AJT. 1990. Isotope-ratio and background corrections for accelerator mass spectrometry radiocarbon measurements. *Radiocarbon* 32(2):135–42.

Gulliksen S, Scott EM. 1995. Report of the TIRI workshop, Saturday 13 August 1994. *Radiocarbon* 37(2): 820–1.

Kelvin WT. 1893. *Popular Lectures and Addresses, Volume 1: Electrical Units of Measurement*. Lecture given to the Institute of Civil Engineers. London: McMillan & Company. p 80–143.

Le Clercq M, van der Plicht J, Gröning M. 1998. New ^{14}C reference materials with activities of 15 and 50 pMC. *Radiocarbon* 40(1):295–7.

Mook W, van der Plicht J. 1999. Reporting ^{14}C activities and concentrations. *Radiocarbon* 41(3):227–39.

Royal Society of Chemistry [RSC]. 2003a. Terminology—the key to understanding analytical science. Part 1: accuracy, precision and uncertainty. AMC technical brief 13. Available online http://www.rsc.org/images/brief13_tcm18-25955.pdf.

Royal Society of Chemistry [RSC]. 2003b. Is my uncertainty estimate realistic? AMC technical brief 15. Available online at http://www.rsc.org/images/brief15_tcm18-25958.pdf.

Rozanski K, Stichler W, Gonfiantini R, Scott EM, Beukens RP, Kromer B, van der Plicht J. 1992. The IAEA ^{14}C intercomparison exercise 1990. *Radiocarbon* 34(3):506–19.

Scott EM. 2003. The Third and Fourth International Radiocarbon Intercomparisons. *Radiocarbon* 45(2):135–328.

Stuiver M, Polach H. 1977. Discussion: reporting of ^{14}C data. *Radiocarbon* 19(3):355–63.

Taylor BN, Kuyatt CE. 1994. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical note 1297. Available online at http://physics.nist.gov/Pubs/guidelines/contents.html.

Ward GK, Wilson SR. 1978. Procedures for comparing and combining radiocarbon age determinations: a critique. *Archaeometry* 20(1):19–31.