ESTIMATORS OF PROBABILITY DENSITY AND THEIR UNCERTAINTIES

Manuscript submitted to the American Journal of Science

(this paper can also be downloaded from `http://pangea.stanford.edu/research/noble/epdu`)

July 21, 2004

Pieter Vermeesch

Department of Geological and Environmental Sciences, Stanford University

Braun Hall, room 320-305, 450 Serra Mall, Stanford, CA 94305.

tel: (650) 723 1010, e-mail: pvermees@pangea.stanford.edu

ABSTRACT. Histograms and kernel methods are popular ways for estimating probability density. There are two sources of uncertainty that affect such estimates: counting statistics and measurement uncertainties. For the histogram, a Bayesian method is proposed to estimate the effect of counting statistics by calculating simultaneous confidence bands. Such Bayesian credibility bands have non-zero width, even for empty histogram bins. However, the Bayesian method is not easily applicable to kernel density estimators. Therefore, it is not easy to calculate the effect of counting statistics on the latter. On the other hand, the histogram does not take into account measurement uncertainties, while the kernel density estimator does. The kernel-histogram density estimator is proposed as a hybrid solution, taking into account both counting statistics and measurement uncertainties. The methods for the construction of simultaneous confidence bands for density estimators are easily extended to estimators of cumulative probability.

INTRODUCTION

When reporting experimental data, it is good scientific practice to also give an estimate of their analytical uncertainty. More often than not, this is done by assuming that the uncertainties are normally distributed, and drawing one or two standard deviations wide error bars around their mean. In certain subdisciplines of the Earth Sciences such as petrography and detrital geochronology, the relevant experimental data are not simple measurements, but estimates of their probability density such as histograms. Significant uncertainty can be associated with this form of data, but this uncertainty is rarely or never reported. Howarth (1998) discusses the construction of confidence intervals for individual proportions. However, one is rarely interested in just a single proportion. This paper discusses and compares several ways to compute *simultaneous* confidence bands on both histograms and kernel density estimates, as well as their cumulative equivalents.

METHODS FOR ESTIMATING PROBABILITY DENSITY

Histograms

The histogram is a stepwise estimator of probability density. If $\mathcal{X}_N$ = {X$_1$,...,X$_N$} is a sample from an unknown distribution f(x), the histogram is defined as (Scott, 1992):

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} I_{[t_j, t_{j+1}]}(X_i) \text{ for x} \in \text{the j}^{th} \text{ bin (j=1...m).} \quad (1)$$

Where $t_{j+1} - t_j = h$ is the bin-width and $I_{[a,z]}(X)$ is an indicator function:

$$I_{[a,z]}(X) = 1 \text{ if } X \in [a, z]$$

$$I_{[a,z]}(X) = 0 \text{ if } X \notin [a, z]$$

If $n_{1...m}$ are the bin counts of a histogram, then $n_{1...m}$/N are the maximum likelihood estimates of a multinomial distribution (Rice, 1995). Each of the individual bin counts n_j (with $1 \leq j \leq m$) is distributed according to a binomial marginal of this multinomial distribution. It is also intuitively clear that n_j is a binomial random variable: a measurement either falls within the jth bin or it doesn't. The same is true for the cumulative distribution:

$$\hat{F}(x) = \frac{1}{Nh} \sum_{i=1}^{N} I_{[-\infty, t_{j+1}]}(X_i) \text{ for x} \in \text{ one of the first j bins.} \quad (2)$$

A disadvantage of the histogram as an estimator of probability density is its discreteness. Also, the standard histogram doesn't take into account the possible presence of measurement uncertainties. The kernel density function is an alternative, smooth density estimator that resolves these issues. It is discussed in the next subsection. However the histogram does have some significant advantages to the kernel density estimator, but they will be discussed later.

Kernel density estimates

Written in a similar form as equation 1, the kernel density function is defined as (Silverman, 1986):

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} w_h(x - X_i), \tag{3}$$

where $w_h(\cdot)$ is the *kernel*. Although many other functions can be used, this is typically taken to be the normal distribution:

$$w_h(x - X_i) = \frac{1}{h\sqrt{2\pi}} e^{-(x-X_i)^2/(2h^2)} \tag{4}$$

where h is the *bandwidth* of the kernel. In provenance geology, kernel density functions are often used to estimate detrital age distributions from a relatively small number (typically 10-100) of age measurements and their uncertainties. h is then substituted by the standard deviation of the analytical uncertainties, which makes these kernel densities *variable bandwith* estimates. As opposed to the standard histogram, these kinds of kernel density functions are smooth functions that take into account measurement uncertainties. However, the latter feature can also be incorporated into histograms by "binning" kernel density estimates (fig 1). Such a hybrid density estimator will be called a *kernel-histogram*. Similar to the histogram, it is also possible to construct a cumulative density estimator for the kernel method :

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} W_h(x - X_i), \tag{5}$$

with $W_h(\cdot)$ a "cumulative kernel function", e.g. for the Gaussian kernel:

$$W_h(x - X_i) = \frac{1}{h\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(\chi - X_i)^2/(2h^2)} d\chi \tag{6}$$

THE DEFINITION OF A CONFIDENCE INTERVAL

Surprisingly enough, there is no general agreement in the statistics community on the definition of a confidence interval. There are two points of view.

The Frequentist approach

According to the "Frequentist", a confidence interval for a parameter θ *"consists precisely of all those values of* θ_0 *for which the null hypothesis* H_0*:* $\theta=\theta_0$ *is accepted"* (Rice, 1995). For example, we saw earlier that the histogram represents the outcome of a multinomial experiment. The probability distribution of each of the bin counts of a histogram is the marginal of a multinomial distribution, which is the binomial distribution. Consider a bin containing n out of N measurements. The maximum likelihood estimate for the binomial parameter p then is $\hat{p}_{mle} = n/N$. Now consider the null hypothesis $H_0 : p = p_o$ versus the alternative $H_a : p \neq p_o$. H_0 is accepted on a 100(1-α)% confidence level iff:

$$\sum_{x=n}^{N} \binom{N}{x} p_o^x p_o^{N-x} < \frac{\alpha}{2} < \sum_{x=0}^{n} \binom{N}{x} p_o^x p_o^{N-x} \tag{7}$$

Now, according to the definition, a two-sided confidence interval contains all those values for p_o which pass the test given by equation 7. The solution can be found by numerical iteration and/or interpolation (Clopper and Pearson, 1934). An example for N=50, n=20 and α=0.1 is given in figure 2. It can be shown (e.g. Blyth, 1986), that equation 7 is mathematically equivalent to:

$$\underline{p} = B(1 - \frac{\alpha}{2}, n+1, N-n) < p < B(\frac{\alpha}{2}, n, N-n+1) = \bar{p} \tag{8}$$

Where $B(\alpha, a, b)$ is an α percentile of the β distribution with parameters a and b:

$$\beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \tag{9}$$

where Γ(x) is the gamma function, which can be considered the continuous version of the factorial operator. For example, if x is an integer, then Γ(x+1)=x!. Likewise, the β distribution can be thought of as being the continuous version of the binomial distribution. Notice that for n=0 and n=N, equation 8 breaks down. Instead, the following expressions should be used:

$$\underline{p} = \quad 0 < p < 1 - \alpha^{1/N} = \bar{p} \quad \text{if n=0, or} \tag{10}$$

$$\underline{p} = \quad (1-\alpha)^{1/N} < p < 1 = \bar{p} \quad \text{if n=N} \tag{11}$$

The Bayesian approach

For a "Bayesian", a 100(1 - α)% confidence (or *credibility*) interval for a parameter θ given some data **x** is an interval for θ that covers 100(1 - α)% of its *posterior* distribution P(θ|**x**), where the latter is given by:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)P(\theta) \tag{12}$$

with P(θ) a *prior distribution* on θ and P(**x**|θ) the *sampling distribution* of the data given the parameter. The subjectivity of the Bayesian approach lies in the choice of the prior distribution. A uniform distribution ("flat prior") is often taken if no prior information exists as to what the value of θ should be. However, whether or not this is a good *non-informative* prior has been challenged. The uniform distribution does not yield posterior distributions that are invariant under reparameterization (Jeffreys, 1946). We will soon see an example of an alternative prior distribution that does have this invariance.

We now return to the problem of independent credibility intervals for multinomial proportions. Again, we consider a bin with n counts out of N and want to construct a 100(1-α)% credibility interval for p=n/N. The sampling distribution is binomial: $P(n|p) = \binom{N}{n} p^x p^{N-n}$. If we take a flat prior for P(p), then the posterior is a $\beta(n+1, N-n+1)$ distribution (Bayes, 1763):

$$P(q < p < r|n) = \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)} \int_q^r p^n (1-p)^{N-n} dp \tag{13}$$

Therefore:

$$B(1 - \frac{\alpha}{2}, n+1, N-n+1) < p < B(\frac{\alpha}{2}, n+1, N-n+1) \tag{14}$$

Notice the similarities between equations 8 and 14. However, as opposed to the frequentist equation 8, the Bayesian equation 13 does not require a special case for n=0 and n=N. The β distribution is an example of a *conjugate prior*.

This means that if we take a β-distributed prior, and a binomial sampling distribution, then the posterior will also have a β distribution. The uniform distribution is a special case of the β distribution for a=b=1 (i.e. $\beta(1,1)$). $\beta(\frac{1}{2},\frac{1}{2})$ is a noninformative prior (*Jeffreys' prior*) for the binomial distribution that is invariant under reparameterization (e.g. Gill, 2002, p 124). The posterior distribution then becomes β(n+$\frac{1}{2}$,N-n+$\frac{1}{2}$), resulting in confidence intervals that are not very different from equation 14, except if n≈0 or n≈N. Taking the same example as before (i.e. n=20, N=50, α=0.1), figure 3 shows a two-sided Bayesian credibility interval for p.

SIMULTANEOUS CONFIDENCE INTERVALS FOR MULTINOMIAL PROPORTIONS

As was shown in the previous section, it is relatively easy to construct *independent* confidence intervals for each of the m bin counts n_j ($1 \leq j \leq m$) that make up a histogram, both under the frequentist and the Bayesian paradigm. However, we need to be more ambitious than that. In order to be able to compare two samples and test if they are significantly different, we would like to construct *simultaneous* confidence intervals for all of the m histogram bins. We would also like to have a similar option for the kernel density function.

Frequentist Confidence Regions

As discussed before, histograms are representations of multinomial distributions. Suppose we have N numbers, distributed over m bins, corresponding to m multinomial proportions. The bin counts $(n_1, ..., n_m)$ must fulfill the condition $\sum_{j=1}^{m} n_j$=N. Therefore, all possible multinomial distributions must fall on an *m-simplex* Δ_{m-1}. An example of a 3-simplex is shown on figure 6. Consider a histogram with m bins, representing a sample of N numbers: $\mathcal{X}_N = \{X_1, ..., X_N\}$. This histogram corresponds to one point on Δ_{m-1}, the *maximum likelihood estimate* (mle). Under the frequentist paradigm, a 100(1-α)% confidence region on Δ_{m-1} consists of all those probability vectors **p** = $(p_1, ..., p_m \mid \sum_{j=1}^{m} p_j = 1)$ which are capable of yielding observations as extreme as **n** = $(n_1, ..., n_m \mid \sum_{j=1}^{m} n_j = N)$ with at least 100(1-α)% probability.

In order to find this region, a grid of possible $\mathbf{p}^{kl} = (p_1^{kl}, ..., p_m^{kl} \mid \sum_{j=1}^{m} p_j^{kl} = 1)$ is evaluated. For each of these grid points (k=1,2,...,K , l=1,2,...,L), a large number B of synthetic multinomial samples is generated by a "bars and stars" procedure:

1. Generate the following vector of m+1 numbers: $\mathcal{A} = \left(0, (p_1^{kl}), (p_1^{kl} + p_2^{kl}), ..., (\sum_{j=1}^{m} p_j^{kl} = 1)\right)$. This represents the edges ("bars") of a histogram. The gaps between subsequent entries in this array represent the multinomial probabilities (p_1^{kl},...,p_m^{kl}).

2. Create a matrix $\mathcal{B}$ of size B×N with random numbers between 0 and 1, drawn from a uniform distribution. This represents B synthetic *samples* of N values ("stars").

3. For each row of $\mathcal{B}$, calculate the number of "stars" that fall in between the "bars" of $\mathcal{A}$. This procedure yields a new matrix $\mathcal{H}^*$ of size B×m with multinomial replications of $\mathbf{p}^{kl}$.

Next, the 100α% *convex hull* of $\mathcal{H}^*$ is calculated. This is a polygon around the smallest set containing 100α% (the so-called "hull-percentile") of the rows of $\mathcal{H}^*$. An example for m=3 on Δ_2 is given in figure 4. We test to see if p^{mle} falls within the 100α% convex hull of p^{kl}. If this is not the case, then p^{kl} falls outside the 100α% confidence region of p^{mle}. This procedure is repeated for the entire grid (k=1...K , l=1...L). On figure 5, the contour line contains all those grid points for which the mle falls within their 95 percentile hull. Projecting the frequentist confidence region onto the axes of the simplex would not represent that region, but the smallest polygon circumscribing it. In higher dimensions (m≫3), smallest circumscribing polygons are mostly void and this procedure would grossly exaggerate the size of the confidence region. Therefore, it is not possible to accurately "translate" a frequentist contour plot into error bars on

a histogram, which makes it impossible to easily visualize the frequentist uncertainties of histograms with more than three bins.

Bayesian Credibility Regions

It is relatively easy to generalize the Bayesian methodology outlined above from a binomial to a multinomial situation. Recall that the conjugate prior to a binomial distribution is the β-distribution. The conjugate prior to a multinomial distribution is the Dirichlet distribution:

$$D_{\mathbf{a}}(p_1, ..., p_m) = \frac{\Gamma(\sum_{i=1}^{i=m} a_i)}{\prod_{i=1}^{i=m} \Gamma(a_i)} \prod_{i=1}^{i=m} p_i^{a_i - 1} \tag{15}$$

The multinomial uniform distribution is a special case of the Dirichlet distribution with all a_i=1. If **n** is a vector of m bin counts, then the posterior distribution is $D_{\mathbf{n}+1}(p_1,...,p_m)$. The choice of a prior that is truely non-informative and invariant under reparameterization is more controversial for the Dirichlet prior than it was for the β prior. Jeffreys suggested taking a_i=1/2, while Perks recommended using a_i=1/m ($\forall$ i=1...m) (Good, 1965). The differences for the resulting posterior distribution are slight, except for empty bins (figs 9 and 10). Similar to the binomial case, simultaneous Bayesian confidence bands for the multinomial distribution are intervals that cover 100(1-α)% of the area under the posterior distribution. As opposed to the β distribution, there are no tables of the percentiles of the Dirichlet distribution. In order to integrate this multi-dimensional function ourselves, we have to numerically sample from it. The following procedure takes care of this: generate a vector **x** = (x_1,...,x_j,...,x_m) by drawing each of the x_js from a gamma distribution with shape parameter a_j. Then Θ = (θ_1,...,θ_j,...,θ_m) with $\theta_j = x_j / \sum_{j=1}^{m} x_j$ has the desired Dirichlet distribution (Devroye, 1986). Alternatively, it is also possible to obtain a sample of the posterior distribution using a procedure named the *Bayesian bootstrap* (BB)(Rubin, 1981).

Thus, a B$\times$m matrix $\mathcal{H}^*$ of B "bootstrap histograms" can be constructed, representing B samples from the posterior Dirichlet distribution. All these histograms correspond to points on Δ_{m-1}. Both the "traditional" as the BB way to generate $\mathcal{H}^*$ yield the same results. Asymptotically, *independent* $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles for each of the histogram bin counts will converge to the independent confidence intervals of equation 14. However, it is also possible to obtain *simultaneous* confidence bands. The Bayesian way of doing this is to find m confidence intervals that define a polygon on Δ_{m-1} containing 100(1-α)% of the posterior distribution (fig 6), using the following recursive procedure:

1. Given a matrix $\mathcal{H}^*$ containing B bootstrap histograms of m bins each;

2. Construct a two-sided 100(1-γ)% credibility interval for each of the columns ("bins") of this matrix. This can be done either analytically with equation 14, or numerically by computing the 100(α/2) and 100(1-α/2) percentiles. This yields m independent credibility intervals.

3. For each column, accept those values (rows) that fall within its respective credibility interval and reject those rows that fall outside of it. Divide the number of rejected rows by B (the total number of rows), call this fraction ρ. If $\delta = \rho - \alpha > 0$, repeat step #2 for a larger γ. If $\delta < 0$, repeat it for a smaller γ.

4. Stop the iteration if δ is small enough (e.g. $<$0.001). The independent 100(1-γ)% credibility intervals for each of the bins then correspond to simultaneous 100(1-α)% credibility bands for the entire histogram.

This Bayesian method yields non-zero credibility intervals, even for empty bins. It only works for histograms and not for kernel density estimates, which are continuous functions that cannot be easily represented on a simplex.

As histograms traditionally do not take into account measurement uncertainties, the Bayesian credibility bands only reflect the uncertainties induced by the counting statistics, and not those caused by analytical imprecision. A final remark to be made is that, strictly speaking, the way we have defined simultaneous Bayesian credibility regions is only exact for *categorical* histograms, such as those obtained by point-counting mineral assemblages. However, if the histogram represents a time series, which is the case in detrital thermochronology, it will have some *smoothness* to it. This effect will not be captured by the Bayesian credibility regions discussed before. The categorical Bayesian credibility bands will be on the conservative side if applied to such *autocorrelated* data. The next section discusses this issue.

Bayesian credibility bands for smooth histograms

Strictly speaking, Bayesian credibility bands are only applicable to non-smooth or categorical data. In this section, we will discuss the importance of this problem and a way to solve it. We can express the *roughness* r of a time series f(t) in terms of its second derivative:

$$r(f(t)) = \int \left(\frac{d^2 f(t)}{dt^2} \right)^2 dt \tag{16}$$

For the discrete case, for example a histogram with m bins **n** = (n_1,...,n_m | $\sum_{j=1}^{m} n_j = N$), this becomes:

$$r(\mathbf{n}) = \sum_{j=2}^{m-1} \left(n_{j-1} - 2n_j + n_{j+1} \right)^2 \tag{17}$$

We now define the *smoothness weights* w as follows:

$$w = \left(\frac{1}{r} \right)^s = \frac{1}{\left(\int \left(\frac{d^2 f(t)}{dt^2} \right)^2 dt \right)^s} = \frac{1}{\left(\sum_{j=2}^{m-1} \left(n_{j-1} - 2n_j + n_{j+1} \right)^2 \right)^s} \tag{18}$$

where s is the *smoothing parameter*. Figure 11 shows the trinomial smoothing weights for different values of s. The distribution of the weights can be used to *filter* the posterior distribution, thereby in effect serving as a *prior* distribution. Figure 12 shows the results of this kind of posterior filtering for a trinomial distribution on the 3-simplex. It turns out that unless very strong smoothing is applied, its reducing effect on the width of the simultaneous confidence intervals is mild (fig 13). Most histograms encountered in detrital geochronology are not autocorrelated beyond a distance of more than a few bins (fig 17). In such cases, it may not be necessary to apply much smoothing at all. The effect of smoothing will obviously be greater on smooth data sets such as the sine function shown on figure 14. The amount of smoothing is completely controlled by only one parameter, s. The choice of this smoothing parameter is somewhat arbitrary. A useful upper bound for s is the highest value for which the observed histogram still completely falls within the simultaneous confidence band. Confidence bands for which this is not the case are *oversmoothed*. An example of the latter is shown on figure 13. If the purpose of constructing credibility bands is to constrain the true population, oversmoothed histograms are not necessarily a bad thing. The oversmoothed credibility bands of both figure 13 and 14 correctly contain their smooth parent population. However, if the objective is merely to see whether two samples could have been derived from the same population, it may not be necessary to smooth much at all.

CONFIDENCE ENVELOPES ON KERNEL DENSITY ESTIMATES

As discussed in the previous section, the Bayesian method for constructing simultaneous bands cannot easily be

used for kernel density estimates. An alternative method that is often used for the calculation of confidence bounds on kernel regression and density estimates is the *bootstrap* (Efron, 1979; Efron and Tibshirani, 1993). The bootstrap is a way to asses the uncertainty on a variety of statistics without having to make assumptions about their distribution. This is done by using the sample $\mathcal{X}_N = \{X_1, ..., X_N\}$ as a proxy for its distribution. Then, a large number B of "bootstrap samples" $\mathcal{X}_N^{*b}$ are generated by randomly sampling B times N numbers from $\mathcal{X}_N$, *with replacement*.

$$\mathcal{X}_N^{*b} = \{X_1^{*b}, ..., X_N^{*b}\}, \text{ where } X_i^{*b} = X_1|X_2|...|X_N \text{ for } i = 1...N, \text{ and b=1...B.} \tag{19}$$

B replicates of a statistic $\phi(\mathcal{X}_N)$ of interest can be obtained by using the "plug-in principle": $\phi^{*b}(\mathcal{X}_N) = \phi(\mathcal{X}_N^{*b})$.

To calculate simultaneous confidence bands on a kernel density function, it suffices to generate B bootstrap replications of the kernel density estimate, evaluated at a certain number m of discrete points. Suppose that we are given some data $\mathcal{X}_N = \{X_1, ..., X_N\}$ and their associated uncertainties $\mathcal{S}_N = \{\sigma_1, ..., \sigma_N\}$. Then, one bootstrap replicate of the kernel density estimate is obtained by sampling N times, with replacement, from the set of Gaussians with means $\mathcal{X}_N$ and standard deviations $\mathcal{S}_N$, and using equation 3. Evaluating such a bootstrap replicate at m discrete points yields one row of the B$\times$m matrix $\mathcal{K}^*$. Going through the recursive elimination procedure described before results in an optimal confidence level $\alpha \leq \gamma \leq 1$ at which independent confidence intervals for the columns of $\mathcal{K}^*$ will yield a 100(1-α)% simultaneous confidence band for the entire matrix. Such bootstrapped confidence bands are smooth, and take into account measurement uncertainties. However, over horizontal intervals without observations, the bootstrap will yield unrealistic confidence bands of zero width. This is illustrated by figure 15. The bootstrap replicates $\mathcal{X}_N^{*b}$ represent an (approximate) non-parametric, noninformative Bayes' posterior distribution (Hastie and others, 2001). Kernel density estimates, much like non-parametric regression, have real problems with their *consistency* properties in a Bayesian setting (Diaconis and Freedman, 1986). The zero-width confidence bands are indicative of these problems.

A COMPROMISE: THE KERNEL-HISTOGRAM

We saw that it is relatively easy to compute simultaneous credibility bands for histograms. These bands are of non-zero width, even for empty bins. They properly represent the uncertainty that is caused by counting statistics. However, histograms ignore measurement uncertainties, which can be quite substantial. If these measurement uncertainties are greater than the width of the histogram bins, it is possible that some non-empty bins of the sample histogram in fact correspond to empty population bins. The kernel density method takes care of measurement uncertainties. It has the added advantage that it is a smooth, continuous estimator, as opposed to the histogram. Unfortunately, it is not easy to calculate simultaneous confidence intervals for kernel density estimators. The bootstrap method does a good job at estimating counting-statistics induced uncertainty when there are a lot of data. However, when there are few datapoints, and gaps exist in the sampling distribution, it gives unrealistic, zero-width confidence bands. As briefly introduced in before, it is possible to construct an *ad hoc* density estimator for which it is possible to compute confidence bands that incorporate both the effect of counting statistics and measurement uncertainties. This hybrid estimator is the kernel-histogram. It is obtained by "binning" the kernel density. Simultaneous credibility intervals can be computed using the Bayesian methods described before. Figure 16 shows an example of a kernel-histogram. The added uncertainty induced by the measurement errors, and the fact that this figure (as well as figs 13 and 10) is only based on a very small sample of 25 measurements causes the credibility bands to be relatively wide. A more realistic example on real data is shown in figure 17.

CONFIDENCE BOUNDS ON CUMULATIVE DISTRIBUTIONS

For the construction of simultaneous confidence bands for probability density functions, most of the methods that were discussed before involved the generation of a large number of possible multinomial outcomes. For the bootstrap, this was done by sampling, with replacement, from the sampling distribution. For the Bayesian method, numerical samples were generated from a posterior distribution. These methods can easily be extended from density functions to cumulative distributions by taking the cumulative sum of each of the numerical (re)samples and applying the recursive elimination procedure outlined above.

An alternative to these numerical methods is derived from the *Kolmogorov-Smirnov* (K-S) test, a non-parametric way to decide if two distributions are compatible with each other (Conover, 1980). If $A_1(x)$ and $A_2(x)$ are two *cumulative* distributions, the two-sided test statistic T is defined as the greatest (supremum) vertical distance between the two curves:

$$T = sup_x|A_1(x) - A_2(x)| \tag{20}$$

The maximum allowed values of T for different confidence levels α can be looked up from tables. Similar to the dualitity between hypothesis tests and confidence intervals that was demonstrated for the frequentist case, also the Kolmogorov-Smirnov test can be used for constructing simultaneous confidence bands. In order to find 100(1-α)% confidence evelopes for a distribution, simply draw two curves above and below it, at a vertical distance equal to the critical value for T.

A comparison of the K-S, Bayesian, and bootstrapped kernel methods is shown in figure 18. The K-S and Bayesian methods yield stepwise cumulative probability estimates, while the cumulative kernel function and its confidence bands are continuous. The confidence bands of the K-S and Bayesian methods have non-zero width everywhere, also outside the data range. The larger the number of samples (N), the narrower these confidence bands will be. In contrast with this, the bootstrap method yields confidence bands that have zero width outside the data range. The K-S and cumulative kernel confidence bands have a constant width wherever there are no data, while the Bayesian credibility band steadily increases, even over a range of empty bins.

CONCLUSIONS

This paper showed how to construct simultaneous confidence bands for histogram and kernel density estimators, as well as estimators of cumulative probability. Confidence bands provide a measure of the influence of counting statistics on measured distributions. Confidence bands also allow a better judgement of the possible similarities between different populations. If measurement errors are small, the histogram is a good estimator of probability density, for which exact Bayesian credibility bands can be calculated. These have non-zero width even over intervals that were not sampled. This is important in disciplines such as detrital thermochronology, where not just the presence, but also the *absence* of certain (age) components is important. The degree of confidence that certain age intervals are absent in a detrital population can be calculated analytically (Vermeesch, 2004). For example, if 100 mineral grains are counted, there is up to 11% chance that at least one fraction ≥ 0.05 of the population was missed by that sample. (Bayesian) credibility bands such as those on figure 17 are an alternative way of expressing this kind of uncertainty. If the measurement errors are significant, and there are indications that the population is somewhat smooth, it may be appropriate to use the kernel density method. However, if there are long intervals without data, the corresponding confidence bands will have zero width. The kernel-histogram is a hybrid solution that takes into account both counting

statistics and measurement uncertainties.

This paper comes with a computer program named EPDU (Estimator of Probability Density and its Uncertainties) that runs on both PC and Macintosh computers. Along with the Matlab code that generated the figures in this paper, this program is available online (`http://pangea.stanford.edu/research/noble/epdu`).

REFERENCES.

Avigad, D., Kolodner, K., McWilliams, M., Persing, H., and Weissbrod, T., 2003, Provenance of northern Gondwana Cambrian sandstone revealed by detrital zircon SHRIMP dating.: Geology, v. 31, no. 3, p. 227-230.

Bayes, F. R. S., 1763, An essay towards solving a problem in the doctrine of chances: Philosophical Transactions, v. 53, p. 370-418.

Blyth, C. R., 1986, Approximate binomial confidence limits: Journal of the American Statistical Association, v. 81, no. 395, p. 843-855.

Clopper, C. J., and Pearson, E. S., 1934, The use of confidence or fiducial limits illustrated in the case of the binomial: Biometrika, v. 26, no. 4, p. 404-413.

Conover, W. J., 1999, Practical nonparametric statistics: New York, John Wiley, 584 p.

Devroye, L., 1986, Non-uniform random variate generation: New York, Springer-Verlag, xvi, 843 p.

Diaconis, P., and Freedman, D., 1986, On the consistency of Bayes estimates: The Annals of Statistics, v. 14, no. 1, p. 1-26.

Efron, B., 1979, Bootstrap methods: another look at the jackknife: Annals of Statistics, v. 7, p. 1-26.

Efron, B., and Tibshirani, R., 1993, An introduction to the bootstrap, Monographs on statistics and applied probability ; 57: New York, Chapman and Hall, 436 p.

Gill, J., 2002, Bayesian methods : a social and behavioral sciences approach: Boca Raton, FL, Chapman and Hall, 459 p.

Good, J. I., 1965, The estimation of probabilities: an essay on modern Bayesian methods: Cambridge, Massachusetts, MIT Press, 109 p.

Hastie, T., Tibshirani, R., and Friedman, J. H., 2001, The elements of statistical learning : data mining, inference, and prediction: New York, Springer, 533 p.

Howarth, R. J., 1998, Improved estimators of uncertainty in proportions, point-counting, and pass-fail test results: American Journal of Science, v. 298, no. 7, p. 594-607.

Jeffreys, H., 1946, An invariant form for the prior probability in estimation problems: Proceedings of the Royal Society of London A, v. 186, no. 1007, p. 453-461.

Rice, J. A., 1995, Mathematical statistics and data analysis: Belmont, CA, Duxbury Press, 602p.

Rubin, D. B., 1981, The Bayesian Bootstrap: Annals of Statistics, v. 9, no. 1, p. 130-134.

Scott, D. W., 1992, Multivariate density estimation : theory, practice, and visualization, Wiley series in probability and mathematical statistics: New York, Wiley, 317 p.

Silverman, B. W., 1986, Density estimation for statistics and data analysis, Monographs on statistics and applied probability: London ; New York, Chapman and Hall, 175 p.

Vermeesch, P., 2004, How many grains are needed for a provenance study? Earth and Planetary Science Letters (in press).

FIGURES

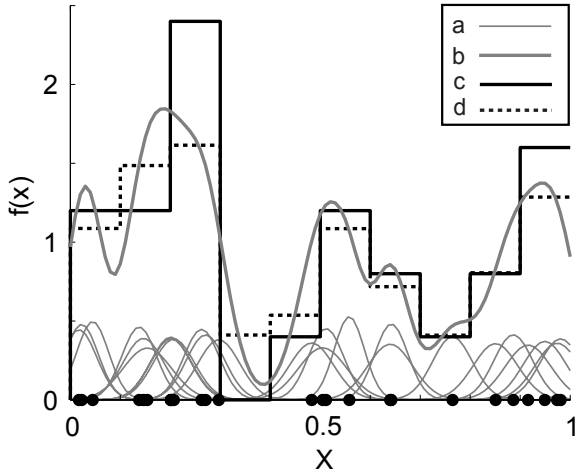

Figure 1: The black dots represent a synthetic sample of 25 numbers (”measurements”), randomly drawn from a uniform distribution. The measurement uncertainties are normally distributed and heteroscedastic (uniformly distributed standard deviations). Therefore, the natural kernel for this sample is the Gaussian kernel, which is shown for each measurment (a). The kernel density estimate is then simply the result of stacking these curves (b). The histogram is a stepwise density estimator that doesn't take into account the measurement uncertainties (c). However, it is possible to modify the histogram and discretize the kernel function, which yields a hybrid density estimator, the ”kernel-histogram” (d).

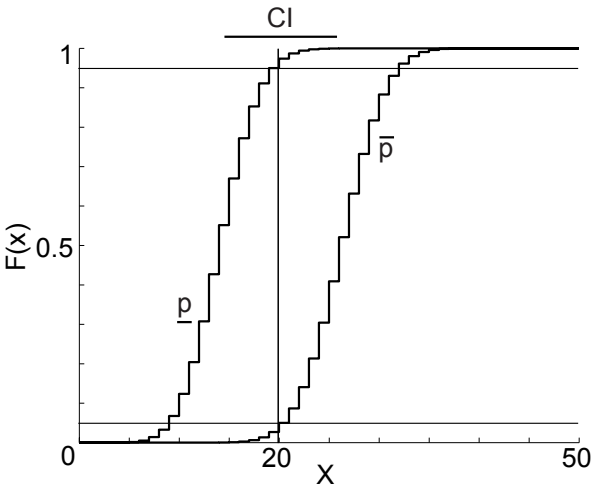

Figure 2: 90% frequentist confidence bounds on p for n=20,N=50. The step-functions represent the cumulative binomial distribution with parameters (N,$\underline{p}$) and (N,$\bar{p}$), respectively.

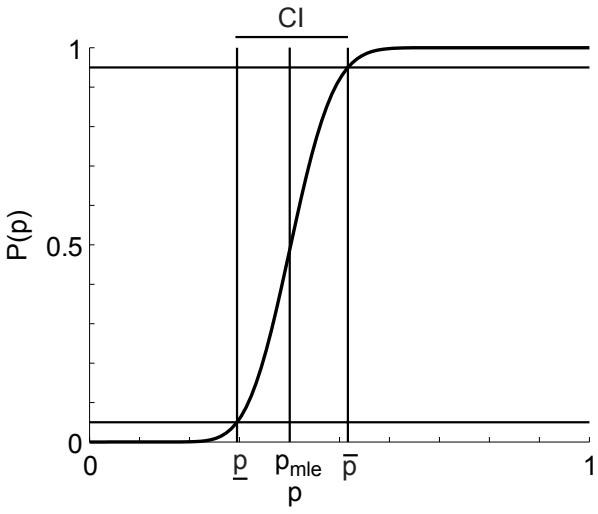

Figure 3: 90% Bayesian credibility bounds on p for n=20,N=50. The curve represents the cumulative β distribution function with parameters n+1 and N-n+1 (i.e. a flat prior).

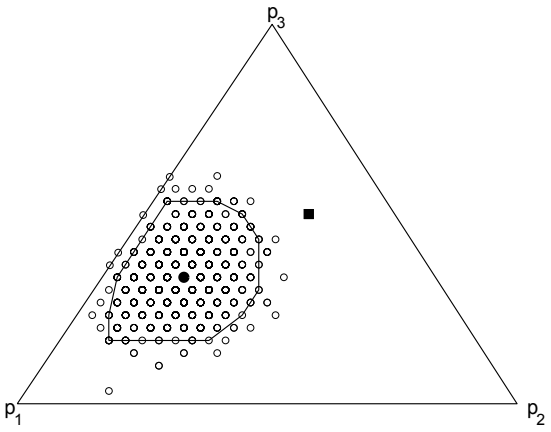

Figure 4: To test if the trinomial distribution marked by the black dot (p_1=1/2,p_2=1/6,p_3=1/3) belongs to the 95% confidence region of the trinomial experiment marked by the black square (n_1=5,n_2=10,n_3=15), a large number (1000) of trinomial samples of N=30 numbers was generated from this distribution. They are represented by the open circles. The black contour line represents the 95% convex hull. Since the black square does not fall within this hull, the black dot falls outside the 95% confidence region of the trinomial experiment.

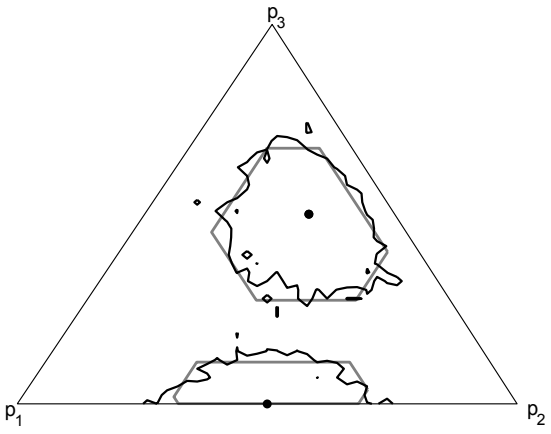

Figure 5: The black dots represent the maximum likelihood estimates (mle) for two trinomial experiments (n_1=5,n_2=10,n_3=15) and (n_1=15,n_2=15,n_3=0). The black contours represent the frequentist confidence regions, obtained by repeating the experiment shown in figure 4 on a 1250 point grid. For each of the grid points, B=200 trinomial samples were generated. The gray lines outline the Bayesian credibility regions (using a flat prior). The agreement between the two methods is surprisingly good.

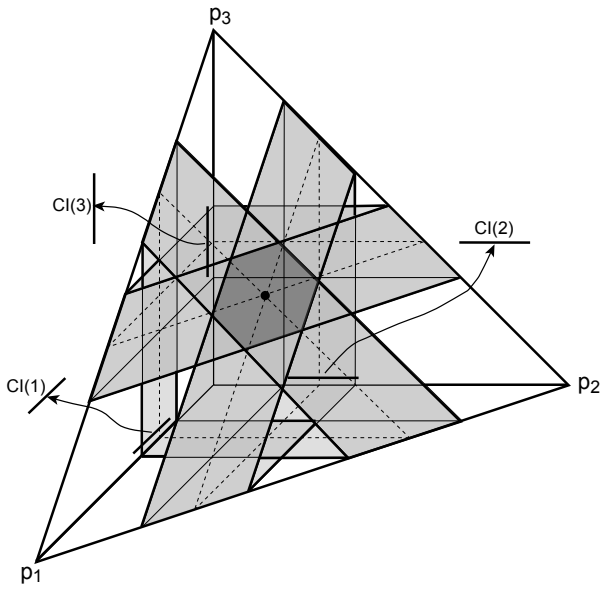

Figure 6: All the possible outcomes of a trinomial experiment, for example a 3-bin histogram of N measurements X_i plot on a 3-simplex. The maximum likelihood estimate for the multinomial proportions is given by p_j^{mle}=(#X_is in j^{th}bin)/N (j=1,2 and 3). The mle is represented by a black dot. The posterior distribution of the unknown parameters p_1, p_2 and p_3 is given by a Dirichlet distribution. To find simultaneous 100(1-α)% confidence bounds for these parameters, we need to find a polygon on the simplex that contains 100(1-α)% of the posterior distribution.

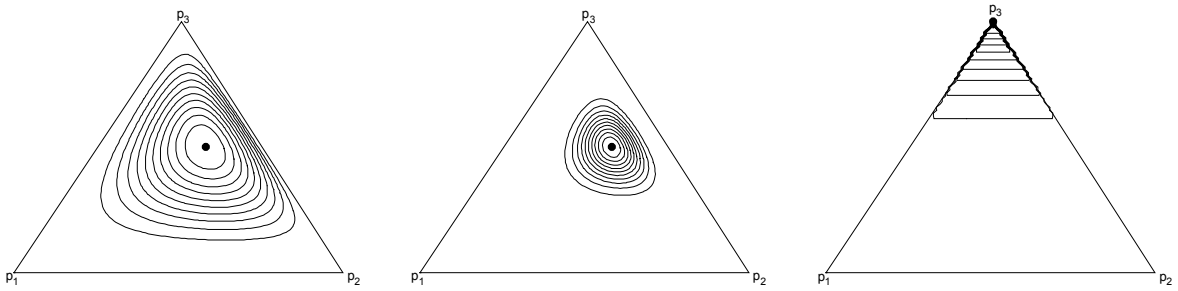

Figure 7: Analytical convex hulls for different posterior Dirichlet distributions. The parameters are, from left to right: (1,2,3), (5,10,15) and (0,0,5). Comparison of the leftmost two figures shows how the posterior Dirichlet distribution is more tightly constrained when more data are used. The rightmost plot shows how, even when two bins are empty, meaningful confidence intervals can be computed.

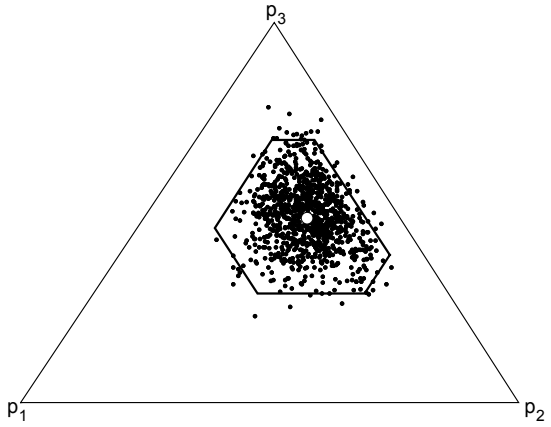

Figure 8: The area under posterior distributions such as those in figure 7 can be numerically integrated. The white dot shows the trinomial distribution for a particular sample (n_1=5, n_2=10, n_3=15). The black dots show 1000 random samples from the Dirichlet posterior distribution $D_{6,11,16}$ on the 3-simplex. The polygon contains 95% of these points. The projection of this polygon onto the three parameter axis yields the simultaneous Bayesian credibility bands.

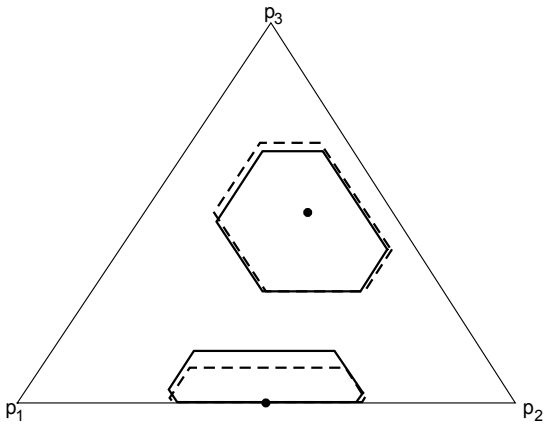

Figure 9: The black dots on this 3-simplex represent sample bin counts (5,10,15) and (15,15,0). The solid black polygons represent their respective simultaneous 95% Bayesian credibility polygons with uniform prior $D_{1,1,1}$, while for the dashed polygons, Perks' prior $D_{\frac{1}{3},\frac{1}{3},\frac{1}{3}}$ was used.

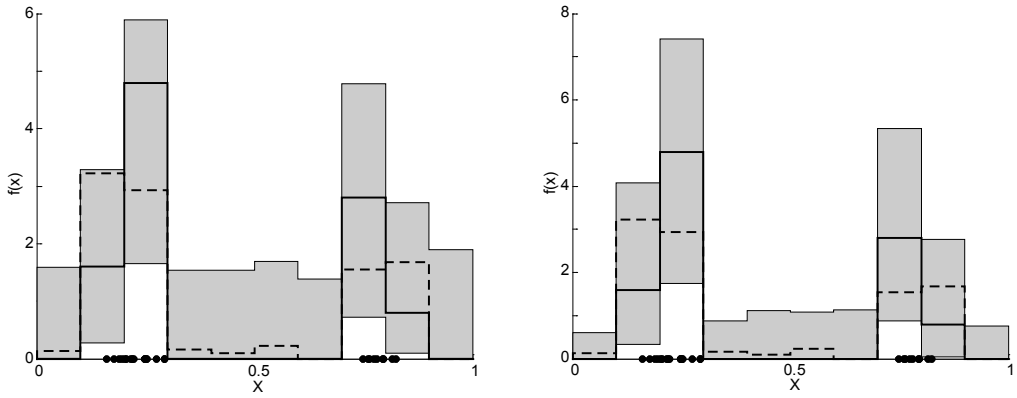

Figure 10: The dashed line represents the histogram of a synthetic, trimodal population. The black dots are 25 random samples from this population. The central black histogram was calculated from the sample. The gray area covers 95% of the posterior distribution. For the left figure, a flat (uniform) prior, and for the right figure, Perks' prior was used. Both credibility bands correctly contain the original population.

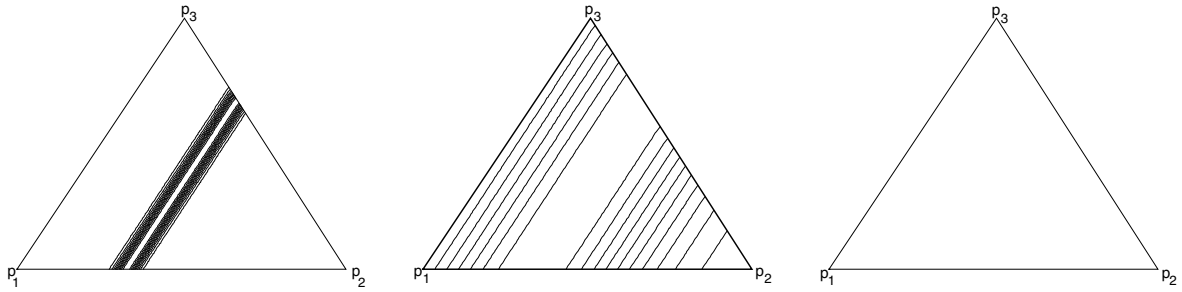

Figure 11: Contoured smoothing priors for different smoothing parameters. From left to right: s=0.95 (strong smoothing), s=0.25 (intermediate smoothing), and s=0.05 (weak smoothing). The strongest weights are located along the p_2=1/3 line, which connects all possible histograms with zero second derivative.

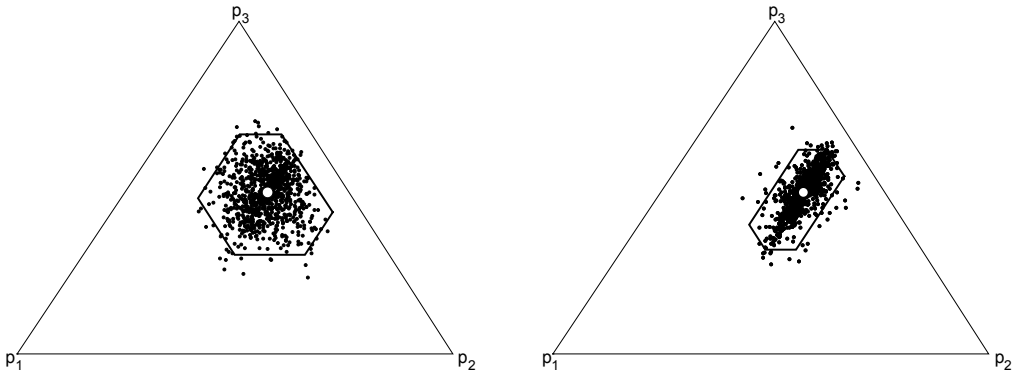

Figure 12: Smoothed versions of figure 8, using s=0.25 (left) and s=0.95 (right). It takes fairly strong smoothing to significantly change the size of the confidence polygon.

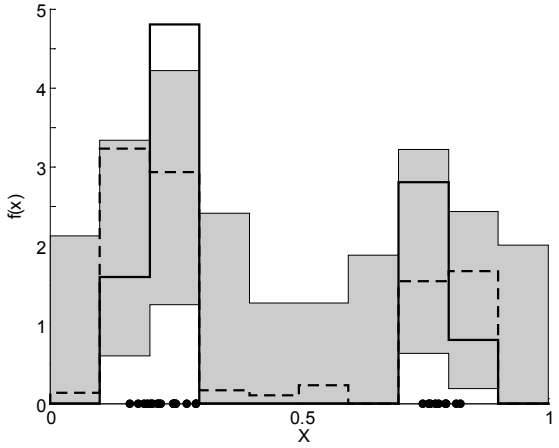

Figure 13: The trimodal population shown by the dashed histogram is the same as was used for figure 10. The gray area represents the 95% smoothed (s=5) simultaneous Bayesian credibility band, using a flat prior. The observed number of observations (black histogram) in the third bin falls *outside* the 95% confidence band, indicating that the confidence band was *oversmoothed*. Other than that, the oversmoothed histogram is only mildly different from the non-smoothed one (fig 10). Notably the height of the confidence band around X=0.5 hasn't changed.

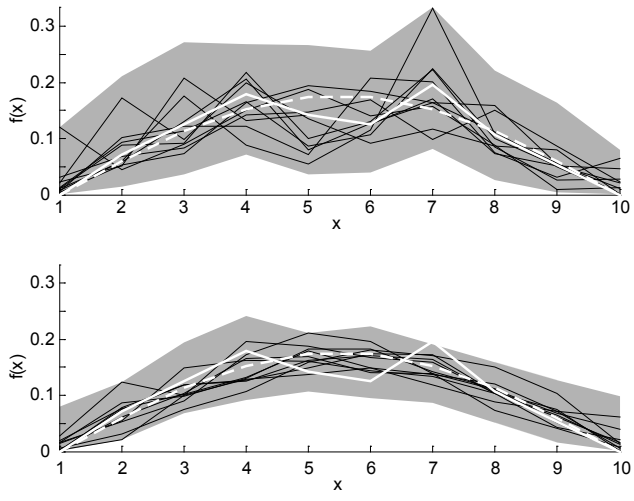

Figure 14: The dashed white lines show the histograms - or rather *frequency polygons* (Scott, 1992) - of an extremely smooth population (sine function). The solid white lines show the frequency polygons of a sample of 57 numbers that were randomly drawn from this population. As for figure 13, very strong smoothing (s=5) was applied. The gray areas mark the simultaneous 95% credibility bands, obtained by the Bayesian method and based on 200 samples from the posterior distribution. 10 of these samples are shown in black to illustrate the effect of the smoothing prior. The non-smoothed Bayesian credibility band (top) is about twice as wide as the smoothed one (bottom).

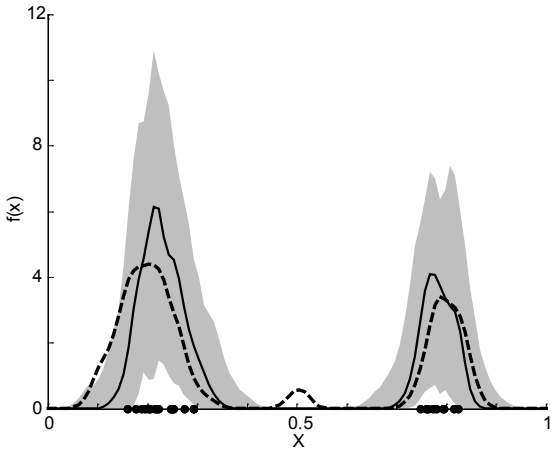

Figure 15: The dashed black line is the kernel density of a synthetic trimodal population. The black dots represent a small (N=25) sample from this population. The kernel density estimate for this sample is marked by the solid black line. A two-sided simultaneous 95% confidence band is shown in gray. It gives a realistic estimate of the uncertainty on the sample-based kernel estimate, except for those areas that had not been sampled. The second mode of the population is completely missed by the sample, yielding an erroneous zero width confidence band over that interval.

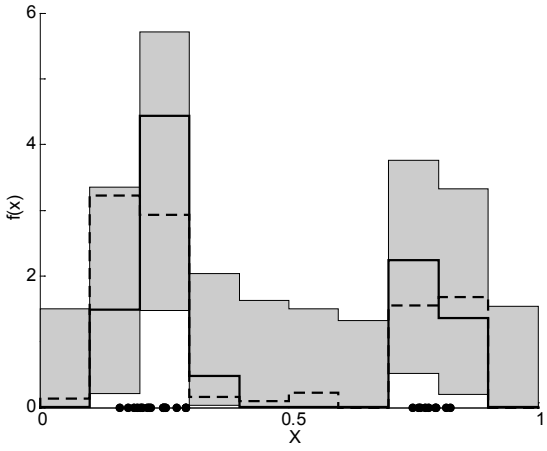

Figure 16: The kernel-histogram applied to the 25-point dataset of figure 10. Incorporating the measurement uncertainties in a histogram has a smoothing effect, and increases the width of the confidence band.

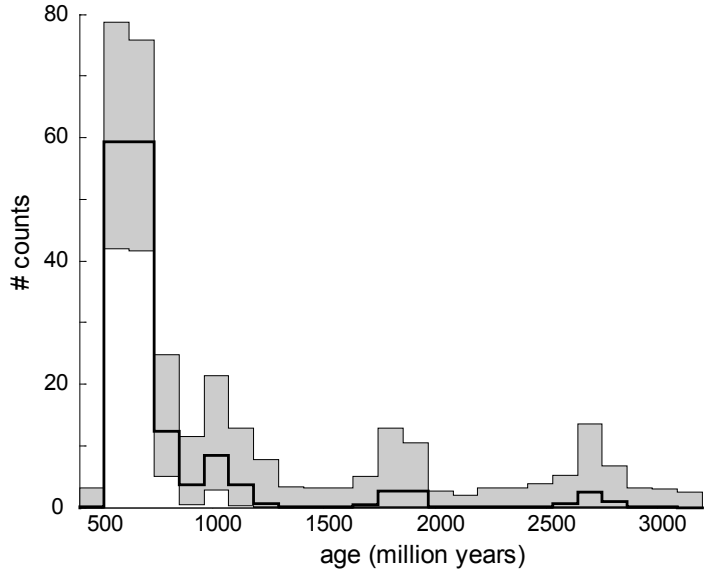

Figure 17: The kernel-histogram of 157 concordant U-Pb zircon ages from the Nubian Sandstone, Israel (Avigad and others, 2003), using Perks' prior. The black line shows the sample histogram. The gray band contains 95% of the posterior distribution. It shows that none of the apparent pre-Pan-African (500-900Ma) age peaks are statistically very significant. The credibility band also gives an idea of the probability that certain ages that are absent in the sample are in fact present in the population.

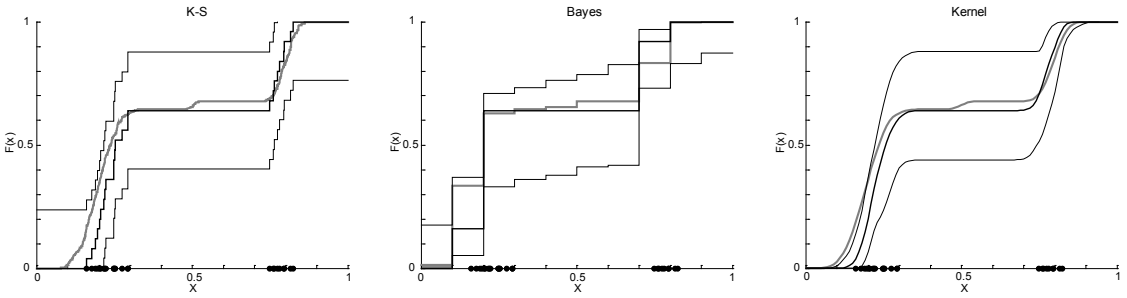

Figure 18: Comparison of three methods for calculating simultaneous 90% confidence bands on cumulative distributions. The gray lines show the trimodal population that was discussed earlier in figures 10, 13 and 15. The black dots show the location of a 25 points, drawn at random from this population. The width of the confidence bands is comparable for all three methods. See the text for further discussion.