Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Nonparametric Random Variate Generation Using a Piecewise-Linear Cumulative Distribution Function

W. Kaczynski [a], L. Leemis [b], N. Loehr [c] & J. McQueston [b]

[a] Department of Mathematical Sciences, United States Military Academy, West Point, New York, USA

[b] Department of Mathematics, The College of William & Mary, Williamsburg, Virginia, USA

[c] Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

Available online: 20 Dec 2011

To cite this article: W. Kaczynski, L. Leemis, N. Loehr & J. McQueston (2012): Nonparametric Random Variate Generation Using a Piecewise-Linear Cumulative Distribution Function, Communications in Statistics - Simulation and Computation, 41:4, 449-468

To link to this article: http://dx.doi.org/10.1080/03610918.2011.606947

Communications in Statistics—Simulation and Computation®, 41: 449–468, 2012

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2011.606947

Nonparametric Random Variate Generation Using a Piecewise-Linear Cumulative Distribution Function

W. KACZYNSKI[1], L. LEEMIS[2], N. LOEHR[3], AND J. McQUESTON[2]

[1]Department of Mathematical Sciences, United States Military Academy, West Point, New York, USA
[2]Department of Mathematics, The College of William & Mary, Williamsburg, Virginia, USA
[3]Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

The standard approach to solving the interpolation problem for a trace-driven simulation involving a continuous random variable is to construct a piecewise-linear cdf that fills in the gaps between the data values. Some probabilistic properties of this estimator are derived, and three extensions to the standard approach (matching moments, weighted values, and right-censored data) are presented, along with associated random variate generation algorithms. The algorithm is a nonparametric blackbox variate generator requiring only observed data from the user.

Keywords Lifetime distributions; Modeling; Piecewise-linear functions; Simulation.

Mathematics Subject Classification 62-04; 62-07; 62G99.

1. Introduction

Simulation practitioners often advocate a "trace-driven" approach to input modeling, in which data values are sampled with equal probability. In the univariate case, this approach is equivalent to generating variates from the empirical cumulative distribution function (cdf)

$$\widehat{F}(x) = \frac{N(x)}{n} \quad -\infty < x < \infty,$$

where n is the sample size, $N(x)$ is the number of data values less than or equal to x, and $x_1, x_2, \ldots, x_n$ denote the data values. We limit the discussion here to the case of raw data, rather than grouped data.

Received November 9, 2010; Accepted July 15, 2011
Address correspondence to W. Kaczynski, Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996, USA; E-mail: william.kaczynski@usma.edu

The advantages to the trace-driven approach are that (a) it avoids any error that might be introduced by fitting the data with an approximate parametric model, and (b) the sampling technique is identical to bootstrapping (Efron and Tibshirani, 1993) and, hence, has well-established statistical properties.

The disadvantages to the trace-driven approach are that (a) no random variate can be generated between the data values, known as the interpolation problem, and (b) no random variate can be generated that is smaller than the smallest data value or larger than the largest data value, known as the extrapolation problem.

A standard technique for overcoming the interpolation problem is to replace the empirical cdf with a cdf which is piecewise linear between the data values (Banks et al., 2001, pp. 296–300; Law, 2007, pp. 309–310, 458; Leemis and Park, 2006, pp. 409–411). Since the $n-1$ gaps between the data values should assume equal weighting, the piecewise-linear cdf has the form

$$\tilde{F}(x) = \begin{cases} 0 & x < x_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - x_{(i)}}{(n-1)(x_{(i+1)} - x_{(i)})} & x_{(i)} \le x < x_{(i+1)};\ i = 1, 2, \ldots, n-1 \\ 1 & x \ge x_{(n)}, \end{cases}$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the order statistics, i.e., the data values sorted into ascending order. This cdf passes through the points

$$(x_{(1)}, 0), \left(x_{(2)}, \frac{1}{n-1}\right), \left(x_{(3)}, \frac{2}{n-1}\right), \ldots, (x_{(n)}, 1),$$

which we refer to as "knot points."

Example 1.1. Consider the univariate data set of $n = 6$ observations:

$$1 \quad 2 \quad 5 \quad 7 \quad 8 \quad 9.$$

We assume that these data values are drawn from a continuous population. The empirical cdf and piecewise-linear cdf are shown in Fig. 1. The piecewise-linear cdf strikes the risers of the empirical cdf; the first intersection occurs 1/5 of the way up the riser at $x = 2$ and the second intersection occurs 2/5 of the way up the riser at $x = 5$. This pattern continues until the piecewise-linear cdf strikes the top of the last riser at $x = 9$.

The probability density function (pdf) associated with the piecewise-linear cdf is constant between the data values:

$$\tilde{f}(x) = \frac{1}{(n-1)(x_{(i+1)} - x_{(i)})} \quad x_{(i)} \le x < x_{(i+1)};\ \ i = 1, 2, \ldots, n-1,$$

and it can be shown that the mean of this distribution is

$$E[X] = \frac{x_{(1)} + 2x_{(2)} + 2x_{(3)} + \cdots + 2x_{(n-1)} + x_{(n)}}{2(n-1)}.$$

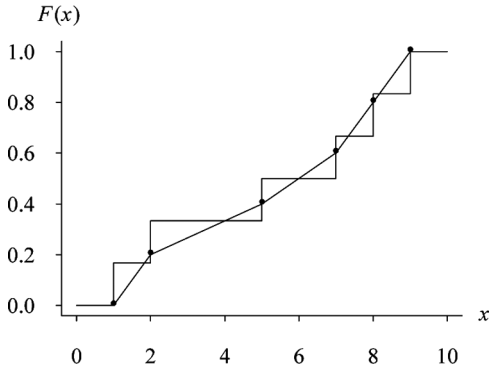

Figure 1. Empirical and piecewise-linear cdfs.

This weighted average of the data values places less weight on the extreme values, and equals $\bar{x}$, the sample mean of the data values, in only rare cases (e.g., a symmetric data set, where $\bar{x} - x_{(i)} = x_{(n+1-i)} - \bar{x}$ for $i = 1, 2, \ldots, \lfloor (n+1)/2 \rfloor$). The value of $E[X]$ approaches the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ in the limit as $n \to \infty$. (The match between the coefficients in the expression for $E[X]$ and the coefficients in the trapezoidal rule is discussed in Appendix A.) Likewise, a closed-form expression for the second moment is

$$E[X^2] = \sum_{i=1}^{n-1} \frac{x_{(i)}^2 + x_{(i)}x_{(i+1)} + x_{(i+1)}^2}{3(n-1)},$$

which can be utilized for computing the variance of the distribution. For the data set from Example 1.1, the mean and variance of the piecewise-linear estimate are $E[X] = 16/3$ and $\mathrm{Var}[X] = 71/9$.

Random variates can be generated efficiently by inverting the piecewise-linear cdf. Given $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and a random number generator, an $O(1)$ variate generation algorithm is:

Generate $U \sim U(0, 1)$

$i \leftarrow \lceil (n-1)U \rceil$

Return $\left(x_{(i)} + ((n-1)\,U - (i-1))\left(x_{(i+1)} - x_{(i)}\right)\right)$

The index i, which assumes one of the integers $1, 2, \ldots, n-1$ with equal likelihood, determines which linear segment to invert. Although this $O(1)$ algorithm is synchronized, monotone, and fast, there are four potential weaknesses that are described in the paragraphs below.

One potential weakness that arises with the piecewise-linear cdf $\widetilde{F}(x)$ occurs when there are tied values in the data set. These tied values result in a discontinuity in $\widetilde{F}(x)$. More specifically, d tied values at $x_{(i)}$ results in a discontinuity of height $d/(n-1)$ at $x_{(i)}$. The associated random variable is mixed (i.e., part discrete and part continuous), and the random variate generation algorithm will generate $x_{(i)}$ with probability $d/(n-1)$. If the modeler requires an absolutely continuous distribution, then it might be reasonable to use the midpoint of the discontinuity at $\widetilde{F}(x_{(i)})$ as the

knot point for the modified cdf. The variate generation algorithm would need to be modified appropriately.

A second weakness of the piecewise-linear approach is that data values that are close together (a common occurrence) lead to high peaks in the estimated density and an associated clustering of random variates near these particular data values. Two ways to overcome this weakness are to (a) use kernel density estimation, and (b) use the piecewise-linear approach on order statistics selected by discarding those with, for example, even indices. The pros and cons on these two alternative methods are addressed later in this article.

A third weakness is the extrapolation problem. Due to the finite endpoints of the piecewise-linear cdf, generating a variate below the first order statistic, $x_{(1)}$, and above the last order statistic of the sample, $x_{(n)}$, is impossible. Bratley et al. (1987) offered Marsaglia's tail algorithm as an elegant way to generate from the tail of a distribution. This approach proves useful in extending possible variate generation beyond just the sample range of a data set.

A fourth weakness of the standard piecewise linear approach is that it can only be applied to a complete data set. Right-censored data sets are commonly encountered in survival analysis; there is not an established technique for adapting the estimator to this type of data set.

In this article, we present three alternatives that overcome these weaknesses. The alternatives to the piecewise-linear cdf are nonparametric, thus avoiding potential error associated with a parametric model. They also allow some extrapolation below the minimum and maximum data values by stretching and translating observed data values such that the estimator's mean and variance match the sample mean and variance. Section 2 develops these variants in detail and Sec. 3 compares resulting estimators with estimates based on kernel density estimation. Section 4 gives a piecewise-linear survivor function associated with the Kaplan–Meier estimate.

2. Moment Matching and Weighted Observations

We consider two variations on the piecewise-linear cdf as a probabilistic model for a data set drawn from a continuous population. The first variation adjusts the knot points horizontally in the piecewise-linear cdf so that its first and second moments match those from the data set. The second variation adjusts the knot points vertically in the piecewise-linear cdf by allowing different weights for each of the data values.

2.1. *Matching Moments*

Occasions might arise when a modeler would like to (a) maintain the piecewise-linear nature of the cdf, (b) maintain the heights of the knot points at 0, $\frac{1}{n-1}, \frac{2}{n-1}, \ldots, 1$ (which implies fast variate generation), and (*c*) match the mean and variance of the piecewise-linear cdf to the sample mean and sample variance of the observations. This can only be achieved by adjusting the horizontal values of the knot points.

The expansion of the support of the piecewise-linear cumulative distribution function beyond the outermost data values may not be appropriate for all modeling situations. If the data values collected are service times in a queuing model, for instance, spreading the observations might result in a support that includes negative

service times. For the occasions when matching means and variances is appropriate, we derive the values of the knot points below. This derivation will maintain the ratios of the gaps between the data values so that their spreading is accomplished in the same way a bellows is spread on an accordion. We stretch the data to match variances first, then shift the data to match the means.

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the ordered raw data values as before and let $g_i = x_{(i+1)} - x_{(i)}$ for $i = 1, 2, \ldots, n-1$ be the ith gap between the observations. Let $g'_i = g_i / \sum_{j=1}^{n-1} g_j = g_i / \left(x_{(n)} - x_{(1)}\right)$ for $i = 1, 2, \ldots, n-1$ be the normalized gap values. If $x_{(1)}$ is shifted to $x'_{(1)} = x_{(1)} - \delta$ and $x_{(n)}$ is shifted to $x'_{(n)} = x_{(n)} + \delta$, the width of the support of the adjusted piecewise-linear cdf is $w = x_{(n)} - x_{(1)} + 2\delta$. To maintain the ratios of the normalized gap values, the adjusted data values are

$$x'_{(i)} = x_{(1)} - \delta + w \sum_{j=1}^{i-1} g'_j$$

for $i = 1, 2, \ldots, n$. The root finding problem now reduces to finding the value δ such that the unbiased sample variance of the original data values $x_1, x_2, \ldots, x_n$ matches the variance of the piecewise-linear cdf associated with the adjusted data values.

Once the variances have been matched, the means are easily matched by shifting each adjusted data value

$$x''_{(i)} = x'_{(i)} - \left[\frac{x'_{(i)} + 2x'_{(2)} + \cdots + 2x'_{(n-1)} + x'_{(n)}}{2(n-1)} - \bar{x} \right]$$

for $i = 1, 2, \ldots, n$. So finally, the knot points of the piecewise-linear cdf that matches first and second moments with the data are

$$\left(x''_{(1)}, 0\right), \left(x''_{(2)}, \frac{1}{n-1}\right), \left(x''_{(3)}, \frac{2}{n-1}\right), \ldots, \left(x''_{(n)}, 1\right).$$

Random variate generation via inversion is performed by the algorithm given in the introduction using the $x''_{(i)}$. Since the differences between the heights of adjacent knot points is constant, variate generation is fast. The stretching and shifting partially solves the extrapolation problem by allowing random variates to be generated outside of the range of the data values. Additionally, in the limit as $n \to \infty$, the sample variance s^2 approaches the population variance σ^2. Therefore, with increasing n, the value of δ is decreasing and as $n \to \infty$, $\delta \to 0$. Additionally, δ must exist since it is well known that the variance of the piecewise-linear estimator is always less than the variance of the sample data, and therefore, by construction, there exists $\delta > 0$ such that the adjusted data points equate the variance of the piecewise linear estimator and the sample variance of the data.

Example 2.1. Consider again the $n = 6$ data values $\{1, 2, 5, 7, 8, 9\}$. Find the piecewise-linear cdf knot values with matching means and variances. In order to match both the mean and variance, we first match the variances by stretching the data, then apply a shift that matches the means. For the ordered data values $x_{(1)} = 1$, $x_{(2)} = 2$, $x_{(3)} = 5$, $x_{(4)} = 7$, $x_{(5)} = 8$, $x_{(6)} = 9$, with gaps, $g_1 = 1$, $g_2 = 3$, $g_3 = 2$,

$g_4 = 1$, $g_5 = 1$, and associated normalized gaps, $g'_1 = 1/8$, $g'_2 = 3/8$, $g'_3 = 2/8$, $g'_4 = 1/8$, $g'_5 = 1/8$, the adjusted data values are

$$\begin{aligned}
x'_{(1)} &= 1 - \delta \\
x'_{(2)} &= 1 - \delta + (8 + 2\delta)\frac{1}{8} = 2 - \frac{3\delta}{4} \\
x'_{(3)} &= 1 - \delta + (8 + 2\delta)\frac{4}{8} = 5 \\
x'_{(4)} &= 1 - \delta + (8 + 2\delta)\frac{6}{8} = 7 + \frac{\delta}{2} \\
x'_{(5)} &= 1 - \delta + (8 + 2\delta)\frac{7}{8} = 8 + \frac{3\delta}{4} \\
x'_{(6)} &= 1 - \delta + (8 + 2\delta) = 9 + \delta.
\end{aligned}$$

The sample mean of the data is

$$\bar{x} = \frac{1 + 2 + 5 + 7 + 8 + 9}{6} = \frac{16}{3}$$

and the unbiased sample variance of the data is

$$s^2 = \frac{1}{5}\left[\left(1 - \frac{16}{3}\right)^2 + \left(2 - \frac{16}{3}\right)^2 + \cdots + \left(8 - \frac{16}{3}\right)^2 + \left(9 - \frac{16}{3}\right)^2\right] = \frac{32}{3}.$$

When the adjusted data values are used as arguments in the formula for the variance of the piecewise-linear cdf, the value of δ must satisfy the quadratic equation

$$\left[\frac{(1-\delta)^2 + (1-\delta)(2 - 3\delta/4) + 2(2 - 3\delta/4)^2 + \cdots + (8 + 3\delta/4)(9 + \delta) + (9 + \delta)^2}{(3)(5)}\right] - \left[\frac{(1-\delta) + 2(2 - 3\delta/4) + \cdots + 2(8 + 3\delta/4) + (9 + \delta)}{(2)(5)}\right]^2 = \frac{32}{3}$$

which reduces to

$$\frac{518}{75} + \frac{259\delta}{75} + \frac{259\delta^2}{600} = \frac{32}{3}.$$

This quadratic equation has positive root

$$\delta = -4 + \frac{80}{259}\sqrt{259} \cong 0.9710.$$

Selecting the negative root still matches the variance to that of the piecewise-linear cdf. However, selecting the negative root of the quadratic equation projects each of the original ordered data values about $(x_{(1)} + x_{(n)})/2$, which is only harmless for a symmetric data set. Finally, to match means,

$$x''_{(1)} = \frac{16}{3} - \frac{88}{259}\sqrt{259} \cong -0.1347$$

$$x''_{(2)} = \frac{16}{3} - \frac{68}{259}\sqrt{259} \cong 1.1080$$
$$x''_{(3)} = \frac{16}{3} - \frac{8}{259}\sqrt{259} \cong 4.8362$$
$$x''_{(4)} = \frac{16}{3} + \frac{32}{259}\sqrt{259} \cong 7.3217$$
$$x''_{(5)} = \frac{16}{3} + \frac{52}{259}\sqrt{259} \cong 8.5645$$
$$x''_{(6)} = \frac{16}{3} + \frac{72}{259}\sqrt{259} \cong 9.8072$$

are the x-values associated with the knot points.

An algorithm for adjusting the data values so that the first two moments of the piecewise-linear model match those of the raw data is given below (indentation denotes nesting).

Input data values $x_1, x_2, \ldots, x_n$

$$\bar{x} \leftarrow \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 \leftarrow \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sort the data values yielding $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

$w \leftarrow x_{(n)} - x_{(1)} + 2\delta$

for $i \leftarrow 1$ to $n-1$

$g_i \leftarrow x_{(i+1)} - x_{(i)}$

$g'_i \leftarrow g_i/(x_{(n)} - x_{(1)})$

for $i \leftarrow 1$ to n

$$x'_{(i)} \leftarrow x_{(1)} - \delta + w\sum_{j=1}^{i-1} g'_j$$

Find the positive root δ of the quadratic equation

$$\sum_{i=1}^{n-1} \frac{(x'_{(i)})^2 + x'_{(i)}x'_{(i+1)} + (x'_{(i+1)})^2}{3(n-1)} - \left[\frac{x'_{(1)} + 2\sum_{i=2}^{n-1} x'_{(i)} + x'_{(n)}}{2(n-1)}\right]^2 = s^2$$

for $i \leftarrow 1$ to n

$$x''_{(i)} \leftarrow x'_{(i)} - \left[\frac{x'_{(1)} + 2\sum_{i=2}^{n-1} x'_{(i)} + x'_{(n)}}{2(n-1)} - \bar{x}\right]$$

This piecewise-linear model associated with data values $x''_{(1)}, x''_{(2)}, \ldots, x''_{(n)}$ has a mean and variance that matches the mean and variance of the original data values. Appendix B in Kaczynski et al. (2012) contains an algorithm and associated S-Plus/R code for computing δ and $x''_{(1)}, x''_{(2)}, \ldots, x''_{(n)}$.

2.2. *Weighted Data Values*

An algorithm in the companion article Kaczynski et al. (2012), which concerns the generation of bivariate observations, requires a variant of the univariate piecewise-linear cdf approach which allows for the data values to be weighted. For $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, let $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$, where $\sum_{i=1}^{n} w_{(i)} = 1$, be the corresponding positive-valued weights. Any estimated cdf should collapse to $\widetilde{F}(x)$ when $w_{(i)} = 1/n$, $i = 1, 2, \ldots, n$. Although there is no claim made to the uniqueness of the estimator presented here, one approach is to first draw the empirical cdf associated with a discrete random variable X with support values $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and corresponding mass values $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$. Points on each of the risers can be connected to form a piecewise-linear estimated cdf. The only question that remains is what the heights of these points should be. One reasonable approach is to place the first knot point at $(x_{(1)}, 0)$, the second knot point $\frac{1}{n-1}$ of the way up the second riser (which is associated with $x_{(2)}$), the third knot point $\frac{2}{n-1}$ of the way up the third riser (which is associated with $x_{(3)}$), and so on. Using this approach is equivalent to connecting the points

$$\left(x_{(1)}, 0\right), \left(x_{(2)}, w_{(1)} + \frac{w_{(2)}}{n-1}\right), \left(x_{(3)}, w_{(1)} + w_{(2)} + \frac{2w_{(3)}}{n-1}\right), \ldots, \left(x_{(n)}, 1\right)$$

to form the piecewise-linear cdf. Define

$$y_{(i)} = w_{(1)} + w_{(2)} + \cdots + w_{(i-1)} + \frac{(i-1)w_{(i)}}{n-1} \quad i = 1, 2, \ldots, n,$$

as the height of each knot point. The piecewise-linear cdf for the weighted data values is

$$F^*(x) = \begin{cases} 0 & x < x_{(1)} \\ y_{(i)} + \dfrac{(y_{(i+1)} - y_{(i)})(x - x_{(i)})}{x_{(i+1)} - x_{(i)}} & x_{(i)} \le x < x_{(i+1)}; \; i = 1, 2, \ldots, n-1 \\ 1 & x \ge x_{(n)}. \end{cases}$$

This cdf reduces to $\widetilde{F}(x)$ in the equal-weighting case when $w_{(i)} = 1/n$, for $i = 1, 2, \ldots, n$. Using the associated pdf, it can be shown that $E[X]$ and $E[X^2]$ are

$$E[X] = \frac{1}{2} \sum_{i=1}^{n-1} \left(w_{(i)} + \frac{i w_{(i+1)} - (i-1) w_{(i)}}{n-1} \right) \left(x_{(i+1)} + x_{(i)} \right)$$

$$E[X^2] = \frac{1}{3} \sum_{i=1}^{n-1} \left(w_{(i)} + \frac{i w_{(i+1)} - (i-1) w_{(i)}}{n-1} \right) \left(x_{(i+1)}^2 + x_{(i+1)} x_{(i)} + x_{(i)}^2 \right).$$

To formulate an algorithm for variate generation, first sort the data, yielding the $x_{(i)}$ and $w_{(i)}$ values. Then, at the beginning of a simulation, calculate the $y_{(i)}$ values. The $O(n)$ algorithm for generating random variates given below also uses inversion.

Generate $U \sim U(0, 1)$

$i \leftarrow 1$

```
while (U > y_(i+1))
  i ← i + 1
return (x_(i) + (U − y_(i)) (x_(i+1) − x_(i)) / (y_(i+1) − y_(i)))
```

As expected, this algorithm collapses to the equally weighted algorithm given in Sec. 1 because $y_{(i)} = (i-1)/(n-1)$, for $i = 1, 2, \ldots, n$ in the equally weighted case. This algorithm can easily be modified to a $O(\log n)$ algorithm by employing a binary search rather than the linear search presented here.

Occasions might arise in which the weights need to be calculated from data. Consider the previous example. The data values 1, 2, 5, 7, 8, and 9 were stretched and translated so that the sample mean and variance matched the mean and variance of the piecewise-linear estimate. This resulted in the lowest data value $x_{(1)} = 1$ being shifted to $x''_{(1)} = -0.1347$. For certain types of data sets (e.g., service times), generating a negative service time might be unacceptable. So the only recourse for a modeler who wants to (a) keep the x-coordinates of the knot points at the data values and (b) match moments, is to adjust the weights $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$ to values other than the usual equally likely weights $1/n$. As seen earlier, the effect of moving from a data set to the piecewise-linear estimator is to decrease the variance. Thus, adjusting the weights will place increased weight on the extreme values (and therefore less weight on the middle values) so as to increase the variance.

One problem that arises from this approach to matching moments is that there will typically not be a unique solution for the weights that will match moments. We therefore introduce the objective function

$$\frac{\prod_{i=1}^{n} w_{(i)}}{\prod_{i=1}^{n} 1/n}$$

from the empirical likelihood literature (Owen, 2001) to achieve a unique solution. Thus, the optimization problem is nonlinear and is written with constraints as:

$$\begin{aligned}
\text{maximize} \quad & n^n \prod_{i=1}^{n} w_{(i)} \\
\text{subjectto} \quad & \frac{1}{2}\sum_{i=1}^{n-1}\left(w_{(i)} + \frac{iw_{(i+1)} - (i-1)w_{(i)}}{n-1}\right)\left(x_{(i+1)} + x_{(i)}\right) = \bar{x} \\
& \frac{1}{3}\sum_{i=1}^{n-1}\left(w_{(i)} + \frac{iw_{(i+1)} - (i-1)w_{(i)}}{n-1}\right)\left(x_{(i+1)}^2 + x_{(i+1)}x_{(i)} + x_{(i)}^2\right) \\
& \quad - \left(\frac{1}{2}\sum_{i=1}^{n-1}\left(w_{(i)} + \frac{iw_{(i+1)} - (i-1)w_{(i)}}{n-1}\right)\left(x_{(i+1)} + x_{(i)}\right)\right)^2 = s^2 \\
& \sum_{i=1}^{n} w_{(i)} = 1 \\
& w_{(i)} \geq 0.
\end{aligned}$$

This method is advantageous for certain types of positive data that might be close to zero, ensuring that negative x values are not created by stretching the data (e.g., positive service times). By choosing this method the x_i values are not affected.

Example 2.2. Consider again the univariate data set of $n = 6$ observations $\{1, 2, 5, 7, 8, 9\}$. Just as in Example 2.1, we assume that these data values are drawn from a continuous population. The sample mean and sample variance of the data are $\bar{x} = 16/3$ and $s^2 = 32/3$. Find the corresponding weights, w_i, for $i = 1, 2, \ldots, 6$ that solve the above nonlinear program.

This problem was solved in Microsoft Excel and Matlab, yielding the optimal weights $w_{(1)} = 0.3721$, $w_{(2)} = 0.0519$, $w_{(3)} = 0.0391$, $w_{(4)} = 0.0444$, $w_{(5)} = 0.0761$, and $w_{(6)} = 0.4165$. These weights maximize the objective function and match the sample mean and variance of the data to the mean and variance of the weighted piecewise-linear cdf. The small sample size results in heavy weights being placed on the extreme values in order to match the moments.

Because this is a nonlinear optimization program, the solution achieved is quite sensitive to the solver chosen and starting point provided. As expected, as the number of observations n increases, the optimization problem becomes more difficult to solve. The next example uses a common data set from survival analysis.

Example 2.3. Consider the univariate data set of $n = 23$ ball bearing failure times in millions of revolutions (Lieblein and Zelen, 1956):

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80,

68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

We assume that these data values are drawn from a continuous population. Find the corresponding weights, $w_{(i)}$ for $i = 1, 2, \ldots, 23$ that solve the above nonlinear

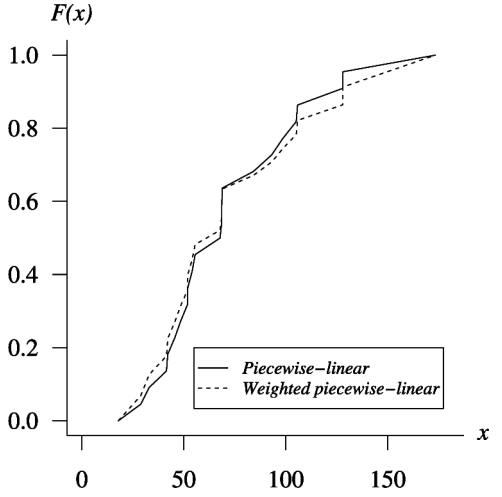

Figure 2. Piecewise-linear and optimal weighted piecewise-linear cdfs.

program. This problem was again solved in Microsoft Excel and Matlab, yielding the optimal weights $w_{(1)} = 0.0665$, $w_{(2)} = 0.0552, \ldots, w_{(22)} = 0.0471$, $w_{(23)} = 0.0850$.

Figure 2 shows the piecewise-linear cdf for this data and is overlaid with the optimal weighted piecewise-linear cdf matching the sample means and variances.

3. Comparing Estimates

Thus far, four estimates have been suggested for generating from a given continuous data set. They are (a) the piecewise-linear cdf, (b) the piecewise-linear cdf with a mean and variance matched to the data, (c) the weighted piecewise-linear cdf, and (d) the piecewise-linear cdf created by order statistics associated with discarding even indices. These methods all provide a means for variate generation via inversion, thus are fast, synchronized, and exact. Their main competitor in the literature is variate generation from an estimated density known as the kernel density. For a detailed discussion of this method, see Silverman (1986). To compare results for these estimates, a Monte Carlo simulation study was conducted in which estimates were created from six parametric distributions. These distributions were selected to adequately cover decreasing failure rate (DFR), increasing failure rate (IFR), increasing/decreasing failure rate (IFR/DFR), bathtub (BT), and upside-down bathtub (UBT) hazard functions. A sample was generated from each distribution, and the corresponding estimates were created. The metric developed for comparing the cdfs is

$$\frac{1}{nb}\sum_{j=1}^{b}\sum_{i=1}^{n}\left|F\left(x_{(i)}\right)-\ddot{F}_j\left(x_{(i)}\right)\right|,$$

where $F\left(x_{(i)}\right)$ is the cdf for the known population distribution at $x_{(i)}$, $\ddot{F}_j\left(x_{(i)}\right)$ is the corresponding jth cdf estimate at $x_{(i)}$ for one of the estimates listed below, n is the sample size, and b is the number of simulation replications. The average absolute errors for various sample sizes are given in Table 1, each for $b = 1,000,000$ replications. Common random numbers were used in the simulation experiments. The results can be replicated using the S-Plus/R `set.seed(123)` command. The smallest metric in each column is set in boldface type. The four estimators that are compared are: (a) the piecewise-linear estimator $\tilde{F}(x)$; (b) the moment matching piecewise-linear estimator $F^*(x)$; (c) the selected order statistic estimator $F^g(x)$; and (d) the kernel estimator $F^k(x)$.

The selected order statistic estimator breaks up the clumping that occurs with random sampling by deleting every order statistic with an even index and using the piecewise-linear estimator on the remaining order statistics. This is why the sample sizes are chosen to be odd. The weighted piecewise-linear cdf method is not included in the study due to the CPU time required to solve multiple replications of the optimization problem. Two kernel functions were selected for the study: (a) the standard normal and (b) U(−1, 1). The bandwidth parameter used for each kernel density estimate is the optimal bandwidth parameter (Silverman, 1986) described by $b = \alpha(k)1.364\min(s, R/1.34)n^{-1/5}$, where $\alpha(k) = 0.776$ for the Gaussian kernel, $\alpha(k) = 1.351$ for the uniform kernel, s is the sample standard deviation, and R is the sample range. The results for the uniform kernel density were not included because the estimates had gaps in their support. As expected, the kernel density estimate dominates for distributions with a pronounced non-zero mode. However,

Table 1
Average absolute error

	class	$\widetilde{F}(x)$	$F^*(x)$	$F^g(x)$	$F^k(x)$
$n = 9$					
Uniform(0, 1)	IFR	0.112	0.105	0.110	**0.091**
Weibull(1, 1/2)	DFR	0.238	0.229	0.250	**0.213**
Exponential(1)	IFR/DFR	0.112	0.118	**0.110**	**0.110**
Weibull(1, 2)	IFR	0.112	0.099	0.110	**0.092**
Exponential Power(1, 1/2)	BT	0.158	0.169	**0.143**	0.170
Arctan(1, 1)	UBT	0.190	0.164	0.201	**0.161**
$n = 21$					
Uniform(0, 1)	IFR	0.070	0.068	0.069	**0.061**
Weibull(1, 1/2)	DFR	0.219	0.216	0.222	**0.208**
Exponential(1)	IFR/DFR	0.070	0.094	**0.069**	0.088
Weibull(1, 2)	IFR	0.070	0.066	0.069	**0.062**
Exponential Power(1, 1/2)	BT	0.141	0.150	**0.134**	0.151
Arctan(1, 1)	UBT	0.164	**0.150**	0.168	0.151
$n = 45$					
Uniform(0, 1)	IFR	0.047	0.046	0.047	**0.043**
Weibull(1, 1/2)	DFR	0.211	0.210	0.212	**0.206**
Exponential(1)	IFR/DFR	**0.047**	0.082	**0.047**	0.073
Weibull(1, 2)	IFR	0.047	0.046	0.047	**0.043**
Exponential Power(1, 1/2)	BT	0.136	0.142	**0.133**	0.141
Arctan(1, 1)	UBT	0.152	**0.135**	0.154	0.146
$n = 71$					
Uniform(0, 1)	IFR	0.037	0.037	0.037	**0.035**
Weibull(1, 1/2)	DFR	0.208	0.208	0.209	**0.205**
Exponential(1)	IFR/DFR	**0.037**	0.076	**0.037**	0.067
Weibull(1, 2)	IFR	0.037	0.037	0.037	**0.035**
Exponential Power(1, 1/2)	BT	0.134	0.140	**0.132**	0.138
Arctan(1, 1)	UBT	0.148	**0.125**	0.149	0.144
$n = 101$					
Uniform(0, 1)	IFR	0.031	0.031	0.031	**0.030**
Weibull(1, 1/2)	DFR	0.207	0.207	0.208	**0.205**
Exponential(1)	IFR/DFR	**0.031**	0.072	**0.031**	0.062
Weibull(1, 2)	IFR	0.031	0.031	0.031	**0.030**
Exponential Power(1, 1/2)	BT	0.134	0.138	**0.132**	0.137
Arctan(1, 1)	UBT	0.146	**0.118**	0.146	0.143

the arctangent, exponential, and bi-modal exponential power distributions are more accurately estimated by one of the piecewise-linear cdfs. The matching moments estimator $F^*(x)$ for the exponential distribution deserves further explanation. When stretching values to match variances, negative values are possible, causing the excess error in the metric. We decided to leave this result as is in Table 1 with explanation

for emphasis. In conclusion, though we boldface only one error value for each row of the table (except where ties occur), in many cases the average error differences between methods appear to be negligible.

4. Generating Variates From Lifetime Data

Two distinguishing characteristics of lifetime data are a non negative response and the presence of right censoring. An example of right censoring in biostatistics is a cancer patient in remission, in which the time to recurrence is the lifetime of interest; an example of right censoring in reliability engineering is a spare part that has not failed, in which the time to failure is the lifetime of interest. This section develops a piecewise-linear estimate of the survivor function for a data set that contains right-censored observations. In lifetime data analysis, analysts often work with the survivor function (SF) rather than the cdf. The SF is the probability of survival to time x, written as $S(x) = P(X > x)$ for $x \geq 0$, and is complimentary to the cdf because $S(x) = 1 - F(x)$ for continuous populations (Meeker and Escobar, 1998). We use x rather that the more traditional t (for time) used in lifetime data analysis to be consistent with the notation used earlier in the article. The cases of complete and right-censored data sets are considered separately in the subsections that follow.

4.1. *Complete Data*

A complete data set contains no censored observations. Similar to the construction described in Sec. 1 for the empirical cdf, the empirical SF is

$$\widehat{S}(x) = 1 - \frac{N(x)}{n} \quad -\infty < x < \infty.$$

The empirical SF is a nonparametric estimate of the population SF. Generating variates from this estimate is accomplished by randomly sampling the data values with replacement, resulting in the same advantages and disadvantages of sampling from the empirical cdf defined in Sec. 1. The piecewise-linear SF is constructed in a similar manner to the piecewise-linear cdf presented in Sec. 1. The knot points for this SF estimate are:

$$(x_{(1)}, 1), \left(x_{(2)}, 1 - \frac{1}{n-1}\right), \ldots, \left(x_{(i)}, 1 - \frac{i-1}{n-1}\right), \ldots, \left(x_{(n-1)}, \frac{1}{n-1}\right), (x_{(n)}, 0).$$

Random variates can be generated efficiently by the $O(1)$ inversion algorithm given below, which requires the sorted data values $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and a random number generator.

Generate $U \sim U(0, 1)$

$i \leftarrow n - \lceil (n-1)U \rceil$

Return $\left(x_{(i)} + ((n-1)(1-U) - (i-1))\left(x_{(i+1)} - x_{(i)}\right)\right)$

4.2. *Right-Censored Data*

We now consider the more difficult case of a right-censored data set. The most common nonparametric SF estimator when censoring is present is the Kaplan–Meier step function estimator, also known as the product-limit estimator (Kaplan and Meier, 1958). The estimator is defined as

$$\widehat{S}(x) = \prod_{j \mid y_j < x} \left(1 - \frac{d_j}{n_j}\right),$$

where lifetimes are observed at times $y_1 < y_2 < \cdots < y_k$, d_j denotes the number of lifetimes observed at time y_j, $j = 1, 2, \ldots, k$, and n_j denotes the number of subjects on test just prior to time y_j, $j = 1, 2, \ldots, k$. This SF estimator takes downward steps at times $y_1, y_2, \ldots, y_k$. Censored observations between the downward steps in $\widehat{S}(x)$ result in a larger subsequent downward step due to the associated decrease in n_j.

In the special case when $d_k = n_k$ (the largest data value is an observed lifetime), the SF estimator drops to $S(x) = 0$ at its last step because $1 - \frac{d_k}{n_k} = 0$. In this case, no special treatment of the right-hand tail of the distribution is necessary for the piecewise-linear SF estimate.

The remaining problems associated with determining the piecewise-linear SF estimate for a right-censored data set are:

1. Where should the estimator strike the risers of the Kaplan–Meier SF estimate?
2. What should be done with the right-hand tail of the distribution when $d_k < n_k$?

We answer these questions in order.

The simplest approach to determine where the piecewise-linear SF estimate strikes the risers of the Kaplan–Meier SF estimate is to treat the heights of the downward steps in the Kaplan–Meier estimate as weights $w_{(1)}, w_{(2)}, \ldots, w_{(k)}$. The desire is to have the first knot point of the piecewise-linear SF estimate at $(y_1, 1)$, and the last knot point at $(y_k, \widehat{S}(y_k))$. Each subsequent knot point after the first will strike the associated riser an additional $1/(k-1)$ of the way down the riser, that is, the knot points occur at

$$(y_1, 1), \left(y_2, 1 - w_{(1)} - \frac{w_{(2)}}{n-1}\right), \left(y_3, 1 - w_{(1)} - w_{(2)} - \frac{2w_{(3)}}{n-1}\right), \ldots,$$
$$\left(y_k, 1 - w_{(1)} - w_{(2)} - \cdots - w_{(k)}\right).$$

This approach is consistent with the weighted observation approach in Sec. 2, and is illustrated in the following example.

Example 4.1. Consider the remission times from the treatment group of the 6–MP data set with $n = 21$ patients on test and $k = 7$ distinct observed remission times (Cox and Oakes, 1984). The data values, in weeks, are

$$6 \quad 6 \quad 6 \quad 6^* \quad 7 \quad 9^* \quad 10 \quad 10^* \quad 11^* \quad 13 \quad 16$$
$$17^* \quad 19^* \quad 20^* \quad 22 \quad 23 \quad 25^* \quad 32^* \quad 32^* \quad 34^* \quad 35^*,$$

Table 2
Product-limit calculations for 6–MP treatment case

j	y_j	d_j	n_j	$1 - \frac{d_j}{n_j}$
1	6	3	21	$1 - \frac{3}{21}$
2	7	1	17	$1 - \frac{1}{17}$
3	10	1	15	$1 - \frac{1}{15}$
4	13	1	12	$1 - \frac{1}{12}$
5	16	1	11	$1 - \frac{1}{11}$
6	22	1	7	$1 - \frac{1}{7}$
7	23	1	6	$1 - \frac{1}{6}$

where $*$ denotes a right-censored observation. Table 2 gives the values of y_j, d_j, n_j, and $1 - d_j/n_j$ for $j = 1, 2, \ldots, 7$. The knot points for the piecewise-linear SF estimate are

$$(6, 1), \left(7, \frac{101}{119}\right), \left(10, \frac{1408}{1785}\right), \left(13, \frac{184}{255}\right), \left(16, \frac{496}{765}\right), \left(22, \frac{592}{1071}\right), \left(23, \frac{160}{357}\right).$$

The product–limit SF estimate for all x values is plotted in Fig. 3, along with the knot points for the piecewise-linear SF estimate. Downward steps occur only at the $k = 7$ observed remission times. The vertical hash marks on the Kaplan–Meier SF estimate highlight censored values that occur between observed failure times; these occur at times 9, 11, 17, 19, and 20 in Fig. 3. If there is a tie between an observed failure time and censoring time (as there is at time 6 in this example) our convention of including the censored value(s) in the risk set means that there will be a larger downward step in the SF estimate following this tied value. The SF estimate is truncated at time 23, because that is the last observed failure time.

The remaining question concerns how to estimate the right-hand tail of the survivor function, that is, how should $S(x)$ be estimated for x values exceeding y_k

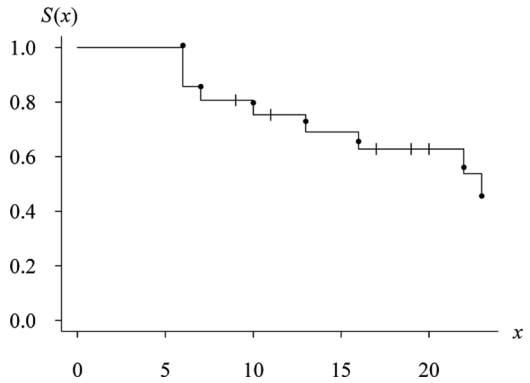

Figure 3. Product–limit survivor function estimate for the 6-MP treatment group.

when extrapolation is reasonable and required. Three possibilities for modeling the right-hand tail when $d_k < y_k$ follow.

- If the slopes of the piecewise-linear segments are nearly the same, it might be reasonable to use a linear segment that passes through the points $(y_1, 1)$ and $(y_k, \hat{S}(y_k))$ which is a line with slope

$$\frac{1 - \prod_{i=1}^{k} \left(1 - d_j/n_j\right)}{y_1 - y_k}.$$

  This method effectively appends the additional knot point

$$\left(y_1 - \frac{y_1 - y_k}{1 - \prod_{i=1}^{k} \left(1 - d_j/n_j\right)}, 0\right).$$

- If the linear segment for the right-hand tail does not seem reasonable, an exponential right-hand tail can be appended to the piecewise-linear SF estimate. The rate parameter for the exponential distribution is estimated by maximum likelihood as the ratio of the number of observed failures to the total time on test. The survivor function for the tail needs to be adjusted vertically or horizontally (both methods are equivalent by the memoryless property) so that it intersects the right-most knot point.
- Any lifetime distribution (for example, the Weibull distribution) can be fitted to the data set and used as a right-hand tail. The adjustment of the right-hand tail of the distribution should be done with care because the vertical and horizontal adjustments result in different tail distributions.

Modeling occasions might arise when it is advantageous to have the support of the piecewise-linear SF begin at 0 (or more generally some arbitrary warranty period w, where $0 < w < y_1$). The easiest way to proceed is to artificially add the data value 0 or w to the data set, although this will induce a significant bias, particularly for small values of n.

5. Conclusions & Further Work

The standard solution to the interpolation problem for Monte Carlo or discrete-event simulation uses a piecewise-linear cdf as a model. The variate generation algorithm is fast and trivial to implement. We have suggested three modifications to the original model: (a) stretching and shifting the original data values so that the mean and variance of the piecewise-linear cdf model matches the mean and variance of the sample values; (b) a modification to the model and variate generation algorithm to account for weighted observations; and (c) a modification to the model to account for right-censored data sets. These modifications could prove to be useful in further work associated with the generation of bivariate samples. Another important consideration is how these modifications are conducted. The first is to adjust the knot points horizontally so as to match the first and second moments. The second is to adjust the knot points vertically by weighting the data so as to match the first and second moments and solve an optimization problem. These two approaches can be combined so that the knot points can be adjusted both horizontally and vertically so as to match moments and optimize some measure

of interest (e.g., the minimum absolute area between the standard piecewise-linear cdf estimator and the adjusted cdf estimator). Although this may require solving a high-dimensional optimization problem, heuristics exist to solve these problems and the problem only needs to be solved once to develop a probability model. Once the model is developed, an $O(\log n)$ algorithm could be utilized to generate random variates.

We conclude with a summary of piecewise-linear and kernel density estimation pros and cons. The advantages to using the piecewise-linear estimator for variate generation include: (a) no decisions from the modeler, completely nonparametric; (b) easily extended to match sample mean and variance; (c) easily smoothed to minimize the effect of clustering of data values; and (d) extends to bivariate data without the assumptions and requirements demanded if using kernel density estimation.

As shown in Sec. 3, kernel density estimation performs better than the piecewise-linear estimators, but not universally so. The drawbacks encountered when using kernel density estimation were: (a) the arbitrary decisions left to the modeler; kernel density functional form; variance of kernel densities (smoothing parameter); (b) normal kernel density function implies an infinite left-hand tail (obviously inappropriate for certain types of data, e.g., survival times); and (c) the uniform density may leave undesired gaps and extend to negative values.

One way to overcome the clustering problem associated with the piecewise linear estimates is to delete even-numbered order statistics from the data set or to group the data into cells. The drawbacks associated with these approaches are: (a) grouping involves arbitrary decisions/parameters from the modeler; and (b) too much grouping may mask the shape of the distribution.

While we recognize the approach presented in this article "not ideal" in density estimation, our goal is not density estimation. The goal is nonparametric variate generation, thus density estimation can be considered as an unnecessary step. The method proposed is a turnkey operation, requiring only the observed data from the modeler. The extension to these methods to the bivariate case are considered in Kaczynski et al. (2012).

Appendix A. Relationship to the Trapezoidal and Simpson's Rule

Suppose that $F = F_X : \mathbb{R} \to [0, 1]$ is an unknown *continuous* cumulative distribution function (CDF) and $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with this distribution. Let

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \tag{1}$$

denote the particular values obtained in a given random sample, sorted into weakly increasing order. Our goal is to use this sample to estimate F, which will then be used to simulate further observations. (We additionally assume that the support of the population is positive for simplicity. If the lower bound of the support happens to be a finite negative value, the results given in this appendix can be achieved by shifting the data values and adjusting the associated moments.)

For convenience, assume that F is strictly increasing on some (unknown) interval of possible values $[a, b]$. Thus, $F : [a, b] \to [0, 1]$ is an invertible function with inverse $F^{-1} : [0, 1] \to [a, b]$. Letting $U \sim U(0, 1)$, the probability integral

transformation (Fishman, 2006, page 77) states that X and each X_i have the same distribution as $F^{-1}(U)$. In particular,

$$E[X^k] = \int_0^1 F^{-1}(u)^k \, du.$$

Furthermore, to simulate random observations from the distribution of X, we need only use a random number generator to generate random numbers $u \in [0, 1]$, and then compute $F^{-1}(u)$.

Let $y_{(i)} = (i-1)/(n-1)$ for $i = 1, 2, \ldots, n$. Given the input data (1), symmetry suggests that we estimate F by a piecewise-linear function for which $F_0(x_{(i)}) = y_{(i)}$ for $i = 1, 2, \ldots, n$. This is equivalent to estimating F^{-1} by a piecewise-linear function F_0^{-1} such that $F_0^{-1}(y_{(i)}) = x_{(i)}$ for $i = 1, 2, \ldots, n$.

More generally, we might postulate that F^{-1} is some continuous function (not necessarily piecewise linear) such that $F^{-1}(y_{(i)}) = x_{(i)}$ for $i = 1, 2, \ldots, n$. We can then use numerical integration techniques to estimate integrals involving the unknown function F^{-1}. This is easy to do, since the $y_{(i)}$'s form a partition of $[0, 1]$ into $n-1$ subintervals of equal length. For example, using the trapezoidal rule to estimate $E[X]$ gives

$$\begin{aligned} E[X] &= E[F^{-1}(U)] \\ &= \int_0^1 F^{-1}(u) \, du \\ &\approx \frac{1-0}{2(n-1)} \left(F_0^{-1}(y_{(1)}) + 2F_0^{-1}(y_{(2)}) + 2F_0^{-1}(y_{(3)}) + \cdots + F_0^{-1}(y_{(n)}) \right) \\ &= \frac{x_{(1)} + 2x_{(2)} + 2x_{(3)} + \cdots + 2x_{(n-1)} + x_{(n)}}{2(n-1)}. \end{aligned}$$

Of course, this is exactly the formula obtained by using a piecewise-linear approximation in Section 2.1. Similarly, the trapezoidal estimate of $E[X^2]$ is

$$E[X^2] = \int_0^1 F^{-1}(u)^2 \, du \approx \frac{x_{(1)}^2 + 2x_{(2)}^2 + \cdots + 2x_{(n-1)}^2 + x_{(n)}^2}{2(n-1)}.$$

We remark that this expression does not necessarily equal $\int_0^1 F_0^{-1}(u)^2 \, du$, but it is certainly one reasonable way to estimate $E[X^2]$. Note that both of our formulas give unbiased estimators for the mean and second moment of X, although these are not the usual unbiased estimators commonly employed in statistics.

The simplest approach to simulating observations from X is to use the piecewise-linear estimate F_0^{-1} for F^{-1}. One more advanced approach is to replace F_0^{-1} by some affine transformation $F_1^{-1} = cF_0^{-1} + d$, for suitable constants c, d. One way to proceed is to choose c and d so that $E[F_1^{-1}(U)]$ equals the sample mean of the $x_{(i)}$'s, and $\mathrm{Var}[F_1^{-1}(U)]$ equals the unbiased sample variance of the $x_{(i)}$'s. A related approach (which is a bit simpler computationally) is to choose c and d so that the trapezoidal estimates of $E[X]$ and $E[X^2]$ (computed with respect to F_1^{-1}) equal the corresponding sample moments (computed using the $x_{(i)}$'s). In more detail,

let $m_1 = \sum_i x_{(i)}$, $m_2 = \sum_i x_{(i)}^2$,

$$t_1 = \frac{x_{(1)} + 2x_{(2)} + \cdots + 2x_{(n-1)} + x_{(n)}}{2(n-1)}, \quad t_2 = \frac{x_{(1)}^2 + 2x_{(2)}^2 + \cdots + 2x_{(n-1)}^2 + x_{(n)}^2}{2(n-1)}.$$

Then we can choose c and d to satisfy

$$m_1 = ct_1 + d, \quad m_2 = c^2 t_2 + 2cdt_1 + d^2.$$

We then simulate random observations from X by generating random numbers $u \in [0, 1]$, and computing simulated values $cF_0^{-1}(u) + d$.

The preceding discussion suggests some tantalizing extensions. What if we used more advanced numerical integration techniques to estimate integrals involving the unknown function F^{-1}? For example, when $n - 1$ is even, we could use Simpson's Rule to estimate $E[X]$ and $E[X^2]$, which amounts to using piecewise-quadratic estimates of the functions $F^{-1}(y)$ and $F^{-1}(y)^2$. This leads to formulas such as

$$E[X] = E[F^{-1}(U)] \approx \frac{x_{(1)} + 4x_{(2)} + 2x_{(3)} + 4x_{(4)} + \cdots + 4x_{(n-1)} + x_{(n)}}{3(n-1)}.$$

One could then try to modify the associated piecewise-quadratic functions by affine transformations to attain a closer match to the sample mean and unbiased sample variance.

Acknowledgments

We thank Michael Lewis for his assistance in formulating and verifying solutions of the nonlinear optimization program. We also thank Barry Nelson for his ideas provided in applying empirical likelihood theory to weighting data values. We acknowledge support for this research from the NSF via grant DUE-0123022 and the Omar Nelson Bradley Foundation.

References

Banks, J., Carson, J. S., Nelson, B. L., Nicol, D. M. (2001). *Discrete-Event System Simulation*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Bratley, P., Fox, B. L., Schrage, L. E. (1987). *A Guide to Simulation*. 2nd ed. New York: Springer-Verlag.

Cox, D. R., Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fishman, G. (2006). *A First Course in Monte Carlo*. Belmont, CA: Duxbury/Thompson/Brooks/Cole.

Kaczynski, W. H., Leemis, L. M., Loehr, N. A., Taber, J. G. (2012). Bivariate nonparametric random variate generation using a piecewise-linear cumulative distribution function. *Communications in Statistics—Simulation and Computation* 41:469–496.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53:457–481.

Law, A. (2007). *Simulation Modeling and Analysis*. 4th ed. New York: McGraw–Hill.

Leemis, L. M., Park, S. K. (2006). *Discrete-Event Simulation: A First Course*. Upper Saddle River, NJ: Prentice Hall.

Lieblein, J., Zelen, M. (1956). Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research NBS* 57:273–316.

Meeker, W. Q., Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: Wiley Interscience.

Owen, A. B. (2001). *Empirical Likelihood*. Boca Raton, FL: CRC Press.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.