On plotting OSL equivalent doses

R. Galbraith

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

(Received 27 November 2009; in final form 18 December 2009)

Abstract

This article is motivated by some recent discussion of the use of so-called "probability density" plots of OSL equivalent doses. Such graphs are not advocated in the statistics literature. I try to explain what they are doing, why they are easy to mis-interpret and why they are not to be recommended. I include discussion of the meaning of dose frequency distributions, statistical research on the problem of estimating a frequency distribution when observations from it have added errors, and the possible role of dose histograms, in addition to radial plots, as data displays.

Introduction

There has been some recent discussion of the use of so-called "probability density" (PD) plots for displaying single grain, or single aliquot, OSL equivalent doses, and it was suggested to me that I might contribute to this. PD plots are used quite widely, as can be seen by perusing articles to be published in *Quaternary Geochronology* arising from the 12th International Luminescence and Electron Spin Resonance dating conference. Some years ago they were used by the fission track community to display single grain fission track ages. I criticised them then on several grounds: they do not estimate the true age distribution, modes in a PD plot do not necessarily correspond to discrete component ages, they obscure good information by combining it with bad, and their reliability was untested (Galbraith, 1998). They have been largely abandoned by the fission track community — I suspect mainly because they have not been found useful in practice.

In principle those criticisms also apply to OSL equivalent doses, though the popularity of PD plots here suggests that some people do consider them to be useful. However, they do not appear to have been advocated in the statistics literature. In this article I will try to explain what I think PD plots are doing, why they are difficult to interpret, and what alternatives there might be. Some of these ideas are also in a book chapter (Roberts and Galbraith, in press) which is to appear, though it was originally written in 2006.

What are the data?

We have a set of *bivariate* observations — an equivalent dose and its standard error for each of *n* quartz grains or aliquots, where *n* might be as low as 20 or 30 or as high as several hundred. A general feature of such data is that both the observed doses and their standard errors vary. Usually they vary together, with a higher standard error associated with a higher dose, the main exception to this being when the observed doses are close to zero.

A natural candidate for a graph is therefore some sort of bivariate plot; and a particularly useful one is a radial plot, which most readers will be familiar with. Descriptions of this method can be found in Galbraith et al. (1999), Galbraith (2005), Roberts and Galbraith (in press) and in other references cited there, so I will not deal with them further here. It is worth emphasising, though, that radial plots have optimal statistical properties (Galbraith, 1988) — they display the data as informatively as is possible and without distortion. They have also been found to be powerful in practice and can reveal features not otherwise apparent. Regardless of what other plots are also made, I would recommend researchers to look carefully at a radial plot of their equivalent doses.

A radial plot, though, does not provide an explicit picture of the *frequency distribution* of equivalent doses, which is perhaps why researchers may want to see some sort of frequency curve. However, we need to think carefully about what the frequency distribution of doses is and whether it has a useful scientific meaning. Sometimes it does not, and it is a strength of the radial plot that it does not force this interpretation on the reader.

Dose frequency distributions

The information in a relative frequency distribution of observed equivalent doses is complicated. It contains mixtures of received doses, natural variation and estimation errors — some of which are multiplicative and some additive — and does not simply relate to the relative numbers of grains in some real population that have received each possible dose. It is much more complicated than, for example, a frequency distribution of heights of men or weights of babies.

Consider a hypothetical situation where we have a sample of single grain equivalent doses from a field sample of quartz that have been measured essentially without error (i.e., with negligible standard errors). The doses received in nature may differ between grains for various reasons, such as differing burial history or partial bleaching. Furthermore, even if each grain had experienced the same radiation dose in nature, the measured doses (even though measured exactly) would vary because of natural variation in luminescence properties between grains. Different scenarios will typically produce different dose distributions — for example unimodal distributions with low dispersion (perhaps representing only natural variation in luminescence), mixtures of two or three such component distributions, or highly heterogeneous, asymmetric or multimodal distributions.

What would knowing the shape of the frequency distribution of the doses tell us?

If we are lucky, it might indicate the type of sample or scenario we have. But before going further, there is another question: does this frequency distribution represent a natural phenomenon or is it largely a result of the process of grain selection and measurement? In the latter case it may be of more limited interest, and possible inferences from the data may also be more limited.

For example, suppose that each grain in our sample had essentially experienced one of two alternative burial histories, so each had received one of two radiation doses (e.g., by mixing of grains from two juxtaposed sedimentary strata that differ significantly in age). We could fit a two component mixture to estimate those doses. But would the estimated *mixing proportions* reflect anything other than artifacts of the experimental procedure, particularly grain selection? After all, only a small fraction of grains in a sample actually produce a measurable luminescence signal and these could be a highly non-random subset. Nevertheless, the component doses themselves should still be meaningful. The same applies to mixtures of more than two components — what do the mixing proportions represent? And by extension, what do the relative frequencies of different doses represent? In particular, does the most frequent dose in a sample have any special scientific relevance or meaning? These are questions for practitioners. The answer to the last one may sometimes be yes and sometimes no.

For aliquots comprising several grains, the concept of a dose frequency distribution is more complicated. It makes some sense if all grains in the same aliquot have experienced the same burial history. Then any differences in “true” single grain doses within an aliquot (had they been observed) would presumably just be due to differences in luminescence properties, and the aliquot equivalent dose would be representative of the burial history. But if grains in the same aliquot had different burial doses, or had experienced different amounts of partial bleaching, then the aliquot dose distribution would be much harder to interpret.

Incidentally, I have seen researchers fit finite mixtures (say with two or three component doses) and then choose the dose with the largest mixing proportion to be the relevant one. This seems like bad logic, especially for samples composed of partially bleached sediments. The relevant dose might be that corresponding to the youngest grains, and these could easily be a minority of the sample. This type of reasoning arises when looking at humps and bumps in frequency distributions too.

Histograms and kernel density estimates

Continuing with the case where our equivalent doses are measured without error, suppose that we *do* want a picture of the shape of the dose distribution. This could be provided (if there were enough grains) by a well-drawn histogram, which is essentially a graph of relative numbers of grains falling into different dose intervals (bins). Histograms are of course very familiar and widely used. A possible alternative is a kernel density estimate (KDE). This is a continuous curve that is an estimate of the probability density function (assumed to be continuous) of the distribution that the observations are supposedly a random sample from.

Denote the sample of true doses by $x_1,x_2,\ldots,x_n$ and imagine that they were drawn randomly from a distribution with probability density function $f(x)$. Now think of a histogram of these with equal bin widths. For a large enough sample, and small enough bin widths, this will give an idea of the shape of $f(x)$. The area of each rectangle, and in this case also its height, is proportional to the number of observations falling in that bin, and (suitably scaled) is an estimate of the relative numbers in that interval in the population.

Now imagine drawing a histogram by starting with a bin at the extreme left (with no data in it) and sliding that bin continuously along the x scale. At each value of x draw a point at height equal to the number of data values in the bin centered at x. The points will trace out a curve that goes through the top middle points of the histogram rectangles plus more in between. That curve is a kernel density estimate of *f(x)* — in this case, using a rectangular "window". If you increase the bin width the curve will be smoother but may lose shape features, and if you decrease the bin width the curve will resolve more shape features but be more erratic. Choice of bin width is a compromise between these two.

Rather than using a rectangular window, many kernel density estimates use a Gaussian window, which does not have discontinuities at the ends. The curve you then get is equivalent to drawing a Gaussian probability density function centered at each data value (each with the same standard deviation *b* which is chosen by you) and then summing them point-wise. This is simply a data smoothing method — as is counting up numbers in a histogram bin — there is no probability interpretation of this Gaussian window.

The quantity *b* is called the *bandwidth* of the window, and is analogous to the bin width of the histogram: the larger *b* is, the smoother the curve but the less resolution in shape there is. Actually there are many types of window around — nowadays they are called kernel functions — including triangular and cosine, but the principle is the same.

Statisticians have found that the shape of the window does not make much difference to the shape of the density estimate. What really matters is the bandwidth, which is a compromise between how much smoothing and how much resolution in shape you want. Choice of bandwidth usually depends on the sample size, with smaller bandwidths used for larger samples. This is like choosing the bin sizes for a histogram. Note that any smoothing distorts the data and loses information. A kernel density estimate is always a biased estimate of *f(x)* and in statistical terms the choice of bandwidth is a compromise between reducing bias and reducing variance. There is some theory about how to choose a bandwidth in order, for example, to minimise mean squared error (which is variance plus squared bias). In general large samples are needed to get reasonably informative kernel density estimates.

As estimates of density functions, KDEs enjoy some theoretical advantages over histograms (Wand and Jones, 1995, p5). The main disadvantage of histograms in this regard is that their shape can depend on where the first bin starts as well as on the bin width. Being continuous, KDEs give an impression of high precision, even for small sample sizes — but often a spurious impression. They have been developed by statisticians for over 50 years and are a useful exploratory tool, but they are not often used to present scientific data. One reason, I think, is that a histogram is better for this purpose. A histogram explicitly displays *proportions* of observations in various intervals as *areas*, whereas a KDE displays relative frequencies as a continuous curve. A KDE does not so easily lend itself to visual comparisons or simple calculation; it emphasises humps and bumps in the frequency curve, many of which have no significance; and it hides information, particularly relating to sample size and variability. As a general-purpose graph, a histogram is nearer to the raw data, easier to use and more convincing.

PD plots

Now let us return to the situation where the standard errors are non-negligible and variable. Denote the observed doses and their standard errors for *n* grains by (y_i, s_i) for i =1,2,...,n.

A PD plot is constructed by replacing each y_i with a Gaussian probability density function centered at y_i and having standard deviation s_i, and then adding these point-wise to obtain a continuous curve. Its construction is similar to that of a KDE, but with a different kernel function (with a different bandwidth) for each observation. The plot has some intuitive sense: you can think of it as plotting for each candidate dose, the "popularity" of that dose, as voted for by the *n* grains in the sample, where each grain spreads its vote (unequally) over several neighbouring doses, with more uncertain grains voting for a wider range of doses. Note that popularity comes both from frequency (y_is close together) and precision (small s_i). Does a particular dose have any special scientific meaning simply because it is measured with high precision? Surely not.

The name "probability density plot" suggests that it is a plot of a probability density. An immediate question is: what probability density? The answer is: that of an equal mixture of *n* Gaussian distributions, where the *i*th component has mean y_i and standard deviation s_i. In other words, a PD plot is plotting the probability density function of a random variable *z* constructed as follows: choose one of the *n* y_is at random and add to it a Gaussian random error with standard deviation s_i.

A second question is whether the random variable *z* (and its associated distribution) is of any interest. To understand this, it is useful to think of a simple statistical model. Suppose that for each given s_i, the observed dose y_i is generated by the equation

$$y_i = x_i + e_i \qquad (1)$$

where x_i is randomly drawn from a distribution with probability density function $f(x)$ and e_i is randomly drawn from a Gaussian distribution with mean 0 and standard deviation s_i. Intuitively, x_i represents the "true" dose (i.e., measured without error) for grain i and e_i is the error in estimating x_i (i.e., the difference between y_i and x_i). Neither x_i nor e_i is observed. The function $f(x)$ is unknown and our aim is to estimate it, or at least some of its features.

This type of model is familiar. If we postulated a parametric form for $f(x)$, such as Gaussian, we would have a version of the central age model. But here we are trying to let the data tell us something about $f(x)$ without assuming a specific form. In the previous section we were essentially thinking about how to do this if we could directly observe the x_is.

Under this model, we can now think of obtaining a value of z by first choosing one of the n x_is at random, adding a random e_i to it to get y_i, and then adding *another* Gaussian random error to y_i to get z. So the distribution of z (i.e., the PD plot) does depend on the n x_is, which have been sampled from $f(x)$. But it also depends on the n s_is — doubly so because two independent random errors, each with standard deviation s_i, have been added to x_i. Its usefulness in practice will depend on whether it provides recognisable and useful information about $f(x)$.

There is a conspicuous lack of published theory about this. I've never seen a proper statistical study of PD plots, or even a reference to such a study — indeed I have never seen them advocated in a statistics journal. But there is some published research in statistics journals on how to estimate $f(x)$. One result of this is that the data (y_i, s_i) in general contain very little information about the shape of $f(x)$. This is a warning against giving much credence to locations and relative heights of peaks in *any* estimate of $f(x)$. I summarise this research below.

My own experience from looking at PD plots, both with real and with simulated data, is that they are not uninformative but nor are they very informative, and their shape can be greatly affected by the s_is. If we observed the same doses, but with different precisions, the curve can look very different. Often s_i tends to increase with y_i. This alone will tend to produce a highly positively skewed curve with the highest peak or peaks near the left hand (lower dose) end. That is, one can often guess its general shape even without seeing any data. In general a high peak will be partly a result of several y_is being close together but partly also a result of s_is being relatively small. Conversely, if there are a substantial number of low-precision (large s_i) grains in the sample, as there often are, these will tend to smooth out the whole curve and dilute the information from the high precision grains. Examples of these effects in the context of fission track ages can be seen in Galbraith (1998).

The force of these effects will of course be less if all or nearly all of the s_is are small compared with differences between y_is. In that case the distribution of y_is will not differ greatly from that of the x_is and a PD plot may be similar to a kernel density estimate (based on the x_is) having the same average bandwidth.

In the above model s_i is unrelated to x_i. But usually in practice the standard error tends to increase with dose. Sometimes the *relative* standard error is approximately independent of the dose. Then equation (1) would apply better with (y_i,s_i) equal to the estimate and standard error of the natural log of the dose. But a PD plot of *log* doses would look very different in shape, and may have different numbers, locations and relative heights of peaks, compared to using a linear dose scale. Which scale should be used and why?

Some pitfalls

I don't think I have ever seen a paper where the author presents a PD plot and then comments that its shape may be reflecting the differing estimation errors rather than how the equivalent doses vary. Nearly always it is interpreted, implicitly or explicitly, in terms of which doses are predominant or indicated. This is understandable, because the graph invites one to do this, but it is misleading. Here are examples of possible pitfalls.

- You draw a PD curve and find that it has a high peak near the left hand end at a dose that plausibly corresponds to the burial dose of that sample (perhaps inferred from other information). So you present the PD plot as if it were pointing to that as the burial dose, or as support for that value. There may well be some relation between where the highest peak occurs and the burial dose — it may even agree closely sometimes — but it is not a reliable one. Often the PD curve is likely to have its highest peak near the left end simply because of the scale on which it is drawn and the nature of the error distributions. Sometimes also the burial dose may be reflected in only a minority of grains, and may not appear as a peak at all.

- In the previous scenario you might argue that *in this case* the PD graph gives the "right" answer, so it is useful *here*. How do you know it is the right answer? Presumably from some other information. Then what use is the PD plot? It's not good enough that the location of a peak in a PD plot might sometimes agree with the burial dose. As Lewis Carroll famously wrote, a stopped clock is right twice every day. A better approach would be to say "The PD graph suggests such and such. How can I investigate that hypothesis more seriously?"

• You look at the grains sitting under a peak of the PD curve and use these to estimate the burial dose, or some dose of interest. Or likewise, you use grains under different peaks to estimate different mixture components and their standard errors. This is like a so-called "classification" method of estimating mixture components. Such methods are known to be biased — sometimes very biased — and to provide unreliable standard errors. Fortunately there are more reliable methods available, such as maximum likelihood estimation.

• Among lots of information and data in a paper are several PD plots, and a commentary that refers to these to support the discussion of some phenomenon or theory of interest. The proposed theory may well be right, but logic tells us that if a PD plot does not reliably estimate the true dose distribution (which it does not) then those graphs do not support the theory.

An important aspect of this is that even if the writer is able to avoid such pitfalls, it may still be hard for readers to do so.

How *can* we get a picture of f(x)?

Suppose we have the scenario given by equation (1) and we want to estimate the function *f(x)*. A PD plot will not do this, so how can we do it? There are two general statistical approaches: parametric and non-parametric. The central age and minimum age models are examples of parametric methods. These assume a specific form for *f(x)*, but with unknown parameters that represent quantities of interest which are then estimated from the data. The idea of using a non-parametric method is to see if the data can tell us what shape *f(x)* has without imposing a particular form.

There has been some research on this. An important general result is that the data (y_i,s_i) contain a very limited amount of information about the *shape* of *f(x)*, as opposed to its location and dispersion (e.g., Goutis, 1996; Madger and Zeger, 1996; Wand and Jones, 1995, p160). The same data can easily arise from quite different *f(x)*s.

Several methods have been suggested. One is the "non-parametric maximum likelihood estimate" (NPMLE). This turns out to yield a discrete probability distribution concentrated on a relatively small number of values — that is, it estimates *f(x)* as a set of *k* different values and their probabilities, where *k* is quite small compared with *n* (Laird 1978). When *f(x)* is assumed to be continuous it is arguable that it would be nice if the estimate of *f(x)* were also continuous. To this end Madger and Zeger (1996) proposed a smoothed version of the NPMLE (called SNPMLE). This assumes that *f(x)* is a mixture of *k* Gaussian distributions (where *k* is unknown) each having a standard deviation greater than or equal to some known value *b*. The thinking behind this is that you can produce a wide variety of different shapes by mixing enough Gaussian distributions in differing proportions. The condition on the standard deviations is necessary in order to guarantee convergence to a solution. You could call it a semi-parametric method. The SNPMLE converges to a mixture (with differing mixing proportions) where, again, *k* is relatively small and all components of this mixture have the *same* standard deviation, equal to *b*. The value of *b* is chosen empirically to achieve a desired amount of smoothing, like a bandwidth of a kernel density estimate — the larger *b* is, the smoother the graph. Other methods have been proposed by Goutis (1996) and Newton (2002). These methods are all computationally intensive to implement. More recent work includes Delaigle and Meister (2008), Staudenmeyer et al. (2008) and Wang et al. (submitted) so theoretical progress is being made in this area.

The general message seems to be: it is hard to infer the shape of an underlying distribution when observations from it have added errors, even when these errors have known standard deviations. A more fruitful approach might be to ask: what specific features of *f(x)* do we really want to know? Then try to ascertain these by appropriate statistical modelling.

Improving a histogram

To help interpret a histogram of single grain equivalent doses, Roberts and Galbraith (in press) suggest adding a scatter plot of s_i against y_i. This is illustrated in Figure 1 for a sample of 82 single quartz grains. Olley et al. (2004) reported that these grains were transported by wind on to the bed of Lake St. Mary (in semi-arid south-eastern Australia) within the last 40 years. Many of the observed equivalent doses are close to zero and some are negative.

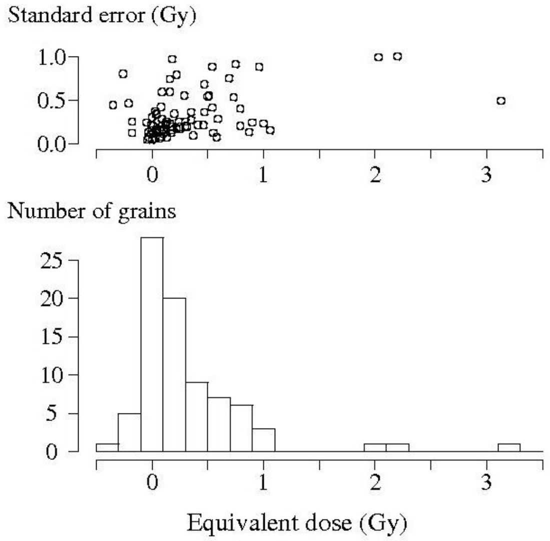

Figure 1: *Histogram and scatter plot of equivalent doses for 82 grains of aeolian quartz (sample SM15 from Olley et al., (2004).*

The histogram has a positively skewed shape for doses below 1 Gy and three more extreme values around 2 and 3 Gy. The scatter plot shows that several grains have standard errors greater than 0.5 Gy, which is quite large compared with differences between the dose estimates, and two of the extreme values have standard errors greater than 1 Gy, so could, in principle be consistent with the values for some of the lower dose grains. It must be emphasised that this graph is simply a plot of the raw data; the histogram in particular is a summary of the y_is and should not be interpreted as a graph of the x_is. The scatter plot helps with this by drawing attention to the s_i associated with each y_i. In fact the histogram is better viewed as an adjunct to the (y_i, s_i) scatter plot, showing the marginal distribution of y_i, rather than the other way round.

This example is presented here simply to illustrate the method. It is unusual in having several negative and near-zero equivalent doses; but their presence serves to remind us that the y_is are not the x_is (the true doses) but just *estimates* of them. For example, the smallest y_i is –0.35 Gy. Because x_i cannot be negative, we can deduce that this y_i underestimates its x_i by at least 0.35 Gy. The s_i for this grain is 0.45 Gy, indicating that its x_i could still be as large or larger than $-0.35 + 2 \times 0.45 = 0.55$ Gy. In general, a histogram of y_is need not look like a histogram of the corresponding x_is.

Graphs like Figure 1 were used, in conjunction with radial plots, by Arnold et al. (2009) to compare a number of samples of differing origin. We found the scatter plot useful for revealing the relationship between s_i and y_i. Sometimes there was a strong positive correlation, indicative of multiplicative errors, and other times there was little or no correlation (especially for small y_i, such as in Figure 1 here) suggesting that the main sources of error were additive. My co-authors also found the histograms useful for indicating some general characteristics of the sample.

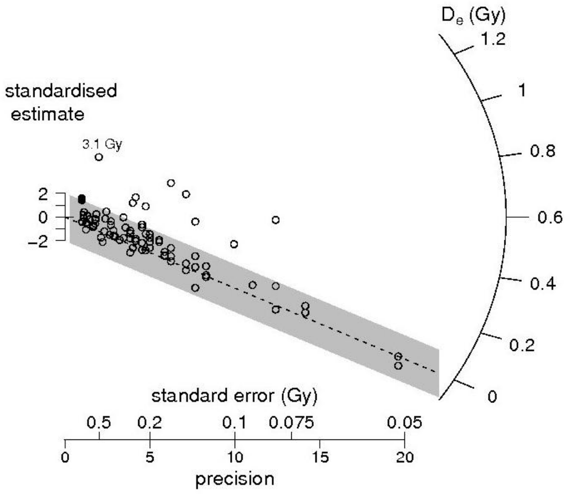

Figure 2. *Radial plot of the data in Figure 1. The two grains with very imprecise equivalent doses near 2 Gy are plotted as filled circles (the points almost coincide).*

Figure 2 shows a radial plot of the same data. This uses a linear dose scale rather than the usual log scale, the latter not being possible with negative estimates. It is not easy to draw this scale in such a way as to accommodate the three extreme values and at the same time to show the rest of the data in detail. Here I have drawn it so as to see the main data clearly; and radii through the three values greater 2 Gy go off the D$_e$ scale. The two values near 2 Gy are seen here to be almost completely uninformative — you hardly notice them — and the vast majority of points (all but about 10) are consistent with having zero dose.

In this example the information in the radial plot is so clear that very little further analysis is necessary. Possible further analysis might be to try to ascertain whether the burial dose is actually zero or some positive value close to zero and perhaps calculate an upper confidence limit for it. This would require proper statistical modelling. In fact Olley et al. (2004) estimated a burial dose of 0.1 Gy with a confidence interval that included 0 Gy.

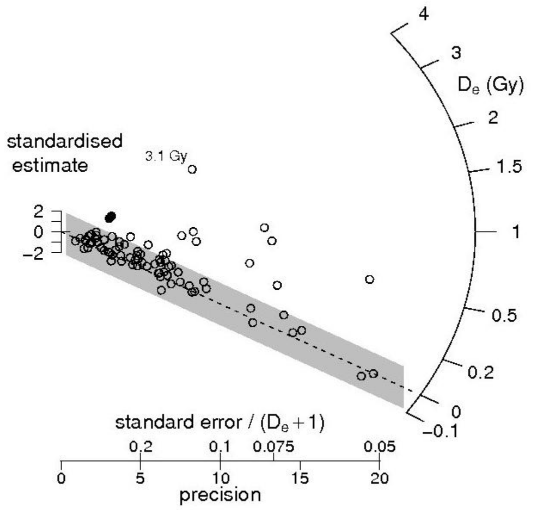

Figure 3. *An alternative radial plot of the data in Figure 1 using the modified log transformation $d = \log(D_e + 1)$. The two D_e values near 2 Gy are again plotted as filled circles.*

For data containing zero or negative estimates, an alternative to using a linear dose scale is to use a modified log transformation given by $d = \log(D_e + a)$ for some suitable a. That is, add a Gy to each observed dose and then take logs. The standard error of d is then approximately $\mathrm{se}(D_e)/(D_e + a)$. Figure 3 illustrates this method for $a = 1$. The dose scale is now non-linear, calculated from the formula $D_e = e^d - a$, and there is no difficulty in including the extreme values on it.

The message from Figure 3 is very similar to that from Figure 2. It looks a bit different because the estimates are plotted with respect to relative, rather than absolute, standard errors; in particular, the three extreme values are more prominent. This method is useful when the data contain some near zero doses and some larger ones.

Choice of bin width

A reviewer raised the question of what bin width to use for the histogram, particularly in relation to consistency of presentation and also whether the standard errors should be used to determine it.

In Figure 1 I have used bins of width 0.2 Gy, located so that 0 Gy comes in the middle of a bin (and consequently, so do 1, 2 and 3 Gy). A reasonable alternative would be to have 0 Gy at the edge of a bin. General guidelines tell us to use smaller bins for larger sample sizes and to try to achieve a reasonable amount of smoothing without losing too much detail, but there is no hard and fast rule. It is helpful to use friendly values; 0.2 Gy is better than 0.23 Gy, say. If I had used 0.1 Gy there would be twice the number of bins with smaller numbers in each, while 0.4 Gy would produce half the number of bins and more data pooled. Here 0.2 Gy seems about right. My personal preference is to err on the side of more bins rather than fewer, so as to reveal more of the raw data.

A histogram does not have to have equal width bins of course. For very highly skewed data it is sometimes suggested to have wider bins in the tail (drawn so that the area of a rectangle, not its height, is proportional to the number of observations in the bin). But equal bin widths are easier to understand and are nearly always used for routine presentation. In Figure 1 it is much better to show the three large values in separate bins rather than combining them into one long bin. Incidentally, if you look at these actual doses in the scatter plot you can see that they do not fall in the middle of each bin; the histogram just tells you that the points are somewhere in the bin, not necessarily in the middle.

What about consistency of presentation? In fission track analysis it is standard practice to measure about 100 track lengths and present them in a histogram with 1 micron bins on a scale that goes from 0 to 20 microns. This is possible because unannealed track lengths have a very tight distribution with mean about 16 microns and standard deviation about 1 micron. You never see a track longer than 20 microns. When tracks are heated they shorten and become more variable in length: the distribution shifts towards zero and becomes more dispersed and skewed. It tells us something about the thermal history that the grain has experienced. This consistency of presentation is a huge advantage and greatly outweighs other criteria for choosing bin widths. Many such histograms are shown in articles and at conferences and it is possible to compare them, not only within the same presentation, but also between different articles and talks, even in different journals and conferences.

Is such a thing possible for equivalent dose distributions? I don't think so. Samples may have doses ranging between, say, 10 and 80 Gy. Using 0.2 Gy bins there would do no smoothing at all. Using bins of width 4 or 5 Gy might be reasonable there but would be useless for the data in Figure 1. But there is some scope for consistency of style, including axis labels and terminology. This is a matter for general discussion. Sometimes it may be useful to plot equivalent doses on a log scale, which raises further questions about style. OSL equivalent doses are far more complicated than fission track lengths! They are more like fission track ages, but more complicated than them too. Fission track ages are routinely presented in radial plots but not in histograms.

Should the s_is be used to determine the bin width? No. The histogram is a graphical display of the observed doses (the y_is). The standard error s_i tells us something about how close y_i is likely to be to its x_i, but this has nothing to do with choosing the bin width for a histogram of y_is. If we had a larger sample size we would want smaller bins (regardless of the s_is) to get a better summary of the data.

This point serves to emphasise that a histogram of y_is is not the same as a histogram of the x_is and should not be seen as such. If the s_is are all small compared with differences between x_is then the two will be similar. If the s_is are non-negligible, then all of the previous discussion and theory is telling us that we just don't have much information about the frequency distribution of x_is. We have some information about its location and dispersion; which is what the central age model is extracting, and we can try to extract other information using parametric models such as the minimum age models. There are non-parametric methods for estimating this frequency distribution but they do not yield either PD plots or histograms of y_is.

Summary

When OSL equivalent doses are observed with non-negligible and differing standard errors they are not easy to compare. A radial plot will display them informatively and without distorting their message. I recommend looking at a radial plot in addition to any other graphs that might be made.

Research has shown that such data contain little information about the form of the underlying frequency distribution of true doses; quite different underlying distributions can easily give rise to the same observed data. Several methods have been suggested for trying to estimate such an underlying distribution, though little is known about how useful they are in practice. A question to consider is what use this frequency distribution would be if it were known. If only some of its features or parameters were of interest then a more fruitful approach might be to try to estimate these directly.

A histogram of observed doses will reflect features of the single grain error distributions and the relationship between observed doses and their standard errors, as well as variation in true doses. In order to interpret it without pitfall it is necessary to add further information, such as an adjacent scatter plot of standard errors against doses. Together these can provide a useful description of the data, but will typically not provide a true picture of the underlying dose distribution.

PD plots also depend on the error distributions and their relationship with dose — more so than histograms because effectively *two* independent errors are added to each true dose. There appears to be no rationale or justification for them in the statistics literature. They too do not provide an estimate of the underlying dose distribution. All you can really do with them is look and see where peaks occur. These may or may not reflect features of the true dose distribution, which in turn may or may not reflect events in nature.

Perhaps their biggest difficulty, though, is that it is hard to avoid the types of pitfalls mentioned above. The reader is faced with a continuous curve that looks meaningful; but it does not mean what it appears to mean and there is no reliable way to extract what we want from it. Someone once said that Wagner's music is better than it sounds. Indeed it may be. But PD plots are not as good as they look. I don't recommend them.

Acknowledgments

I am very grateful for useful comments and suggestions from Geoff Duller, Jane Galbraith, Bert Roberts and Ann Wintle.

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.

Delaigle, A., Meister, A. (2008) Density estimation with heteroscedistic error, *Bernoulli* **14**, 562–579.

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.F. (2005) *Statistics for Fission Track Analysis*. Chapman and Hall/CRC, Interdisciplinary Statistics Series, ISBN: 1-58488-355-5, 224pp.

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

Goutis, C. (1997) Non-parametric estimation of a mixing distribution via the kernel method.

Journal of the American Statistical Association **92**, 1445–1450.

Laird, N.M. (1978) Non-parametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**, 805–811.

Madger, L.S., Zeger, S.L. (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association* **91**, 1141–1151.

Newton, M. A. (2002). A nonparametric recursive estimator of the mixing distribution. *Sankhya Series A* **64**, 306–322.

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.

Roberts, R.G., Galbraith, R.F. (in press), Statistical aspects of equivalent dose and error calculation. In Krbetschek, M. (Ed.) *Luminescence Dating: An Introduction and Handbook*. Springer, Berlin.

Staudenmeyer, J., Ruppert, D., Buonaccorsi, J. D. (2008) Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association* **103**, 726–735.

Wand, M.P., Jones, M.C. (1995) *Kernel Smoothing*, Chapman and Hall, ISBN 0-412-55270-1, 212pp.

Wang, X-F, Sun, J,. Fan Z. (submitted) Deconvolution density estimation with heteroscedastic errors using SIMEX. *Electronic Journal of Statistics* arXiv:0902.2117v1[mathST]

Reviewer
A.G. Wintle