Density Estimation for Censored Economic Data

Meghan O'Malley
Bureau of Labor Statistics, Room 1950, 2 Massachusetts Ave, NE, Washington DC 20212-0001

1. Introduction

This paper describes a simple, closed-form density estimator, derived solely from a single histogram, and explores its performance. The density estimator is defined primarily as a piecewise quadratic polynomial with adjustments made at the right and left extremes of the domain and is intended for use on proportions estimated from interval-censored data. For an intuitive feel of the performance, the density estimator is compared visually to kernel density estimates using corresponding point data. A simulation study is used to test the density estimator on two possible distribution shapes, and with two possible sample sizes. Performance is measured by the Mean Square Errors of the means and a few percentiles.

Disclaimer: Any opinions expressed in this paper are those of the author and do not constitute the policy of the Bureau of Labor Statistics.

1.1 Motivation

To what extent can accurate estimates be derived from coarse, low-detail data? Due to cost constraints, survey designers are often faced with difficult decisions regarding the quality and quantity of data that can be collected within a set budget. Survey designers try to optimize the quality of estimates produced. One of the ways in which this is done is by choosing between a relatively small amount of detailed data, a relatively large amount of coarse data.

The Occupational Employment Statistics Program conducted by the Bureau of Labor Statistics is an example of a survey that takes a fairly coarse, censored measurement for a very large sample. OES collects wage data from approximately 1.2 million business establishments every 3 years using, primarily, a form mailed to each establishment. Rather than collecting specific wage values, twelve intervals are given and the number of employees whose wages fall in each interval is recorded. From this data, OES estimates the mean wage and five percentile values. The mean is estimated by assuming a mean for each interval from an outside source and weighting the assumed means by the estimated proportion of employment within the interval.

Compared to assumed mean methods, which are frequently used in practice, density estimation has several major advantages for this type of data. It can make a more full use of the information available by considering both the proportions in the intervals and their relationship to adjacent intervals. Density estimation can be done with relatively few assumptions about the distribution as a whole and the distributions within particular intervals. Once calculated, the density estimate can be used for a variety of different types of estimates. And furthermore, if desired, the functional format allows for simple alterations or extensions. For instance, for occupational wage distributions one might add spikes at $0.25 increments for lower wages or one might move older data forward in time with higher wages rising somewhat faster than lower wages.

1.2 Data Sources

Three types of data were used for this study. Two types of data collected by the Bureau of Labor Statistics were used to test the density estimation methods described. Wage data collected as specific values through the National Compensation Survey was used to compare interval and point data density estimation methods and wage data collected in intervals by the Occupational Employment Statistics Program was used for studies of aggregation and other related issues which play a major role in the usability of a proposed method. To allow more thorough analyses and more general, reproducible results and to avoid any possible unintended disclosure of confidential BLS data, synthetic data was generated and used to produce the figures and analyses included in this paper. The generating distributions are based on the lognormal distribution, selected to mimic certain features of the collected data. The generating distributions are defined on $[5.15, \infty)$, and OES intervals from 2006 are used to define bins: 5.15, 7.50, 9.50, 12.00, 15.25, 19.25, 24.50, 31, 39.25, 49.75, 63.25, 80.00. Note that the interval boundaries are set so that the maximum relative errors of observations within each interval are approximately equal.

1.3 Available Density Estimation Methods

Density estimation methods tend to focus on smoothing point data rather than extrapolating from censored data (i.e. B.W. Silverman 1986, W. Hardle 1991). Several interesting methods for density estimation from censored data have arisen in the biostatistics field such as Peto's experimental survival curves, Turnbull's CDF, and Braun et. al.'s kernel density estimate for interval censored data. This literature focuses on iterative methods used to estimate the cumulative distribution function. Three aspects of these methods make them difficult to apply to OES and similarly censored data: the nature of the data used, the requirements for automation, and the low order of resulting CDFs.

Density estimation methods for interval censored data, including all three methods mentioned above, often make use of varying bounds for the intervals of each observation. When interval boundaries are well spread throughout the domain, there is substantial information about the shape of the CDF within any interval of the domain. However, when observations are restricted to a small set of possible intervals, although the estimated CDF may be more accurate at those boundaries, less information is available about the shape within intervals.

Available density estimation methods for interval censored data can involve fairly sophisticated iterative methods. Even a quickly converging method may be problematic for a large-scale economic survey where several hundred thousand estimates may be made regularly. OES, for example, describes wage distributions of over 800 occupations in over 600 areas and over 400 industries. Additionally, the quantity of estimates made requires a high degree of automation of the methods used; it is not reasonable to use a method which requires any non-programmable judgment.

In addition to practical concerns arising from the nature of the data and automation of the methods, available density estimation methods for interval censored data give CDFs which are either low order polynomials or heavily parametric. Peto's and Turnbull's methods result in CDFs that are undefined in each interval and horizontal elsewhere and, when all observations are in collected under the same set of fixed intervals, Braun et. al.'s kernel density estimate for interval censored data strongly resembles a histogram. CDFs of low orders result in non-intuitive PDFs, containing discontinuities, counter-intuitive cusps at boundaries, or both.

In light of these differences, it may be worth considering a density estimation method which can better accommodate this type and use of data, even at the expense of the elegant theoretical properties of the available methods for estimating.

2. Density Estimation Method

2.1 Methods Considered

Modeling with polynomial methods is fairly simple, computationally inexpensive, and reasonably flexible in the shapes that could be matched. Piecewise linear, quadratic, and cubic polynomial methods were considered. Piecewise linear methods often proved superior to piecewise quadratic methods on synthetic data but inferior for collected data due to irregularities in shape. Additional constraints at interval boundaries give the piecewise quadratic density estimator an advantage over the piecewise linear density estimator on rough histograms. Piecewise cubic polynomials, although constrained at interval boundaries, involved burdensome algebra and often produced results with counter-intuitive dips and bulges. In both simplicity and performance, piecewise quadratic density estimation was the clear favorite. A more detailed description of this density estimator is given below.

2.2 Piecewise Quadratic Density Estimation

2.2.1 General Description

The piecewise quadratic density estimator (PQDE) is formulated with two guiding principles. First, the area in each interval of the frequency histogram is preserved and, second, the curve should be somewhat smooth with no large spikes or jumps between intervals. Although neither of these guiding principles is necessary for a reasonable density estimator, and, in fact, a method which smoothes out areas across intervals may even be desirable for particularly small samples, these two guiding principles allow for a simple, intuitive, closed form density estimator which seems well suited to the large samples common to OES data and may be altered and improved later for other situations. A general description for how the polynomials are chosen follows:

1) The weighted proportion in each interval is plotted as a histogram.
2) The rightmost, half-open interval is set first. If non-zero, a scaled and shifted exponential distribution is assumed with parameters determined using the prior interval (see Section 2.2.2).
3) Connector points are set at the boundaries of each interval. The y-values are initialized at the average of the heights of the adjacent histogram bars. To maintain continuity, the y-values are set to zero if either adjacent interval is zero.

4) Connector points are algebraically adjusted to reduce differences between slopes of adjacent polynomials (see Section 2.2.4).
5) If non-zero, the leftmost closed interval uses a linear polynomial with parameters determined using the right hand connector point and the area underneath the curve.
6) Center intervals use quadratic polynomials with parameters determined using the right and left hand connector points and the area underneath the curve.

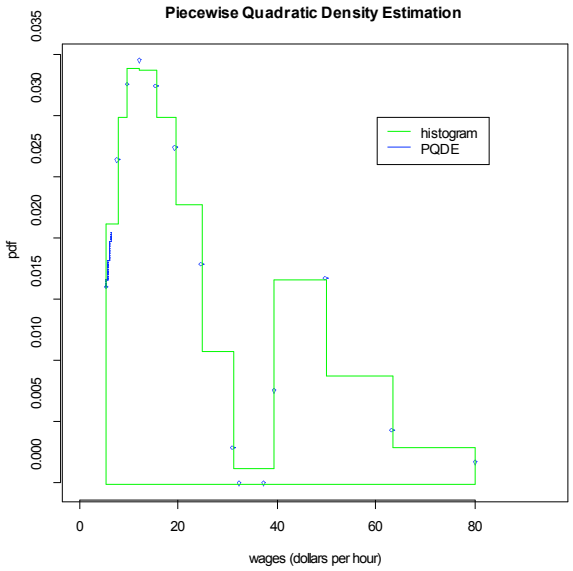


Figure 1: Histogram, adjusted connector points, and piecewise quadratic density estimate with corrections for negative values. From proportions (0.05, 0.06, 0.085, 0.11, 0.12, 0.12, 0.07, 0.01, 0.175, 0.12, 0.05, 0.03).

2.2.2 Handling Negative Regions

Using the piecewise quadratic density estimator method described in 2.2.1, it is possible for the parabolas in certain intervals to go negative. This can happen when an interval with a small proportion is sandwiched between intervals with large proportions or simply when a small proportion is adjacent to a very large proportion. In these situations, the small degree of the polynomial does not allow the curve to rise or fall fast enough for the curve to meet both conditions and remain above the x-axis. Fortunately, since the connector points are always non-negative, the negative region can be easily found by testing for a negative vertex located inside the interval. The parabola is then replaced by one or more lines which drop from the larger interval to zero in a way that preserves area. When one adjacent interval is zero, the line from the non-zero interval is fixed so that it covers the interval's entire area. When two lines are employed, the area is allocated proportionately based on the heights of the connector points on each side of the interval.

2.2.3 Handling Domain Endpoints

The treatment of domain endpoints here is tailored specifically to OES data. The interval data collected by OES has a closed bound on the left hand side at the federal minimum wage and is open on the right hand side, as there is no maximum wage.

The closed bound on the left hand side can be easily handled. For this interval, the only constraints are the area and a point on the right hand side of the interval. Either a connector point can be fixed for the left hand boundary or the parabola can be replaced by a line for the first interval. The latter was chosen here.

The right hand interval is somewhat more difficult to handle. There is very little information about the location and spread of the data in this interval, but it can have a profound impact on the mean and higher percentiles. In order to make estimates from this type of censored data without relying on strong assumptions or an outside source, the extent of this problem must be limited by raising the lower bound of the last interval until it is large enough that only a small tail remains in the rightmost interval. An exponential distribution is used here with the parameter set using the bounds, L_{n-1} and L_n, of the second rightmost interval and the areas, A_{n-1} and A_n of the two rightmost intervals. The curve sets the connector point between the rightmost two intervals and is retained only for the rightmost interval.

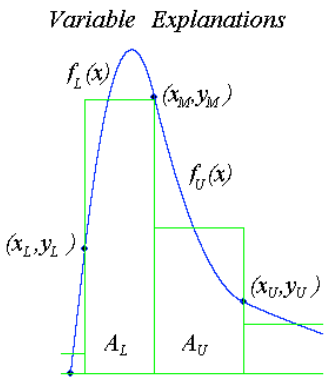
$$\frac{A_{n-1}}{A_{n-1}+A_n} = \int_{L_{n-1}}^{L_n} be^{-b(x-L_{n-1})}dx$$

$$f_n(x) = (A_{n-1}+A_n)be^{-b(x-L_{n-1})}$$

Some problems arise with small sample sizes where the second rightmost interval is empty or nearly empty and the rightmost interval is non-empty. In these cases an exponential distribution might be assumed from the third interval from the right and the combined areas of the second and third intervals from the right might be used to set the parameter for exponential decay in the rightmost interval.

2.2.4 Algebraic Improvements to Connector Points

The following section describes the formula for altering connector points to increase the smoothness of the density estimator. Applying this procedure does not optimize smoothness at boundaries; however, a single



application gives a visible improvement in smoothness. Consider each connector point, (x_m, y_m), beginning with the right hand boundary of the first interval. Lower (L) and upper (U) interval boundaries and quadratic curves are described in relation to (x_m, y_m).

The constraints for the lower polynomial can be written as:

$$A_L = \int_{x_L}^{x_M} f_L(x)dx = a_{L,2}\frac{x_M^3 - x_L^3}{3} + a_{L,1}\frac{x_M^2 - x_L^2}{2} + a_{L,0}(x_M - x_L)$$

$$y_L = f_L(x_L) = a_{L,2}x_L^2 + a_{L,1}x_L + a_{L,0}$$

$$y_M = f_L(x_M) = a_{L,2}x_M^2 + a_{L,1}x_M + a_{L,0}$$

And we can solve for the coefficients:

$$a_{L,2} = 3\frac{y_L + y_M}{(x_L - x_M)^2} - 6\frac{A_L}{(x_L - x_M)^3}$$

$$a_{L,1} = \frac{y_L - y_M}{x_L - x_M} - (x_L + x_M)a_{L,2}$$

$$a_{L,0} = y_M - x_M\frac{y_L - y_M}{x_L - x_M} + (x_L x_M)a_{L,2}$$

And, similarly, the constraints for the upper polynomial can be written as:

$$A_U = \int_{x_M}^{x_U} f_U(x)dx = a_{U,2}\frac{x_U^3 - x_M^3}{3} + a_{M,1}\frac{x_U^2 - x_M^2}{2} + a_{U,0}(x_U - x_M)$$

$$y_U = f_U(x_U) = a_{U,2}x_U^2 + a_{U,1}x_U + a_{U,0}$$

$$y_M = f_U(x_M) = a_{U,2}x_M^2 + a_{U,1}x_M + a_{U,0}$$

And we can solve for the coefficients:

$$a_{U,2} = 3\frac{y_U + y_M}{(x_U - x_M)^2} - 6\frac{A_U}{(x_U - x_M)^3}$$

$$a_{U,1} = \frac{y_U - y_M}{x_U - x_M} - (x_U + x_M)a_{U,2}$$

$$a_{U,0} = y_M - x_M\frac{y_U - y_M}{x_U - x_M} + (x_U x_M)a_{U,2}$$

Temporarily consider y_L and y_U to be fixed. For the leftmost adjusted point, since the leftmost interval is linear, y_L is not necessary in the calculations. For other connector points, the adjusted y_M from the prior calculation can be used as y_L in the current calculation. In the first application of connector point adjustments, the upper connector point y_U can be set at the average value of the histogram bars adjacent to it. For any subsequent iteration, the previously determined connector points can be used. The goal is to choose y_M to minimize the difference between the slopes of the lower and upper polynomials.

Let $\left|f_L'(x_M) - f_U'(x_M)\right| = 0$

$$\left|(2a_{L,2}x_M + a_{L,1}) - (2a_{U,2}x_M + a_{U,1})\right| = 0$$

$$\left|\frac{2y_L + 4y_M}{x_M - x_L} - \frac{6A_L}{(x_M - x_L)^2} + \frac{2y_U + 4y_M}{x_U - x_M} - \frac{6A_U}{(x_U - x_M)^2}\right| = 0$$

$$\left|y_M\left[4\left(\frac{1}{x_M - x_L} + \frac{1}{x_U - x_M}\right)\right] + \left[2\left(\frac{y_L}{x_M - x_L} + \frac{y_U}{x_U - x_M}\right) - 6\left(\frac{A_L}{(x_M - x_L)^2} + \frac{A_U}{(x_U - x_M)^2}\right)\right]\right| = 0$$

$$y_M = \frac{3\left(\frac{A_L}{(x_M - x_L)^2} + \frac{A_U}{(x_U - x_M)^2}\right) - \left(\frac{y_L}{x_M - x_L} + \frac{y_U}{x_U - x_M}\right)}{2\left(\frac{1}{x_M - x_L} + \frac{1}{x_U - x_M}\right)}$$

$$y_M = \max\{y_M, 0\}.$$

3. Results

3.1 Comparison with Kernel Density Estimates

This section is intended to give an intuitive feel for the performance of the piecewise quadratic density estimator (PQDE). Two distributions were selected to mimic certain features of the data. Samples of 100 observations were drawn from each distribution. The kernel density estimate (KDE) was formed from the sample using R function *density()*. The smoothing parameters were calculated using R function *bw.nrd();* this method is not optimal for all distribution shapes but is used throughout so that the comparison of methods is more fair. The proportion of the sample in each interval was then calculated and used to form the PQDE. The kernel density estimate and the PQDE are shown alongside the generating distribution for reference. Please keep in mind while reviewing the graphs, the difference in the precision of the data used for the KDE and that used for the PQDE. The KDE is formed using 100 specific values and the PQDE is formed using only the proportion of observations falling in each interval.

3.1.1 Lognormal Distribution

Since wage data generally takes a roughly lognormal shape (Aitchison and Brown 1957), this distribution is used to show the general behavior of the PQDE and the statistics derived from it. In OES and NCS data, with a few exceptions, most occupational series cover several intervals and have little or no data in the last, open-ended

interval. Parameters LN(3.5,0.3) were selected to reflect this situation.

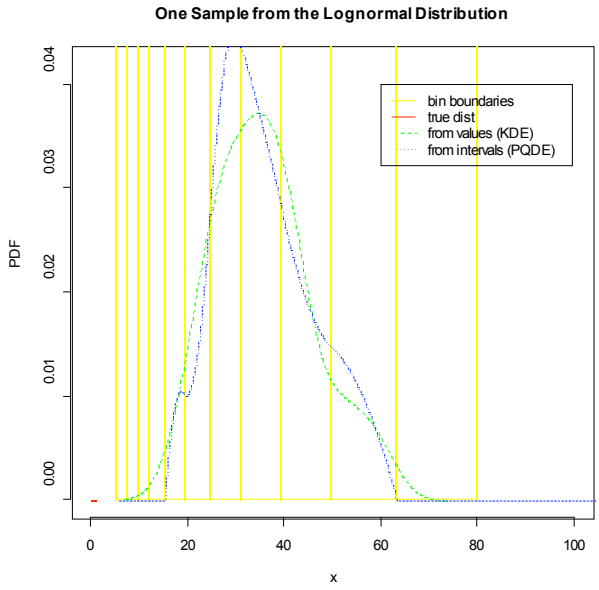


Figure: Typical performance

The graph above is intended to show the typical performance of the PQDE. For the most part, the performance of the PQDE and KDE appear comparable. In the sample shown above, the PQDE seems to have better captured the peak of the generating distribution but has distinct bumps on both tails. The bump on the right tail is due to sampling error in the interval just below 60 which the PQDE does not smooth out as well as the KDE. The bump on the left tail has no apparent cause.

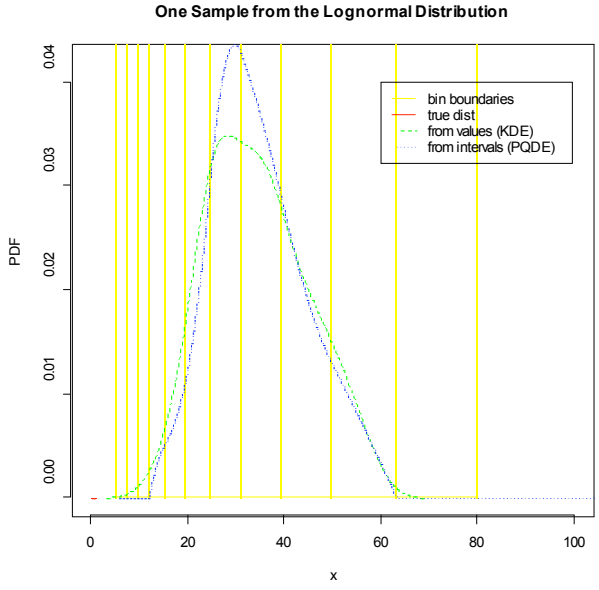


Figure: Good performance

The graph above is intended to show especially good performance of the PQDE. The PQDE is almost everywhere closer to the generating distribution than the KDE.

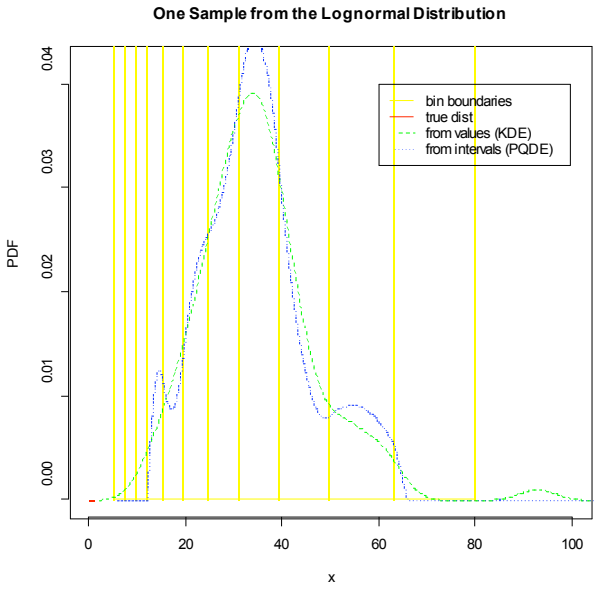


Figure: Poor performance

The graph above is intended to show especially poor performance of the PQDE. Here, there is substantial sampling error; the proportions of observations in several intervals are substantially different than expected. Both the PQDE and the KDE miss the location of the main peak and have heavier than expected tails. The KDE does a better job of smoothing the sampling errors in the tails than the PQDE.

3.1.2 Bimodal Distribution

Although the distributions for occupational wages usually appear fairly smooth and well behaved in OES and NCS data, some distributions have distinct bimodal shapes. Parameters 0.5*LN(3.1,0.2)+0.5*LN(3.8,0.2) were selected to reflect such a situation.

Although bimodal shapes are often viewed as stemming from inadequate classification, bimodal shapes may be particularly important to capture because occupational distributions can be thought of as composed of several distinct distributions, defined, for instance, by area, industry, establishment size and skill level. Changes in distribution shape sometimes appear to be gradual shifts in underlying distributions.

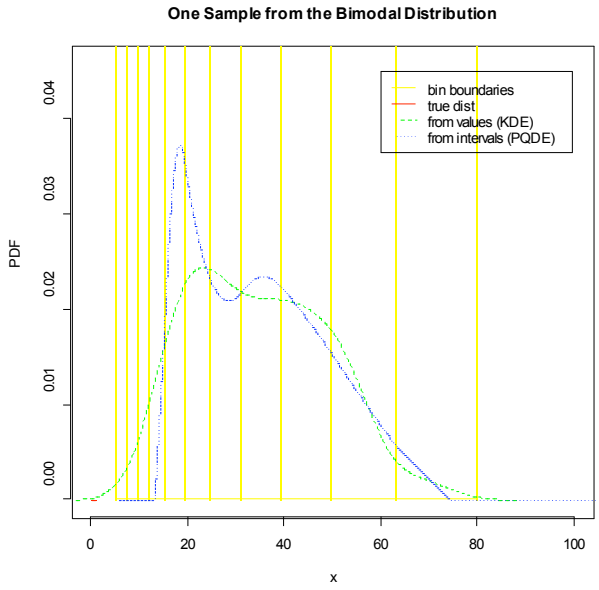


Figure: Typical performance

The sample graphed above has less observations than expected in the interval where the higher peak should be located and more than expected observations in the interval where the valley should be located, making the bimodal shape especially difficult to notice. Although the PQDE missed the location of both peaks, it did better a much better job of displaying the peaks as distinct than the KDE.

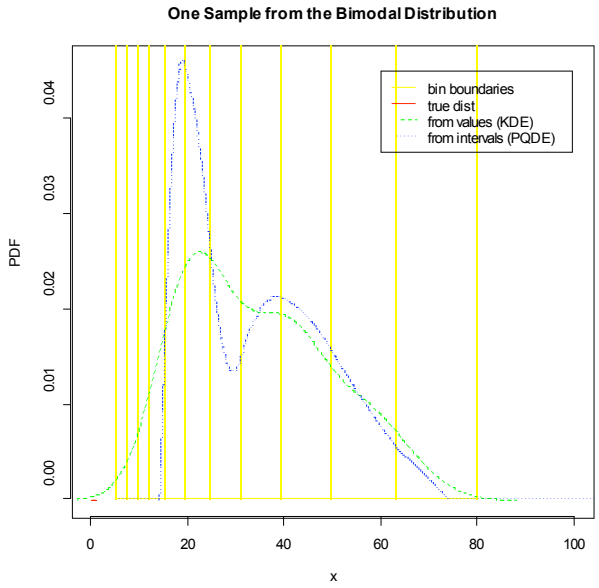


Figure: Good performance

In the sample above, the observations in each interval were fairly close to what was expected. Here the PQDE is slightly off in the location of the peaks but captures the shape of the generating distribution remarkably well. The KDE, by the nature of the smoothing parameter, smoothes sharp rises and falls masking the bimodal shape of the distribution.

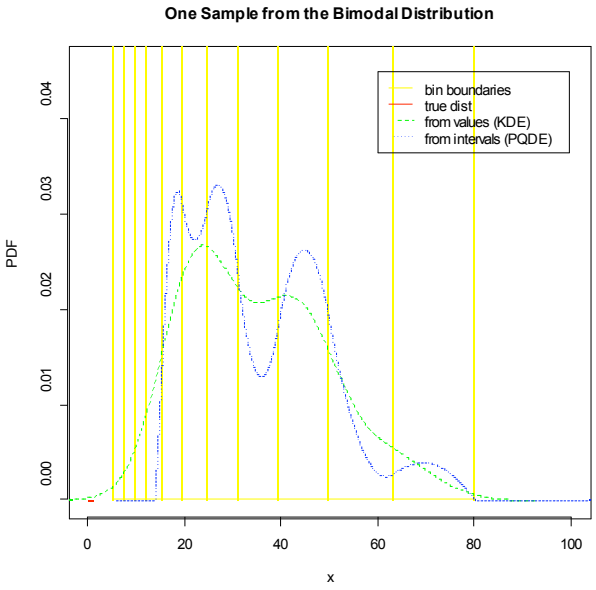


Figure: Poor performance

In the graph above, the PQDE captures the abrupt start in the left tail and the valley between the peaks of the generating distribution better than the KDE but exaggerates the sampling errors in the left peak and the right tail to the extent that the PQDE suggests four peaks rather than two.

3.2 Simulation Results

For each distribution, 5,000 samples of sizes 50 and 500 were drawn. The mean and the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles were estimated directly from each sample. The proportion of the sample in each interval was then calculated and used to derive the piecewise quadratic density estimate (PQDE) described above. The mean and the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles were then estimated again from the PQDE.

3.2.1 Lognormal Distribution

For this distribution, the true mean is 34.640 and the true 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles are 22.545, 27.049, 33.115, 40.542, and 48.641, respectively.

n=50	Rel. MSE (%)	
	PQDE	Samp
mean	20.8	19.3
10th	49.7	52
25th	33.2	32.9
50th	28.7	27.3
75th	32.8	33.5
90th	55.9	55.2

n=500	Rel. MSE (%)	
	PQDE	Samp
mean	2.111	2.033
10th	5.544	6.136
25th	3.524	3.664
50th	3.183	3.298
75th	3.584	3.858
90th	5.169	5.337

For the lognormal distribution with parameters LN(3.5,0.3), the relative mean square errors for the estimates from the piecewise quadratic density estimator are extremely similar to those calculated directly from the sample. This seems true for both sample sizes. In addition to calculating the mean square errors, the distribution of errors was plotted for each statistic. The distributions of the errors of the estimates directly from the sample and from the PQDE are virtually identical. Graphical results for the first quartile are shown below.

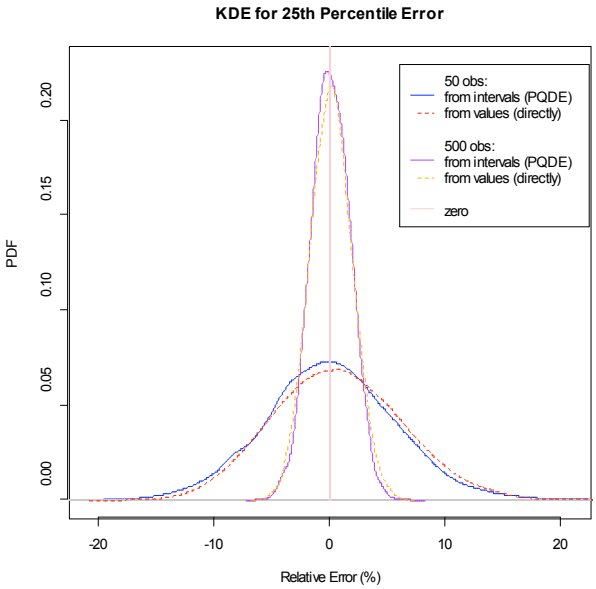


Figure: Kernel Density Estimate of the relative error of the first quartile from 5,000 samples of 50 and of 500 observations.

Several sample sizes in addition to 50 and 500, shown above, were also tested. As would be expected, below a certain threshold, the distributions of the relative error of the estimates from the PQDE become somewhat biased with most or all estimates resulting in a few possible values. This threshold seems to be around 25. I would also expect an upper limit beyond which estimates form the sample would outperform those from the PQDE. Sample sizes up to 5,000 were tested and the distributions of the relative error of the estimates directly from the sample and from the PQDE are virtually identical in each.

3.2.2 Bimodal Distribution

For this distribution, the true mean is 33.450 and the true 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles are 18.759, 22.195, 31.500, 44.706, and 52.896, respectively.

n=50	Rel. MSE	
	PQDE	Samp
mean	35.6	32.5
10th	38.9	42.4
25th	42.9	45.7
50th	157.9	166.1
75th	43.2	41.7
90th	41.2	40.2

n=500	Rel. MSE	
	PQDE	Samp
mean	2.822	2.72
10th	3.706	3.918
25th	3.058	3.614
50th	18.69	22.4
75th	2.915	3.628
90th	4.041	3.768

As with the lognormal distribution in 3.2.1, these distributions of the relative errors of estimates directly from the sample and from the PQDE were virtually identical. Several sample sizes in addition to 50 and 500, shown above, were also tested. Results from these sample sizes were similar to those of the lognormal in section 3.2.1.

4. Discussion

The results above seem to suggest, for practical purposes, that is, for the distribution shapes and sample sizes which tend to occur for the occupational wage distributions in practice, the accuracy of the estimates made from the piecewise quadratic density estimator (PQDE) is essentially the same as the accuracy of estimates calculated directly from a sample using the point data.

The PQDE has some exceedingly nice properties that may make it a powerful tool for large-scale surveys. It allows estimates to be made directly from interval data by incorporating information contained in relationships between adjacent intervals rather than relying on information from outside sources. It is simple both in concept and computation; it can be easily described, easily automated, and quickly run. Furthermore, the results above seem to suggest that, for distributions of practical interest for this survey, and possibly others, mean and percentile estimates from interval censored data using the PQDE are comparable in accuracy to those made directly from the non-censored samples. For surveys currently collection point data where this method could be applied, it could potentially translate to a very large reduction in collection costs with little to no loss in estimate accuracy.

I would be interested in generalizing this density estimator in a way in which consistency properties could be explored. I would also be interested in further exploring ways in which this density estimator could be altered to better smooth sample errors, possibly including bootstrapping.

Additionally, a practical issue particularly important for working surveys which is not covered here is aggregation. If the proposed density estimation method is applied separately to several sub-domains and to the whole domain, a weighted sum of the means of the sub-domains will not necessarily give the mean for the whole domain. Preliminary results in aggregation were encouraging but I have not explored this topic in depth.

Acknowledgements

There are several people who deserve recognition for related work. I'd like to thank my mentor Larry Ernst and supervisor John Eltinge for their generous guidance, the graduate students who worked out and tested the first versions of this density estimator at the 13th Annual Industrial Mathematical and Statistical Modeling Workshop: Zhen Chen, Rachel Fonstad, Suman Sanyal, Svetlana Simakhina, Stephanie Vance, and Chang Xu, especially Svetlana who proposed and coded the first versions of this density estimator, and the BLS' Modeling Team for OES-NCS Integration who read this work and offered insightful suggestions for extension, especially Steve Miller, Alan Dorfman, Matt Dey, and Mike Lettau.

References

Aitchison, J., and Brown, J.A.C.. The Lognormal Distribution with Special Reference to Its Uses in Economics. © 1957 Cambridge University Press.

Braun, John, Duchesne, Thierry, and Stafford, James. "Local Likelihood Density Estimation for Interval Censored Data." *The Canadian Journal of Statistics*. Vol.33. 2005.

"Density Estimation from Interval Data." Industrial Mathematical and Statistical Modeling Workshop for Graduate Students. Presenter: Meghan O'Malley. Students: Zhen Chen, Rachel Fonstad, Suman Sanyal, Svetlana Simakhina, Stephanie Vance, Chang Xu. Faculty Advisors: Ilse Ipsen and Xiaobiao Lin.

Hardle, Wolfgang. Smoothing Techniques: with Implementation in S. © 1991 Springer Verlag, New York.

van der Laan, Mark J., Peterson, Derick. "Smooth Estimation and Inference with Interval Censored Data." Technical Paper for University of California at Berkeley's Division of Biostatistics. 1997.

Peto, Richard. "Experimental Survival Curves for Interval-Censored Data." *Applied Statistics.* Vol. 22, No. 1, 1973, pp.86-91.

Silverman, B.W.. Density Estimation for Statistics and Data Analysis. © 1986 Chapman & Hall/CRC, Washington DC.

Turnbull, Bruce W.. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data." *Journal of the Royal Statistical Society.* Series B, Vol. 38, No. 3, 1976, pp.290-295.

Technical Notes for May 2006 OES Estimates. http://www.bls.gov/oes/current/oes_tec.htm