LETTER TO THE EDITOR

BIAS, ACCURACY, AND PRECISION

2 July 1999

Dear Sir,

In the most recent issue of *Radiocarbon* (Vol 41, Nr 1, 1999), a paper by Rasmussen et al. discusses a blind check of the accuracy of the Copenhagen radiocarbon dating system. The authors report on a comparison of a series of ^{14}C dates from dendrodated samples, measured over a 23-yr period, both with the absolute dendrodates and also with the bidecadal calibration curve (Stuiver and Pearson, 1993).

The authors compare 92 uncalibrated ^{14}C dates of the Copenhagen measurements and the equivalent uncalibrated ^{14}C dates from the Stuiver and Pearson bidecadal calibration curve (1993). The comparison is based on the difference between each pair of measurements and its error (calculated as the square root of the sum of the squared counting errors for each result). From the results, they calculate the average of the differences (or if necessary the weighted average) as 54 yr and present the histogram of the individual differences (see their Figure 2) as well as quantifying the variation in the results by the standard deviation (quoted as ± 72 yr).

Given the symmetry of the histogram, it seems reasonable to assume that the underlying distribution of differences is Normally distributed so that 95% of the individual differences should lie in the range $54 \pm 2 \times 72$, or –90 to 200 yr. This is borne out in the histogram. The authors then conclude "the comparison shows a good agreement, well within 1σ between the ^{14}C measurements" and that "good accuracy can be obtained".

I would like to draw the authors' attention to a paper which deals with the comparison of two sets of measurements (Bland and Altman 1986). Bland and Altman suggest that figures such as Figure 1 of Rasmussen et al. can be difficult to interpret and may be misleading, and that to measure agreement, it is useful to plot the difference between the two measurements against the average of the two measurements. In this case, the summary of agreement is the mean difference (or "bias") and the standard deviation of the differences. The precision of the agreement is then quantified by the *estimated standard error* of the mean difference. Bland and Altman avoid the use of the word "accuracy".

Rasmussen et al. only briefly mention laboratory bias, preferring to use the term accuracy, but it is worth digressing briefly to considering accuracy, bias, and precision, since the definitions of these terms are relevant to the discussion of this paper.

An accurate result is one which is close to the true value and which is precise. An accurate result should have zero bias, where bias in the estimation sense is defined as the difference between the expected value of the statistic (or parameter estimate) and the true value of the population parameter. Here, the authors are trying to estimate the true difference between the Copenhagen and the calibration results, which is estimated by the mean of the differences between the two sets of dates. Precision refers to how varied the measured results are, and so it measures the spread or scatter in the results, but in an estimation sense, precision is the error on the estimate, which in this case would be the error on the mean difference.

Having calculated the mean difference and standard deviation, the authors seem to rest their case concerning the accuracy of the Copenhagen results, since 54 ± 72 yr shows good agreement ("well

https://doi.org/10.1017/S0033822200019561 Published online by Cambridge University Press

within ± 1 σ") with the calibration data and the quoted standard deviation is in reasonable agreement with the estimated value. In fact, they have summarised the agreement between the two laboratories. However, we need to be careful at this point: if we are considering the accuracy of the laboratory, then 54 yr is the estimate of the true difference between the Copenhagen and calibration results, and 72 yr is a measure of the variation within the population of differences; but the precision of the agreement depends on the estimated standard error, which is the standard deviation divided by the square root of the number of samples. Thus the precision with which we are able to estimate the mean difference is 7.5 yr (72 divided by 9.6).

Thus, from the differences, the best estimate of laboratory agreement is 54 yr, and the precision with which we have estimated the agreement is 7.5 yr. Together, a plausible range of values for the laboratory agreement can be estimated, or in other words, a 95% confidence interval can be calculated as 54 ± 15 yr (2σ), or 39–69 yr. This interval is highly significant, since it does not include the value 0; we can conclude that the Copenhagen and Stuiver and Pearson results are on average different (i.e., do not agree), with the Copenhagen results highly likely to be between 39 and 69 yr younger than the Stuiver and Pearson results. Thus I would dispute the description of the results as being accurate.

There is, of course, the question of the validity of this calculation: the samples are not identical, they may not have exactly the same true age, and the measurements were made over a period of 23 yr, so it is likely that laboratory conditions changed in that time. The use of the term "accuracy" by the authors implies the existence of a true value (and hence the possible existence of a bias), but in this case, the authors are really measuring agreement, so that this might have been a better term to use.

These arguments aside, I believe that it is still valid to make this estimate of agreement and that we must calculate the precision of this estimate, irrespective of how it is called, using the appropriate term.

Marian Scott
Department of Statistics
The University of Glasgow

REFERENCES

Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 8476:307–10.

Rasmussen KL, Tauber H, Bonde N, Christensen K, Theodórsson P. 1999. A 23-year retrospective blind check of accuracy of the Copenhagen radiocarbon dating system. *Radiocarbon* 41(1):9–15.

Stuiver M, Pearson GW. 1993. High-precision bidecadal calibration of the radiocarbon time scale, AD 1950–500 BC and 2500–6000 BC. *Radiocarbon* 35(1):1–24.