Letters

Some comments arising from Berger (2010)

R. F. Galbraith

Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK (e-mail: rex@stats.ucl.ac.uk)

(Received 30 November 2010; in final form 22 March 2011)

Abstract

I note a fundamental error in the "alternate form of probability-distribution plot" proposed by Berger (2010a) and comment on some related issues, including transformations, radial plots, empirical distributions, kernel density estimates, weighted means and selected data.

Introduction

In a recent article in this journal (Galbraith 2010) I discussed, among other things, a type of "probability density" graph that has sometimes been used to display OSL equivalent doses for a sample of single grains or aliquots. In this graph, each D_e value is replaced by a Gaussian curve centered on the observed D_e with standard deviation equal to the standard error of D_e and the curves are then summed pointwise. I referred to this specifically as a "PD" graph and distinguished it from a conventional kernel density estimate (KDE). I tried to explain what it is doing and why it is not to be recommended.

Also, and perhaps more importantly, I tried to encourage researchers to think about the *meaning* of an equivalent dose frequency distribution. To what extent does it represent frequencies in a natural population, rather than artefacts of sampling and grain selection or variation in luminescence and experimental procedures? I also distinguished between the distributions of *true* equivalent doses (where hypothetically, D_e values are measured without error) and *observed* D_e values, a distinction necessary for understanding data and making reliable inferences.

In the same issue of this journal, Berger (2010a) proposed an "alternate form of probability-distribution plot" for OSL equivalent doses, which he called a "Transformed-PD" plot, or TPD plot for short. The essence of this was to construct the sum of Gaussian curves using log D_e values and relative standard errors, rather than actual D_e values and their absolute standard errors. But the probability density curves so obtained were presented on a linear D_e scale, having apparently been transformed (without comment) from the log D_e scale. Unfortunately this transformation was not done correctly and they do not represent the intended probability distributions — and they do not have the meaning attributed to them in that paper, as acknowledged by Berger (2010b).

Berger (2010a) also presented some interesting data examples and raised several other issues that are perhaps worth further comment — concerning radial plots, log transformations, empirical distributions, kernel density estimates, weighted means and selected data. He rightly noted that radial plots offer advantages over PD plots. In fact his radial plots are far more informative than his corresponding PD and empirical distribution plots, and his data presentation would be less convincing without them. This is not to say that one should make *only* radial plots of D_e values, but it supports my recommendation to look at them in addition to other plots that might be made. Berger (2010a) also stated that radial plots could not be used for samples containing zero or negative D_e values, because they use a logarithmic transformation. But of course one can make a radial plot without using a log transformation (or indeed any transformation), as acknowledged by Berger (2010b).

However, what Berger (2010a, page 13) saw as the "two main criticisms" of the conventional PD plot are merely to do with the nature of the empirical distributions of D_e values and their errors, and he did not recognise the more fundamental problems noted in Galbraith (1998) and Galbraith (2010). These are to do with their meaning and interpretation. On one level, a PD plot might just be regarded as an empirical smoothing of the data. If so, it is a poor one compared with, say, a conventional KDE where the kernel bandwidth is chosen according to sample size (among other things). But often PD plots are mis-interpreted and lead to fallacious arguments and unconvincing science.

I elaborate on some of these points below. This article is not intended to be a comprehensive critique of Berger (2010a) but rather a discussion of some statistical issues arising there and elsewhere in the OSL literature.

Transforming a probability density function

To illustrate the incorrect transformation mentioned above, look at the solid line in Berger (2010a, Figure 3). That curve is supposed to correspond, on the natural log scale, to an equal mixture (or sum) of four Gaussian probability density functions (pdfs) all with the same standard deviation (equal to 0.10) and means log(5), log(10), log(20) and log(30). If we transform this to the linear scale, it can be shown that we should get an equal mixture (or sum) of four *log-normal* pdfs. You can see that the TPD curve drawn must be wrong because the areas under its component curves should be equal and they are not.

When introducing his method, Berger (2010a, page 14) wrote: "However, application of equation 2 to these same artificial data generates the solid curve in Figure 3, accurately representing their respective probabilities". But his equation 2 does not generate the solid curve in his Figure 3, and that figure does not accurately represent their respective probabilities.

Figure 1 shows the correct distribution on both log and linear scales. The top panel shows the pdf of Z (corresponding to log D_e), denoted by $f(z)$, and the bottom panel shows the pdf of W (corresponding to D_e) denoted by $g(w)$, where $Z = \log(W)$. The formula relating these two pdfs is

$$g(w) = f(\log(w)) / w$$

for positive w. The factor $1/w$ is called the Jacobian of the transformation and is required in order to preserve the validity of probability statements, which are related to areas under the curve. That is, the probability that W lies between a and b must equal the probability that $\log(W)$ lies between $\log(a)$ and $\log(b)$ for any a and b. Different transformations have different Jacobians.

The bottom panel of Figure 1 also shows the pdf of a mixture of normal (rather than log-normal) pdfs as a red dotted line. The dashed curve in Berger (2010a, Figure 3) should be the same as this. The red dotted curve can hardly be seen as it differs only very slightly from $g(w)$. This reflects the fact that if a log-normal distribution has a small dispersion then it is very hard to distinguish it from a normal distribution with the same mean and standard deviation. In the present case, each component has a coefficient of variation of 10%, which is small enough to make the normal and log-normal distributions practically the same. If the coefficients of variation were larger then the two curves would differ more.

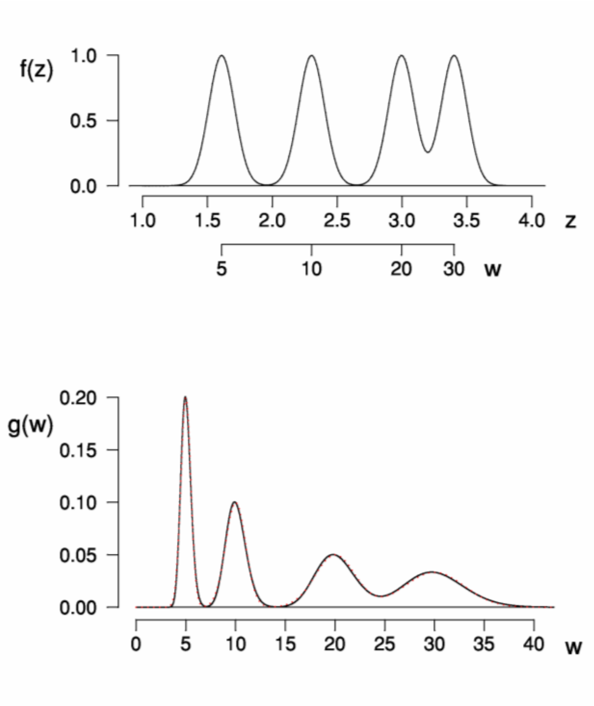

Figure 1: *Upper panel: the pdf $f(z)$ of an equal mixture of four Gaussian pdfs, each with standard deviation 0.10 and means log(5), log(10), log(20) and log(30). Lower panel: the solid curve shows the pdf $g(w)$ of the mixture of log-normal distributions obtained by the transformation $w = \exp(z)$ so that $z = \log(w)$. The dotted red curve shows an equal mixture of Gaussian pdfs with means 5, 10, 20 and 30, and standard deviations 0.5, 1, 2 and 3, respectively. The w axis corresponds to the scale of D_e and z corresponds to $\log(D_e)$.*

This discussion has nothing to do with the merits of PD plots as such, but it is instructive for understanding both log transformations and probability density functions.

What is Berger's TPD plot?

Berger (2010b) confirmed that in drawing his TPD plot on the D_e scale the Jacobian factor was omitted, so that the graph "does not manifest relative probabilities" and he suggested that it shows "rather something more akin to relative 'weighted' frequencies". What is it really a plot of, and is it useful?

Imagine a positive random variable W having a probability density function $g(w)$ describing relative frequencies in a population. Consider plotting a graph of $wg(w)$ against w. The total area under this curve would equal the mean, or expectation, of W; and the area under it between two values a and b would represent the 'contribution' to the overall mean from values of w in that range. However, it is hard to see what practical use this concept might have. In particular, modes or peaks of $wg(w)$ will not coincide with those of $g(w)$.

Now consider a TPD curve as plotted in Berger (2010a, Figures 3, 4, 6, 8 or 10). Those figures do not have numerical scales on their vertical axes, but it can be shown that the TPD curve is effectively a plot of $wg(w)$ against w, where $g(w)$ is an equal mixture (or sum) of n log-normal pdfs, where n is the number of grains or aliquots in the sample. The ith log-normal pdf in this mixture has mean y_i exp(½ r_i^2), where y_i is the observed D_e value for the ith grain and r_i is its relative standard error. The factor multiplying y_i here is slightly greater than 1 and greater for larger r_i. Thus the total area under the TPD curve is the average (or sum) of these means, so is a quantity a bit greater than the un-weighted sample mean (or sum) of observed D_e values. The TPD curve itself would therefore indicate some sort of relative contributions to this quantity from different doses represented in the sample data. The qualification "some sort of" is referring to the rather arbitrary role of the relative standard errors of D_e used in constructing the curve and hence in defining its meaning.

It is clear from this that Berger's TPD plot has no clear-cut interpretation as a frequency distribution of OSL equivalent doses.

Log transformations (or not) in radial plots

A strange idea appears to have arisen that a radial plot must necessarily use a log transformation of D_e and therefore can't be used to represent data containing zero or negative D_e values. Of course it is just as easy to make a radial plot using actual D_e values and their (absolute) standard errors as it is with log D_e values and relative standard errors. Examples of the former can be seen in Arnold et al. (2009) and Galbraith (2010).

Not only is it possible, it is also sometimes *more appropriate* to use a linear D_e scale. For example, when D_e values are close to zero their relative standard errors may happen to be large simply because they are relative to something small, and they may appear to be uninformative on a radial plot that is drawn with respect to relative standard errors, but properly informative when drawn with respect to absolute standard errors. Furthermore, in such cases there may be no clear relationship between D_e values and their standard errors, suggesting that the main sources of error are additive, rather than multiplicative (e.g. Arnold at al., 2009) and hence that comparisons on the linear D_e scale are more straightforward.

For a radial plot, the choice between using a log or linear D_e scale is partly related to whether points are better compared using relative or absolute standard errors. Another transformation, mentioned in Galbraith (2010), is the modified log transformation $z = \log(w+a)$ for some suitably chosen a. This can be useful to plot data having some large and some zero or negative values. If a is small, the transformation is similar to a log transformation, while if a is large it is more like a linear transformation, so you can think of the value of a as making a compromise between these two extremes.

Why are radial plots more informative?

Radial plots are more informative because they exploit the information in the precisions.

In his Figures 1 and 2 and corresponding text, Berger (2010a) cited a radial plot and "weighted histogram" (PD plot) from Galbraith (1988). These used some artificial data from a discrete two component mixture with component means +0.5 and –0.5. He found it "inexplicable" that I then wrote: "The weighted histogram is superficially attractive and suggests a *bimodal* distribution but does not point to the true *mixture* as informatively as the radial plot does".

Here is an explanation. The radial plot (reproduced as Figure 2 here) shows that the data are completely explained by a discrete two-component mixture — i.e. just two distinct values, roughly equal to +0.5 and –0.5. This is because you can easily imagine two radii going to +0.5 and –0.5 and, by referring to the ±2 scale on the vertical axis, see that *all* of the variation about them can be explained by the observation errors. This happens to be the model that was used to generate the data, but even if we did not know this, it is clear that the data are consistent with it.

Of course, it would be easier to see this by explicitly drawing the two radii with a ±2 shaded band about each. Each point would fall in, or very close to, one or other band. Furthermore their scatter looks like a superposition of homoscedastic random scatter about each line. You can confirm this by drawing your own lines and bands. I did not do this in my paper in order to avoid imposing any specific model and to allow the reader the freedom to consider possible mechanisms that might have produced these data.

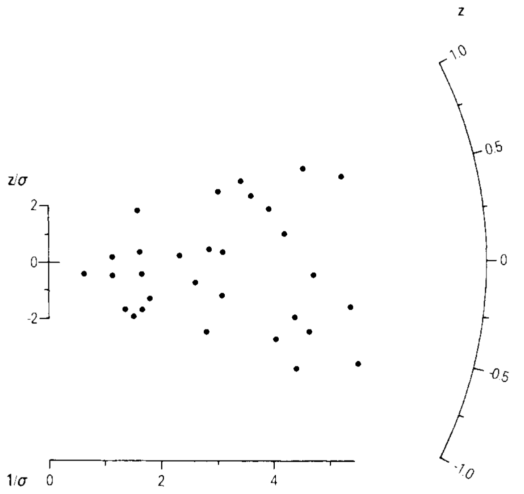

Figure 2: *A radial plot of artificial data from a discrete two component mixture with means +0.5 and −0.5.This is a copy of Figure 2 from Galbraith (1988).*

No such inferences can be made from the PD plot (Galbraith 1988, Figure 9a or Berger 2010a, Figure 2). Although the curve has two modes at about +0.5 and −0.5, it does not tell us either that more than one component is needed to explain the data, or that two is enough. The radial plot tells us both of these things. Furthermore there is no simple relation between the number of components in a mixture distribution and the number of modes its pdf has. For example, it is easy to construct a two component mixture of normal pdfs that is unimodal, and whose mode differs from both component means. Looking at the modes of a PD plot is even more ambiguous because it is not actually estimating the true dose distribution.

In Galbraith (1988, Figure 3) and Berger (2010a, Figure 1) the radial plot is re-drawn with different plotting symbols to show explicitly which observations came from each component. You can see there that most of the low-precision values are consistent with both radii and that many fall closer to the wrong radius (the component that they do not belong to) than to the right one. This reflects uncertainty associated with the observation errors that is inherent in the original data, and that cannot be resolved however you plot them.

That figure is instructive for another reason. Suppose that we wanted to estimate the lowest population component mean value. A method that may naturally spring to mind is to select a subset of points that we think belong to this component and calculate a mean or weighted mean of these. You can see from the figure that however the subset of points is selected it will always contain some from the higher component or omit some from the lower one. I comment further on this below.

I would encourage those interested to read the whole of that section in my original 1988 paper. That paper also explains the close connection between radial plots and least squares regression though the origin, which helps both with understanding and using radial plots.

Empirical distributions and kernel density estimates

Berger (2010a) rightly pointed out that more information is shown by adding a cumulative plot of ranked data with standard error bars. The individual points show the cumulative empirical distribution of the observed D_e values, and the one-sigma error bars display their standard errors. But a PD plot superimposed on it combines these incorrectly. It would make more sense to draw a conventional kernel density estimate (KDE), or a histogram, in order to see the shape of the distribution of observed D_e values.

The upper panel of Figure 3 shows the ranked observations with one-sigma error bars along with a Gaussian KDE for the data that I used in Galbraith (2010). Here I have chosen the kernel bandwidth to correspond to the bin width in my histogram (Galbraith 2010, Figure 1). The histogram there and KDE here both show the smoothed data to essentially the same degree of resolution. They emphasise slightly different aspects. The KDE shows more detail of the shape of the empirical distribution while the histogram shows numbers of grains and areas under the curve more clearly. The standard errors are displayed in Figure 3, but they are not used in the construction of the KDE.

What bandwidth should one use for a KDE? As with choosing the bin width of a histogram, there is no hard and fast rule. It should depend on the data and purpose. But there are general guidelines in the literature (in the R package, in particular). Note that such guidelines are based on the premise that one is trying to see the shape of the underlying frequency distribution from a sample of observations measured without error, which is usually not the case with observed equivalent doses.

The bandwidth of 0.058 in the upper panel of Figure 2 is very close to the value given by the R function bw.ucv (unbiased cross-validation) applied to these data, which is 0.061. In the lower panel I have drawn the graph again but using the bandwidth given by the

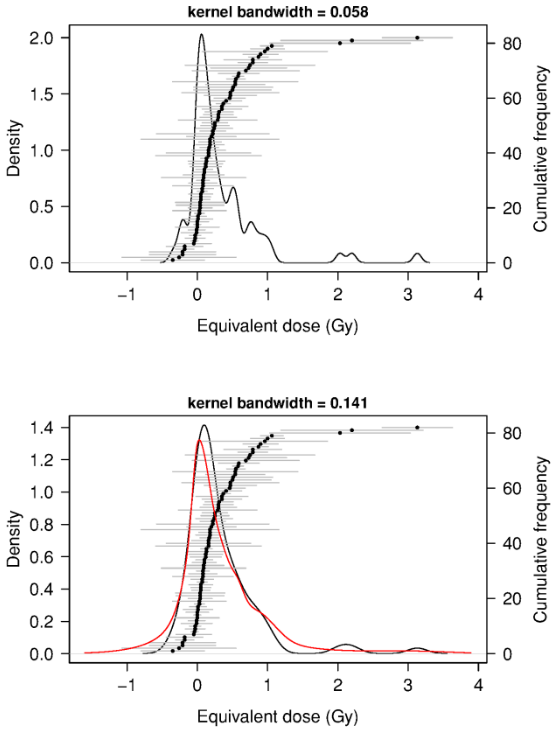

Figure 3: *Kernel density plots with two different bandwidths and empirical cumulative distributions for 82 single grain equivalent doses (data from Olley et al., 2004). The grey bars show ±1 standard error for each point. The red dotted curve shows a PD plot of these data.*

R function bw.nrd (one of several general rules of thumb). This is a larger bandwidth (0.141) giving a smoother graph. In some ways this version looks “nicer”, but of course more resolution is lost. This is good if you think that such resolution is meaningless, but bad if it might be informative. Note that when plotting a KDE there is an implicit assumption that the data were sampled from a *continuous* distribution, and presumably one that is meaningful.

In the lower panel of Figure 3 I have added a PD curve (the red dotted line). The scale is chosen so that the area under it is the same as that for the KDE. Given that it does not represent the distribution of either the true or observed D_e values, what use is it? In this example, it is even smoother than the KDE in that panel and its mode is lower (practically zero). Perhaps its worst feature, though, is its behaviour in the lower tail. We know that the true doses cannot be negative, so a negative observed dose gives us a lower bound on the absolute size of the *actual* error for that grain. For example, if we observed a D_e of –0.2 Gy, then, because the true value for that grain cannot be negative, the estimation error must be negative and not less than 0.2 Gy in absolute value — i.e. the observed value must be (more than) 0.2 Gy below the true value (regardless of what the standard error is). Yet the PD plot still puts more area below even the lowest negative D_e.

There is another distinction between a PD plot and a KDE (or histogram) that is worth repeating: for larger sample sizes one normally uses a smaller bandwidth for a KDE (or bin width for a histogram). But a PD plot does not get any better, in terms of resolution, as the number of grains increases. In general it gets worse because there are more low precision points to obscure the information. For example, the sample sizes in Berger's four examples are, respectively, 22, 63, 56 and 179. It could be argued that the PD plot of the last one, in his Figure 10, is too smooth and that a KDE with a smaller bandwidth would show the data better.

Error bars and confidence intervals

The standard error bars in Figure 3 can be regarded as simply displaying the size of the standard error of each estimate. But they could also be regarded as indicating confidence intervals for the true values. In that case they would be approximate 68% intervals rather than the more conventional 95%, or two-sigma, intervals.

While this may be of some use, it is extremely hard to compare several confidence intervals of differing lengths, both visually and logically. One may be tempted to infer “significant or not” differences from seeing whether intervals overlap, though of course that would not be correct. There is no easy way to interpret a number of univariate confidence intervals together; in principle, a multivariate confidence region is required.

A widely recognised disadvantage of confidence interval plots is that the least precise estimates have the longest intervals and tend to dominate the space on the graph. Sometimes they can be so dense that it is counter-productive to draw them all and it would be better to try to find another method of displaying precisions.

Very often the standard errors increase with dose, which makes it still harder to compare them. In such cases it may be clearer to plot doses (and intervals) on a log scale. Another aspect of this is that a symmetric interval on the log D_e scale will not transform to a symmetric interval on the D_e scale. That is, the symmetric approximate 95% interval $\log(y_i) \pm 2r_i$ for the true log dose corresponds to the non-symmetric interval $y_i \exp(\pm 2r_i)$ for the true dose, which may differ somewhat from the symmetric interval $y_i \pm 2s_i$, where y_i denotes an observed D_e value and r_i and s_i are its relative and absolute standard error, respectively. If we really are regarding the error bars as confidence intervals, then some thought should perhaps be given to how they are best defined and displayed. For example, if the estimation errors were thought to be essentially multiplicative, then it would make more sense to construct symmetric confidence intervals on the log scale.

Weighted means and selected data

In his examples, Berger (2010a) suggested that instead of using a minimum age model, a sufficient estimate of the burial dose can sometimes more simply be obtained from a weighted average of the D_e values for a selected subset of grains; and he further suggested that sometimes using a weighted average of log(D_e) values is better. However, whatever form of average might be used, the crucial questions here are: (a) which subset of grains should be selected? and (b) what are the bias and variance of the resulting estimate?

With respect to (a), many possibilities spring to mind. Among the more sensible would be methods that tried to select the complete group of "youngest" grains whose observed D_e values were consistent having a common burial dose, taking into account estimation error and natural variation between true doses with the same burial history. Ideally, one hopes to select all of the fully bleached grains and no others, though this is usually not possible — see my earlier comments with respect to Figure 2. There will nearly always be partially bleached grains having observed D_e values consistent with those for well bleached grains.

Galbraith (2010) noted that there is no good rationale for choosing the grains whose D_e values are close to the mode of a PD plot (i.e. choosing them because of this) even though that may sometimes produce an estimate close to the correct value. It might be more reliable to choose them by looking at the radial plot, which would at least make it easier to account for their differing precisions. But however you choose them you are bound, except in rare cases, to either include some partly-bleached grains or exclude some well-bleached ones.

With respect to (b), Galbraith (2010) noted that selecting grains with the lowest doses and treating them as if they were properly representative of well-bleached grains leads to biased estimates, sometimes grossly biased. For example, you can imagine that if you tried to be conservative and selected *only* the grains with the very lowest observed D_e values then you are likely to omit some higher values from well-bleached grains and end up with an under-estimate of the burial dose. Many of the lowest observed values will be low because their estimation errors are negative, so they will be lower than the corresponding true values. Hence, such a subset would be biased towards grains whose observed D_e values are lower than the true values.

Furthermore it is not correct to apply the usual standard error formula for an estimate obtained from a sub-sample that has been selected on the basis of the observed D_e values. This would not be an independent sample in its own right and allowance would need to be made for the effect of the selection. Calculation of a valid standard error is difficult for an objectively selected sample and impossible for a subjectively selected one.

Berger (2010a) rightly noted that such estimates are less reliable than those based on the more formal minimum age models. The latter treat the problem as one of extracting a specific component from a mixture. As such, they do not attempt to select a subset of grains at all, but rather they assign to each grain a probability of belonging to the well-bleached component.

On the subject of weighted averages and combining data generally, I recommend the encyclopedia entry by Cox (1982). This is a lucid and insightful article from a high authority.

A note on Sircombe and Hazelton (2004)

An interesting paper by Sircombe and Hazelton (2004), cited by Berger (2010a), adds some further theoretical insight to the question of estimating a frequency distribution from observations measured with error. It is concerned with detrital zircon ages obtained by U-Pb dating. It considers data y_i generated by the equation

$$y_i = x_i + e_i$$

where x_i is sampled from a distribution with pdf $f(x)$ and e_i is randomly drawn from a normal distribution with mean 0 and known standard deviation s_i. Like Galbraith (2010), it discusses how difficult it is to estimate $f(x)$. It then considers two samples of data and proposes a way of measuring the dis-similarity of their two different $f(x)$s without explicitly estimating either of them. Interested readers might like to look at its Figures 1 and 3. The former shows two different true $f(x)$s that have the same observed distribution (when errors are added to the x_is) and the latter shows how their method can nevertheless distinguish between them. Particularly illuminating is the way the standard deviations s_i are used; this is very different from how they are used in a PD plot.

In addition, Figure 5 of that paper shows ten detrital zircon age distribution plots obtained by "summing individual Gaussian distributions" which, as far as I can see, are what we are calling PD plots. They are presented simply to show the ten samples together in a small space so that they can be plotted against measures of dis-similarity between pairs of their underlying $f(x)$s. No inferences about $f(x)$ are made from these plots — indeed the authors explicitly say they are displaying the estimation errors as well as the age variation. That figure is undoubtedly informative, mainly because the estimation errors are small (in some cases very small) compared with differences between the single grain ages. These PD plots are very different from those normally encountered with OSL D_e data. Nevertheless, the error variation, though mostly relatively small, is still confounded with the age variation.

Summary

Frequency distributions of OSL equivalent doses are hard to understand, even when D_e is measured accurately, because they will reflect sampling, experimental and observational effects in addition to the key features of scientific interest. Histograms and kernel density estimates of D_e values are hard to interpret.

The "alternate form of probability-distribution plot" proposed by Berger (2010a) does not represent a proper probability distribution because of an incorrect transformation from the log scale. If it were plotted on a log scale it would be a PD plot in the sense of Galbraith (2010) — but using log(D_e) values and relative standard errors, rather than D_e values and absolute standard errors, as is implied in the abstract of Berger (2010a). If it were correctly transformed to the linear D_e scale it would often not differ greatly from a PD plot directly constructed on that scale. Berger's interpretations of his TPD graphs are based on a misunderstanding of what he plotted and his conclusion that they can "reveal meaningful relative structure in D_e distributions" (Berger 2010a, p.19) is not justified.

Galbraith (1988) and Galbraith (2010) discussed PD plots in the context of fission track ages and OSL equivalent doses, respectively. Such plots do not have a sound statistical basis and they have often been mis-interpreted in the literature. Berger (2010a) noted that PD plots have been criticised but did not recognise the substantive criticisms in those papers. PD plots have sometimes been used as an aid to selecting subsets of grains from which a weighted mean dose is calculated. This is not a reliable practice for the reasons given above.

Acknowledgments

I thank Bert Roberts, Jane Galbraith and Glenn Berger for comments on an earlier draft of this article.

References

Arnold, L.J., Roberts, R.G., Galbraith, R.F., DeLong, S.B. (2009) A revised burial dose estimation procedure for optical dating of young and modern -age sediments. *Quaternary Geochronology* **4**, 306–325.

Berger, G.W. (2010a) An alternate form of probability distribution plot for D_e values. *Ancient TL* **28**, 11–21.

Berger, G.W. (2010b) Errata: An alternate form of probability distribution plot for D_e values. *Ancient TL* **28**, 81.

Cox, D.R. (1982) Combination of Data, In *Encyclopedia of Statistical Sciences* (Vol 2, pp 45–52), Eds S. Kotz and N.L. Johnson, New York, Wiley.

Galbraith, R.F. (1988) Graphical display of estimates having differing standard errors. *Technometrics* **30**, 271–281.

Galbraith, R.F. (1998) The trouble with probability density plots of fission track ages. *Radiation Measurements* **29**, 125–131.

Galbraith, R. (2010) On plotting OSL equivalent doses. *Ancient TL* **28**, 1–9.

Olley, J.M., Pietsch, T., Roberts, R.G. (2004) Optical dating of Holocene sediments from a variety of geomorphic settings using single grains of quartz. *Geomorphology* **60**, 337–358.

Sircombe, K.M., Hazelton, M.L. (2004) Comparison of detrital age distributions by kernel functional estimation. *Sedimentary Geology*,**171**, 91–111.

Response to Galbraith

G.W. Berger

Desert Research Institute, 2215 Raggio Parkway, Reno, NV 89512, USA.

(Received 23 May 2011)

Introduction

Rex Galbraith and I exchanged many e-mails in 2010 concerning my note on so-called TPD plots (Berger, 2010a). As a result I submitted an Erratum (Berger, 2010b) which I think clarifies Berger's note succinctly. That e-mail exchange, as does his recent Comment (Galbraith, 2011), draws out an essential 'philosophical' difference. Perhaps the best way to summarize this difference is to say that I am concerned with "empirical distributions" and how to use visual representations of single-grain paleodose (D_e) estimates as accessible guides to choices of usefully accurate calculations of 'mean' D_e values, whereas he is describing the same 'elephant' from a statistically idealistic viewpoint. Clues to this idealism are provided by the frequent use in Galbraith (2011) of ill-defined (with respect to single-grain D_e data) words such as: 'true', 'useful', 'less informative', 'more informative', 'less meaning', 'less convincing', 'resolved', 'actual', etc. In the context, these words misrepresent the pragmatic message of Berger (2010a, 2010b). While I appreciate his current note as an attempt to educate the reader on the theoretical nuances of the statistical handling of single-grain D_e distributions and their embedded uncertainty estimates, and calculations of weighted means, for some of the reasons outlines above, the Comment (Galbraith, 2011) compels some reply herein.

One of the outcomes of our e-mail exchange was my request that he provide to the community of OSL users a software or spreadsheet 'program' for the ready computation of KDE plots such as shown in Figure 3 of Galbraith (2011). Presumably such plots can be generated (with effort by a novice) from the R statistical package, but most of us don't use that package routinely (I employ it for Arnold's unlogged MAM code: Arnold and Roberts, 2009), if at all. Another request was for dissemination of a software package for generating reasonably high-resolution radial plots (e.g., pdf files, rather than clipboard copies) via a user-friendly interface (GUI) that handles both linear and log D_e scales. He has not supplied that to me. In this context, the Comment of Galbraith (2011) could have been more helpful. The radial-plot software available from John Olley has an excellent GUI but it does not permit use of linear D_e scales, and creates only clipboard images of the plots. The radial-plot software from Vermeesch (2009) does permit creating high-resolution plots (saved as pdf files) and use of linear scales, but lacks many desirable user-selectable options (e.g. choosing centers of $\pm 2\sigma$ bands and band fills) that the Olley package offers.

In addition to these general comments, I have some comments to make on specific sections of Galbraith (2011), under his topic headings.

Introduction

Here he states that Berger (2010a) "does not recognize the more fundamental problems noted in Galbraith (1998) and Galbraith (2010)". This is incorrect and misrepresentative. Berger (2010a) did not attempt a fundamental discussion of the underlying principles expounded in these citations. How can that lacuna then demonstrate a lack of recognition?

Transforming a probability density function

The Erratum (Berger, 2010b) makes it clear that the areas under the peaks of the TPD plot cannot be used as indicators of relative probability, thus much of this section of Galbraith (2011) is redundant.

What is Berger's TPD plot?

There appears to be a logical inconsistency, in that (implied in Galbraith, 2011) it is permissible to adjust bin-widths to construct histograms of data points ('univariate estimates', if you will) lacking equal uncertainties, but it is not permissible to create a visual plot (TPD) free of such forced 'bandwidth' choices (unless once thinks the choice of a Gaussian is 'arbitrary'). It is misleading to state (Galbraith, 2011) that "the role of the relative standard errors of D_e" is "arbitrary". Also, what does Galbraith (2011) mean by "clear-cut interpretation" in the statement that the "TPD plot has no clear-cut interpretation"? Does he mean 'statistically idealistic', or 'empirically pragmatic'?

Log transformations (or not) in radial plots

This section is unnecessary because Berger (2010b) clarified that issue, concisely.

Why are radial plots more informative?

Berger (2010a) gave examples (and the literature has many more) where a radial plot is essential, but Galbraith (2011) repeatedly uses idealistic words such as "completely explained", or "more informative" to imply that other plots are useless. Also, the word "enough" in the phrase "or that two is enough" in paragraph 4 or 5 (depending on what is counted as a paragraph) has a different meaning for a pragmatic geochronologist than for an idealistic statistician. In Figure 9a of Galbraith (1988) (which the reader should read) and Figure 2 of Berger (2010a), I continue to see that the PD plot illustrates the presence of two modes and that this (existence of two modes) is the most parsimonious view of that data distribution. Of course, in a real data set one can easily violate Occam's razor and conceive of many embedded components, but what would be the (geological) meaning of that? In the next paragraph, Galbraith (2011) makes remarks about overlapping data points that apply equally to a PD plot (if one plots the data points and uncertainties with this plot).

Empirical distributions and kernel density estimates
The word "incorrectly" is a statistical usage, whereas in dating practice, these nuances in Galbraith (2011) are likely to be largely of a secondary or tertiary concern, because one often is (or should be) comparing OSL age estimates with numerical or stratigraphic age estimates obtained from other methods. Again, it would have been more helpful if user-friendly code or standalone software were provided for generating such KDE plots (with all their subjectivity). The rest of this section seems to imply that with real data within single-grain D_e distributions (rather than with statistically idealistic data points) every bump and wiggle should be resolved or would be informative (informative of what?). In dating practice, as stated implicitly (if not explicitly) by Berger (2010a), many single-grain D_e data points obtained from non-eolian deposits have no geological meaning. Generally, if non-eolian (not uniformly bleached optically) samples are collected carefully (this topic is addressed below), only the lowest D_e values would have meaning (last daylight exposure), unless there are stratigraphic indicators that a specific multi-depositional history could be preserved, or in carbonate-bearing deposits, evidence of significant β micro-dosimetry. An example would be provided by buried soil horizons, in which case the lowest D_e values might not relate to the main depositional process or event.

In other words, in the Figure 3 of Galbraith (2011) and in many published examples of single-grain D_e distributions, it is not important to 'resolve' (whatever that word might mean to readers) minor clusters of D_e values above the lowest 'age' group. Of course, deviations caused by unrecognized or uncorrectable effects of β micro-dosimetry fold into interpretations in some cases. Further folded into the generation of 'under-estimates' of 'true' single-grain D_e values are the effects of careless sample collection when deposits are heterogeneous. For example, as stated (Berger, 2010a), the use of brute-force tube or pipe sampling may introduce (no one has investigated this effect, to my knowledge) 'too-young' grains (from the sediment face) into the interior of the sample. There are published examples (e.g. supporting online material of Jacobs et al., 2008) of single-grain D_e distributions where the authors are motivated (by stratigraphic or archaeological evidence) to employ a central-age or finite-mixture model and to dismiss widely discordant 'too-young' D_e data points that may be artifacts largely of the sample collection method.

Error bars and confidence intervals
It is not "extremely hard" for me (and presumably most practicing geochronologists) "to compare several confidence intervals of differing lengths, both visually and logically". Also, why is it "counter-productive" to draw confidence intervals? Counter-productive to what? ...to prediction of certain statistical parameters, or to age estimation from empirical data?

Weighted means and selected data
There are several points of disagreement, and I think that merely citing a few standard books (e.g. Bevington and Robinson, 1982; Moroney, 1965; Topping, 1962) on treatment of uncertainties would have sufficed. However, I fail to understand parts of the fourth paragraph. For example, in a geological sense, how can there be "some higher values from well-bleached grains" (apart from Gaussian or other probability effects) if these values indeed have been well-bleached and share the same β micro-dosimetry, unless one considers such and other physical effects, which Galbraith (2011) does not mention? Also, negative D_e values more than one (estimated) standard deviation below zero are possible if one accepts Gaussian probability in the measurement of D_e values close to (and above) zero. Much of that paragraph's argument is hypothetical conjecture: statistical idealism disconnected from empirical settings.

Summary
I disagree with the final sentence in Galbraith (2011): "This" selection of subsets of data points "is not a reliable practice...". Galbraith's definition of 'reliable' apparently is not mine. There are several examples in the single-grain OSL dating literature where selection of subsets provides usefully accurate (geologically, stratigraphically) age estimates. Of course, there are several examples in such literature where selection of subsets by use of visualization plots is too subjective (see some examples in Berger, 2010a) to be useful, and one must resort to more refined statistical calculation schemes (e.g., minimum-age, central-age models, Galbraith et al., 1999) than use of weighted means.

References

Arnold, L. J., Roberts, R. G. (2009) Stochastic modeling of multigrain equivalent dose (D_e) distributions: Implications for OSL dating of sediment mixtures. *Quaternary Geochronology* **4**, 204-230.

Berger, G.W. (2010a) An alternate form of probability distribution plot for D_e values. *Ancient TL* **28**, 11-21.

Berger, G.W. (2010b) Errata: An alternate form of probability distribution plot for D_e values. *Ancient TL* **28**, 81.

Bevington, P.R., Robinson, D.K. (2003) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw Hill, New York.

Galbraith, R.F. (1988) Graphical display of estimates having differing standard errors. *Technometrics* **30**, 271-281.

Galbraith, R. (2010) On plotting OSL equivalent doses. *Ancient TL* **28**, 1-9.

Galbraith, R. (2011) Some comments arising from Berger (2010). *Ancient TL* **29**, 41-47.

Galbraith, R.F., Roberts, R.G., Laslett, G.M., Yoshida, H., Olley, J.M. (1999) Optical dating of single and multiple grains of quartz from Jinmium rock shelter, northern Australia: part I, experimental design and statistical models. *Archaeometry* **41**, 339-364.

Jacobs, Z., Roberts, R. G., Galbraith, R. F., Deacon, H. J., Grün, R., Mackay, A. W., Mitchell, P., Vogelsang, R., Wadley, L. (2008) Ages for the Middle Stone Age of Southern Africa: Implications for human behavior and dispersal. *Science* **322**, 733-735.

Moroney, M.J. (1965) *Facts from Figures*. Penguin Books, London, UK.

Topping, J. (1962) *Errors of Observation and their Treatment*. Chapman and Hall, London, UK.

Vermeesch, P. (2009) Radial Plotter: A Java application for fission track, luminescence and other radial plots. *Radiation Measurements* **44**, 409-410.

Message from the Editor

Following the long discourse that has ensued in the last two issues of Ancient TL, the Editorial Board has decided to clarify the maximum length of Letters and Replies. The purpose of this is not to stifle discussion, but rather to ensure that readers are able to clearly follow the line of argument arising from the original article. In the future, Letters to Ancient TL will be limited to a maximum of two printed pages, including diagrams, tables and references (equivalent to about 1400 words of text). Replies will have the same limit.

G.A.T. Duller