# Statistical Issues on Simulation Techniques in Structural Engineering

**Professor Y. P. Mack, Ph.D.**
**Department of Statistics, University of California**
**Davis, California 95616, USA**

mack@wald.ucdavis.edu

**Abstract**

The use of computer simulation techniques represents a major advancement in reliability assessment in structural engineering. While Monte Carlo methods have gradually come to be recognized as an important tool for many applied disciplines, its widespread acceptance has only occurred in recent years due to advancements in computer technology in terms of memory, speed, and cost. The excellent text by Marek, Guštar and Anagnos [1] put forward a road map as well as numerous examples to guide the reader who is interested in simulation-based assessment of reliability of structural designs subject to various stress loading schemes. Not only are Monte Carlo methods material-saving and time-saving for actual experimentation, they also provide useful teaching tools for the engineer-in-training in university curriculum.

One outstanding feature mentioned in [1] is the generation of random samples based on bounded histograms corresponding to different loading characteristics. Many well-known parametric models such as the normal and gamma distributions have unbounded supports. In order to employ these distributions to model structural engineering, truncation and normalization become necessary to produce bounded histograms. Moreover, in situations where parametric modelling is unsatisfactory, the program M-Star mentioned in [1] actually generates random data (for example, the wind loading effect may be presented by the WIND1 histogram) which may not correspond to any common parametric distributions. (There are many such histograms deployed in [1].) This feature, the simulation of data from a bounded histogram, not necessarily of the parametric type, is at the heart of the improved Monte Carlo techniques in reliability assessment.

In the same time period when there was significant advancement in computer technology, the subfield in statistics know as "nonparametric function estimation" has seen important developments as well. Various techniques in estimation such as the kernel method, splines, local polynomials, and wavelets have now become mainstream statistical tools. In particular, the theory and implementation of kernel density estimation, including the setting when the underlying distribution has bounded support, are by now well-understood.

Here, we propose to improve on the ”bounded histograms” technique mentioned earlier by the kernel method. A nonparametric method in the context of reliability assessment is especially relevant since the data sets are typically of sizes where asymptotics become effective. We demonstrate by an example the effect of generating random samples from a kernel density estimate(the smoothed histogram) on subsequent assessment of reliability.

**Key Words:** Density, support, histogram, binwidth, kernel, mean square error, smoothing, boundary, bootstrap.

# 1 Introduction

The excellent text by Marek, Guštar and Anagnos [1] represents a path-breaking effort towards the use of Monte Carlo techniques in structural engineering. Compelling discussions have already been presented in [1] as well as Marek and Brozzetti [2], Marek, Brozzetti and Guštar [3] in regard to the role and recognition of stochasticity in many situations encountered in reliability assessment.

At the heart of this advancing technology is the generation of random samples corresponding to the distribution of various characteristics in structural analysis. Traditionally, simulations were based on parametric models such as the Gaussian, the Poisson, and the Gamma distributions. Frequently a truncation/normalization is necessary to satisfy the bounded support requirement in real applications. An important contribution of [1] and [3] is the generation of random variates from distributions which may not be easily described in parametric forms. Loading characteristics such as described by the histograms of high wind (WIND1) and yield stress of steel A36 (A572) are among many such distributions included in the two texts mentioned above for Monte Carlo studies.

Presently in [1] and [3], the generation of random samples from these types of continuous models is accomplished via a piecewise uniform quantile approximation (see [3] and also Popela [4]). The resulting histogram is constructed with as many as 256 bins and the original data set can be of sizes up to 60,000. For most applications, whether pedagogical or professional, these specifications are more in keeping with the spirit of ”asymptotics” than many other applications of statistical estimation in disciplines elsewhere.

As structural engineers are moving from a deterministic to a probabilistic way of thinking, statisticians have also been breaking new grounds in nonparametric function estimation in recent times. Starting from the seminal work of Rosenblatt [5] on density estimation, which ushered in the by now well-established kernel technology, related developments that followed include splines, local polynomials and wavelets. These methods, though with their origins in numerical analysis, have become mainstream statistical tools nowadays. It is thus of some interest to explore Monte Carlo simulation by using these new tools. In the sequel, we will consider only the kernel method for definiteness and because the theory and practice behind this method are better understood. The motivation of our exercise is one of possible refinement. We endorse wholeheartedly the mission of this colloquium - that of incorporating randomness in reliability assessment in structural engineering.

## 2 Histogram and kernel estimation of a density

In this section, we will introduce the concept and notations of the histogram and kernel density estimation. Let $X_1, \ldots, X_n$ denote an i.i.d. sample with common cdf $F(x)$, density $f(x)$. Since we are interested in observations which are typically bounded, without loss of generality, we will assume the support of the distribution to be the unit interval [0,1]. For constructing the bounded (support) histogram (for approximating $f(x)$), we assume the unit interval $I$ is partitioned into $m$ equally-spaced bins as follows:

$$I = I_1 \cup I_2 \cup ... \cup I_m$$

where $I_k$= $[\frac{k-1}{m}, \frac{k}{m})$ for $k = 1, ..., m-1$; and $I_m$= $[\frac{m-1}{m}, 1]$. The binwidth is therefore $1/m$. Define the histogram estimate of $f(x)$ by the following: Let $F_n(x)$ be the empirical cdf based on the original data, i.e., let $F_n(x)$ be the relative cumulative frequency of the original observations falling below $x$. Also let $k(x)$ be that integer such that $I_{k(x)}$ contains $x$. Then

$$f_n(x) = F_n(I_{k(x)})/(\frac{1}{m}). \tag{1}$$

Thus $f_n(x)$ is the relative frequency of the interval $I_{k(x)}$ normalized by the binwidth. Note that $n \cdot F_n(x)$ is a binomial r.v. with $n$ trials and success probability $P(X_j \in I_{k(x)})$. For $n$ sufficiently large, a Taylor expansion of $f$ in a neighborhood of $x$ plus standard calculations yield

$$bias(f_n(x)) = Ef_n(x) - f(x) = \frac{f'(x)}{2}[2(\frac{k(x)}{m} - x) - \frac{1}{m}] + O(\frac{1}{m^2}), \tag{2}$$

$$var(f_n(x)) = f(x) \cdot \frac{m}{n} + o(\frac{m}{n}). \tag{3}$$

Hence the (asymptotic) mean square error

$$MSE(f_n(x)) = (bias)^2 + variance$$

converges to 0 if $m = m(n) \to \infty$ sufficiently slowly so that $m/n \to 0$ as $n \to \infty$. Next, we define the *kernel estimate* (see [5])

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{b}w\left(\frac{x - X_j}{b}\right), \tag{4}$$

where $w$ is called the *kernel* ($w$ is typically a symmetric density) and $b$ is called the *binwidth*. $\tilde{f}_n$ can be thought of as the generalization of a "moving histogram" when $w$ is the uniform density on [-1,1], for then

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2b} I\{\left|\frac{x - X_j}{b}\right| \leq 1\} = \frac{1}{n}\sum_{j=1}^{n} I\{X_j \in [x - b, x + b]\} \div (2b), \quad (5)$$

where $I\{A\}$ is the indicator of an event A. In words, the moving histogram is just the relative frequency of the interval $[x - b, x + b]$ normalized by binwidth. For the kernel estimate given by (4), if $w$ is a symmetric kernel, for $n$ sufficiently large, standard calculations involving a Taylor expansion just as in the histogram consideration lead to

$$bias(\tilde{f}_n(x)) = [\frac{f''(x)}{2}\int u^2 w(u)du] \cdot b^2 + o(b^2), \quad (6)$$

$$var(\tilde{f}_n(x)) = [f(x)\int w^2(u)du] \cdot \frac{1}{nb} + o(\frac{1}{nb}.) \quad (7)$$

As in the histogram case, if $b = b(n) \to 0$ in such a way that $nb \to \infty$ as $n \to \infty$, then the MSE$(\tilde{f}_n(x)) \to 0$ as $n \to \infty$. Since $b$ is a gauging quantity balancing between the bias and the variance, it is also called a *smoothing parameter.* In other words, the kernel method is a smoothing procedure. An inspection of expressions (2), (3), (6) and (7) reveals (by equating $1/m$ and $b$) that the variance of both estimates are of the same rate of convergence, but the smoothed estimate has a bias which decays faster than the histogram. Of course the MSE consideration is only a basic performance criterion. For a more detailed and thorough treatment of the theory and practical implementation of the kernel method, please refer to Silverman [6], Härdle [7], Scott [8], and Wand and Jones [9].

## 3 Estimating a density with bounded support

In Sec. 2, the construction of a histogram assumes $f$ has bounded support *à priori*. For the kernel method, formulae (6) and (7) are valid if $x$ is an *interior* point of the support, i.e., $x$ is away from either endpoint of the support by at least one binwidth. Otherwise $x$ is said to be a *boundary* point. It can be shown that the bias of $\tilde{f}_n(x)$ at a boundary point becomes $O(binwidth)$ rather than $O((binwidth)^2)$. Intuitively this happens since a boundary point "sees" more sample observations towards the interior than the boundary, so that the cancellation effect of a symmetric kernel cannot be exploited (as in the case of an interior point). One way to correct this is to use the so-called "boundary kernels" (see Müller [10]). Another method is to use the Richardson extrapolation (see Schucany and Sommers [11]). For either method, the correction must be made one boundary point at a time, at the cost of computational efficiency. The bias will become $O((binwidth)^2)$ again, but the variance will also increase.

Nevertheless, boundary correction is effective, especially when the sample size is moderate.

However, for sample sizes and binwidths employed in [1] in constructing a histogram (or a smoothed density estimate) for use subsequently in generating simulated samples, the variance term typically dominates the $(bias)^2$ term in the MSE decomposition. Thus boundary correction may not be that critical under this scenario. This means that the bounded support requirement does not pose a severe problem for either the histogram or the kernel approach, for the binwidths and sample sizes so chosen.

## 4 Histogram and kernel density estimate based reliability assessment

For simplicity, we will only consider reliability assessment based on a single loading r.v. $X$. We will assume $X$ is continuous with cdf $F(x)$, density $f(x)$, support $= [0, 1]$. In general, a reliability measure can be formulated as a functional $T(F)$. If $T$ is linear, then there is a function $t(x)$ such that

$$T(F) = \int t(x)dF(x) = \int t(x)f(x)dx. \tag{8}$$

When $F$ is unknown, but an i.i.d. sample of $X$ is available, then there are at least three possible estimators of $T(F)$

(i) $T(F_n) = \int t(x)dF_n(x) = \frac{1}{n}\sum_{j=1}^{n} t(X_j)$;
(ii) $T(f_n) = \int t(x)f_n(x)dx$;
(iii) $T(\tilde{f}_n) = \int t(x)\tilde{f}_n(x)dx$.

If repeated samples from $F$ are available, one can assess the standard error, or even the sampling distribution of $T(F_n)$ (likewise for $T(f_n)$ and $T(\tilde{f}_n)$ ). Absent repeated samples from $F$, one can generate samples from $F_n$ (or $f_n$, or $\tilde{f}_n$). This procedure is known as the *bootstrap* (see Efron [12]). If one were only comparing $T(F_n)$ with $T(\tilde{f}_n)$, there exists some study by Silverman and Young [13] who demonstrated that if $t(x)$ satisfies a certain differentiability condition, then there are situations under which the *smooth* bootstrap (the sample generated by a smoothed density estimate) performs better than the standard bootstrap (the samples generated by $F_n$). But their study did not consider $T(f_n)$, and their results are not applicable to functionals such as

$$T(F) = P(X > a) = \int I\{x > a\}dF(x) \tag{9}$$

for some known threshold value $a$. This particular functional corresponds to exceedance probability - a frequently encountered reliability measure.

For the remainder of this discussion, we will focus on the estimation of the quantity

$$\theta = P(X \leq a) = 1 - T(F).$$

Let $X_1(1), \ldots, X_N(1)$ be an i.i.d. sample from $f_n$, and let $X_1(2), \ldots, X_N(2)$ be an i.i.d. sample from $\tilde{f}_n$. Consider estimators

$$\hat{\theta}_1 = \frac{1}{N} \sum_{j=1}^{N} I\{X_j(1) \leq a\},$$

$$\hat{\theta}_2 = \frac{1}{N} \sum_{j=1}^{N} I\{X_j(2) \leq a\}.$$

Without going into details, it can be shown that

$$bias(\hat{\theta}_1) = -\frac{f'(a)}{2}[\frac{k(a)}{m} - a]^2 + \frac{f'(a)}{2m}[\frac{k(a)}{m} - a] + o(\frac{1}{m^2}), \tag{10}$$

$$bias(\hat{\theta}_2) = [\frac{f'(a)}{2} \int u^2 w(u) du] \cdot b^2 + o(b^2), \tag{11}$$

while the asymptotic variance of both estimators are, to the first order

$$F(a)[1 - F(a)] \cdot (\frac{1}{N} + \frac{1}{n}). \tag{12}$$

In terms of convergence rates, by corresponding $b$ with $1/m$, the bias terms (10),(11) are comparable. More importantly, (10) - (12) suggest that the MSE of both estimators of $\theta$ (hence in assessing the reliability $P(X > a)$) is dictated by the binwidth and sample size of the original data set, no matter how much larger the simulated sample size N is, relative to n.

# 5 Conclusion

Our preliminary investigation showed that for a reliability measure such as (9), using a smoothed (kernel) density estimate instead of a histogram at the initial stage in generating Monte Carlo samples does not necessarily produce substantial advantages in terms of MSE performance. For other functionals for which $t(x)$ (see(8)) satisfies a certain differentiability condition, one conjectures that results similar to those in [13] might hold, although such a conjecture still awaits a more rigorous verification. Finally, we conclude with the following comments:

(a) For some mixed distributions which have a discrete and a continuous component, such as those one occasionally encounters in certain failure models, it might be easier to use a histogram approximation rather than a procedure in which one has to deal with estimating a discrete component and then smoothing the continuous component separately.

(b) Ultimately, in order to gain wide acceptance of this Monte Carlo method, certainly much work is needed on the educational front. But for serious professional applications, it is imperative that the integrity of the initial samples (from which subsequent Monte Carlo simulations are based) be of the highest standard. One might even envision that in the future, specifications of the stochastic features of loading materials are routinely supplied by manufacturers, in compliance with official requirements.

### References

1. Marek, P., Guštar, M., Anagnos, T. *Simulation-Based Reliability Assessment for Structural Engineers.* CRC Press, Inc. Boca Baton, Florida, 1995.

2. Marek, P., Brozzetti, J. *From Deterministic to Probabilistic Way of Thinking in Structural Engineering.* AECEF Newsletter 1/2001. The Assoc. of European Civil Engineering Faculties, 2001.

3. Marek, P., Brozzetti, J., Guštar, M. *Probabilistic Assessment of Structures Using Monte Carlo Simulation. Background, Exercises, Software.* Published by ITAM CAS CZ, Prague, Czech Republic, 2001.

4. Popela, P. *Random Number Generators in Technical Applications.* Report GACR (Appendix A), 103/94/0562, Prague ITAM CAS, Czech Republic, 1966.

5. Rosenblatt, M. *Remarks on Some Nonparametric Estimates of A Density Function.* Annals of Math. Statist. Vol. 27, 832 - 837, 1956.

6. Silverman, B.W. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

7. Härdle, W. *Smoothing Techniques with Implementation in S.* Springer-Verlag, New York, 1990.

8. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York, 1992.

9. Wand, M.P., Jones, M.C.*Kernel Smoothing.* Chapman and Hall, London, 1995.

10. Müller, H.-G. *Smooth Optimal Kernel Estimators near Endpoints.* Biometrika. Vol. 78, 521-530, 1991.

11. Schucany, W.R., Sommers, J.P. *Improvement of Kernel Type Density Estimators.* J. of Amer. Statist, Assoc. Vol. 72, 420-423, 1977.

12. Silverman, B.W., Young, G.A. *The Bootstrap: To Smooth or Not to Smooth?* Biometrika. Vol.74(3), 469-479, 1987.

13. Efron, B. *Bootstrap Methods: Another Look at The Jackknife.* Annals Statist. Vol. 7, 1-26, 1979.