2008

Length Bias in the Measurements of Carbon Nanotubes

Paul H. Kvam
University of Richmond, pkvam@richmond.edu

Length Bias in the Measurements of Carbon Nanotubes

Paul Kvam
H. Milton Stewart School of Industrial Engineering
Georgia Institute of Technology
Atlanta, GA, USA

Abstract

To measure carbon nanotube lengths, atomic force microscopy and special software are used to identify and measure nanotubes on a square grid. Current practice does not include nanotubes that cross the grid, and as a result, the sample is length biased. The selection bias model can be demonstrated through Buffon's Needle Problem, which was extended to general curves that more realistically represent the shape of nanotubes observed on a grid. In this paper, the nonparametric maximum likelihood estimator is constructed for the length distribution of the nanotubes, and the consequences of the length bias are examined. Probability plots reveal that the corrected length distribution estimate provides a better fit to the Weibull distribution than the original selection-biased observations, thus reinforcing a previous claim about the underlying distribution of synthesized nanotube lengths.

Keywords: Buffon's Needle Problem, Censoring, Goodness of Fit, Nonparametric, Probability Plot, Sampling, Selection Bias, Weibull, Weighted Distributions

1 Introduction

Carbon nanotubes are rolled-up nano-scale sheets of graphitic carbon used to enhance materials with tensile strength, thermal conductivity and electronic conductivity. In terms of tensile strength alone, they can be stiffer and stronger than any micro-scaled fibres previously manufactured. Nanotubes and similar nano-sized devices have garnered great attention in the industrial and academic world, although statistical research in nano-manufacturing is still in its early stages. In contrast to laboratory research, commercial applications have developed at a slow pace, primarily because of the enormous production costs of high-quality nanotubes.

A major challenge in the production of nanotubes is controlling and analyzing the important qualities such as length, diameter and strength. Nanotube lengths can be measured using atomic force microscopy (AFM) and special software developed by SIMAGIS ®, which processes three dimensional images of the bulk nanotube material samples. The software identifies and classifies nanotubes and then analyzes nanotube lengths while recognizing the intersections, contacts and three dimensions of curvature of the tubes. The solution features also include the ability to filter objects according to selected length and height. In this paper, we will focus on length measurements. Figure 1 represents an example of an AFM profile image of carbon nanotubes in a $5\mu m$ x $5\mu m$ square interval. The average size of these nanotubes is nearly 0.5 μm, or 500 nm. This recent technological innovation gives experimenters the ability to ascertain distributional properties of carbon nanotubes beyond the simple mean lengths, as characterized in Ziegler, et al. (2005). Important mechanical properties that are attainable include tensile strength, elastic modulus and fracture toughness.

Wang, et al. (2006) present a statistical characterization of nanotube length using an AFM sample and the SIMAGIS software. The set of length measurements for 651 detected nanotubes is summarized in a histogram in Figure 2. Like the image in Figure 1, the sampling area consisted of a square grid with sides measuring 5 μm. While it has been verified in smaller samples that the software accurately finds and measures nanotubes less than 10 nanometers, the software's count does not include nanotubes that cross the boundary of the square and the data are therefore length-censored. Although the censoring mechanism can be easily included in the statistical analysis, this sampling method introduces size bias into the analysis as well.

2 Length Bias

For a positive random variable X with density function $f(x)$, if the random variable is observed with a probability proportional to its size $w(x)$, then the distribution of the random sample obtained with this bias has density

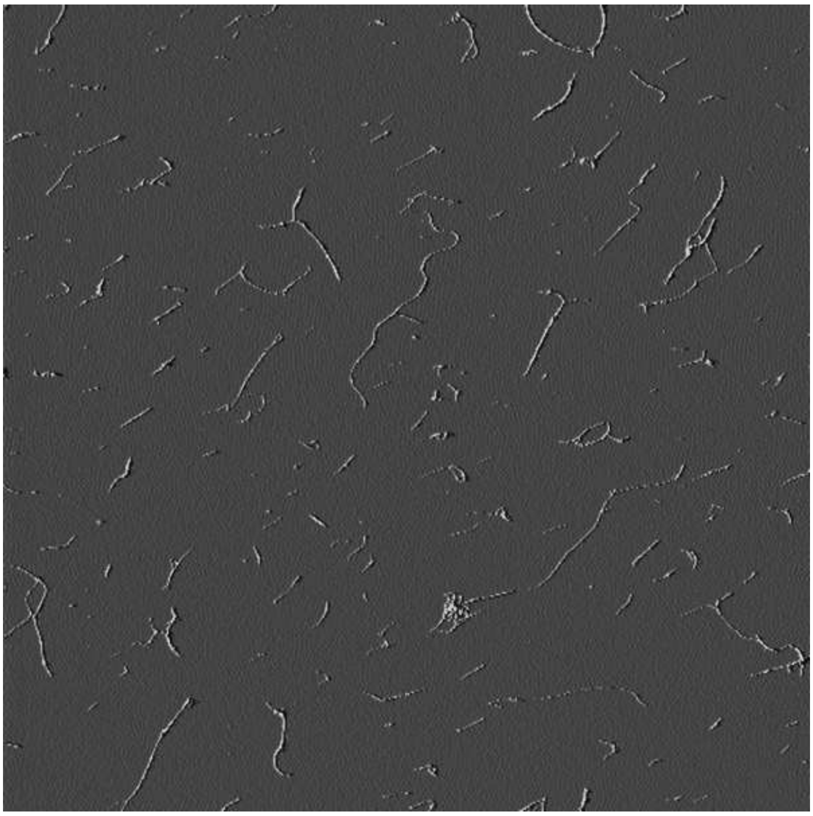

Figure 1: AFM image of Carbon Nanotubes.

$$f_w(x) = \frac{w(x)f(x)}{\int w(x)f(x)dx}.$$

This kind of weighted bias in sampling was first discussed by Rao (1965) with a particular interest in length bias, where $w(x) = x$. This holds for discrete distributions as well, and we will use $dF(x)$ to denote the more general class of distributions.

Length bias is a common phenomenon in survival studies in which prevalent cases are determined through cross-sectional study. For example, a test tends to detect more slow-growing cancers that take longer to become symptomatic compared to aggressive, fast-growing cancers. Cancers that grow slowly are easier to detect because they have a longer pre-symptomatic period when they are detectable, and thus show up with more relative frequency in such cancer studies. Horvath (1985) and Asgharian (2002) contain more detailed discussion on sampling and analysis problems induced by length bias.

2.1 Buffon's Needle Problem

Buffon's famous Needle Problem can be used to illustrate the length bias that takes place in this kind of two-dimensional sample. In 1733, Georges-Louis Buffon, a French naturalist

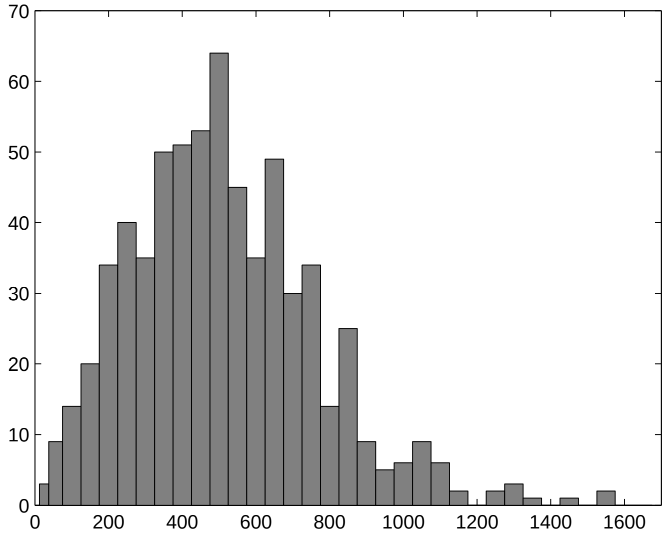

Figure 2: Histogram of dispersed SWNT lengths.

and mathematician, posed the following problem. Given a needle of length r and a grid of parallel lines spaced by an equal distance a, what is the probability that a randomly tossed needle will cross one of these lines? The problem is solved with rudimentary geometry based on assuming that the center of the needle and its angle with respect to the grid are uniformly distributed. The example has been featured as a classic probability problem and as an experimental way of obtaining an estimate of the constant π; see Perlman and Wichura (1975), for example. In this paper, we focus on the case in which $r \leq a$, where the censoring probability is $2r/(a\pi)$.

By 1812, Pierre-Simon Laplace considered a second grid of parallel lines constructed at right angles to the first set of parallel lines and with the same distance a, and found the crossing probability doubled. That is, the probability of a line crossing on a square boundary is $4r/(a\pi)$. A geometrical proof is offered by Mantel (1952).

Gnedenko (1962) provided a more general extension of the crossing probability to multi-sided convex polygons under the constraint that the diameter was less than a. Ramely (1969) reframed this extension in terms of arbitrary curves, and it is sometimes referred to as "Buffon's Noodle Problem". With polygons and crooked lines, the crossing frequency changes, but the number of expected crossings remains constant. To illustrate this, think of a needle bent in half. The probability the folded-over needle crosses a grid line is halved, but the expected number of crossings stays the same because the the needle will necessarily cross the the grid line twice.

To represent a nanotube of length r, we could treat it as a straight needle with crossing probability of $4r/(a\pi)$, or a crooked, curving tube that connects to itself, in which case the crossing probability is reduced to $2r/(a\pi)$. From Figure 2, the nanotubes exhibit only a modest curvature - much more like a straight needle than a connected polygon.

To illustrate the effect of curvature in a simple way, we assume any nanotube consists of two segments of equal length. Figure 3 shows the nanotubes with segment angles of $\theta = \{0, \pi/8, \pi/4\}$. If a nanotube consists of two segments of length $r/2$, the sum of the three side lengths of the inscribed triangle is $r(1 + \cos(\theta))$. Accordingly, the crossing probability would be reduced to

$$\frac{4r}{a\pi}\lambda, \tag{1}$$

where $\lambda = (\cos(\theta) + 1)/2$ represents the shrinkage in the crossing probability due to the bending of the nanotube. For $\theta = \{0, \pi/8, \pi/4\}$, we have $\lambda = \{1, 0.9619, 0.8536\}$. It might be sensible to treat λ as a random effect, reflecting the apparent heterogeneity of shapes observed in the sample of synthesized nanotubes. In this paper, we focus on a fixed, known λ, emphasizing the case of $\lambda = 1$ (straight nanotubes). The effects of curvature and randomness in the shape properties are discussed further in Section 4.

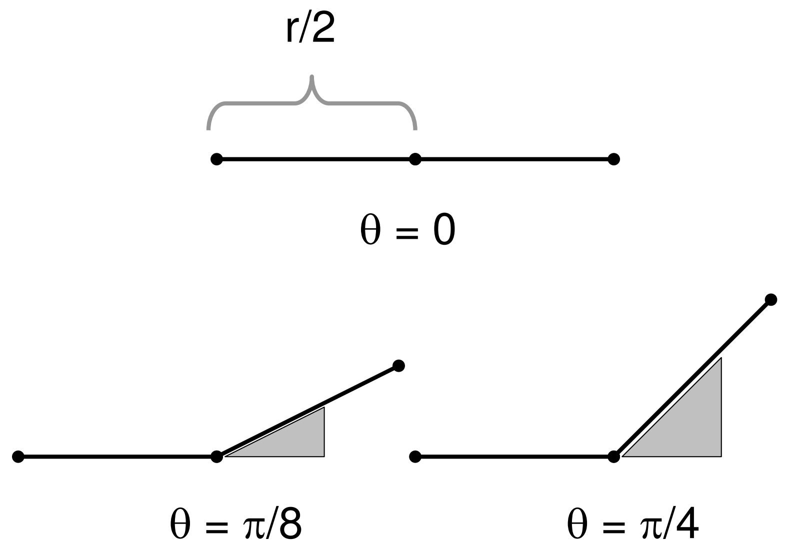

Figure 3: Nanotubes represented as bisected needles with angles $\theta = \{0, \pi/8, \pi/4\}$.

2.2 Censored Nanotubes

Obviously, a longer nanotube is more likely to be censored (and thus not measured) than a smaller one. Consequently, the sample that ignores the censored lines is biased according to the nanotube's length. Suppose the carbon nanotube lengths have cumulative distribution function $F_L(t)$. We will assume $F_L(t_0) = 1$ for some $t_0 < a$, though this constraint is not necessary for demonstrating the selection bias. Let $\mu_L = \int t dF_L(t)$. Based on the approximation in (1), the overall probability of censoring is

$$p = \int \frac{4\lambda t}{a\pi} dF_L(t) = \frac{4\lambda\mu_L}{a\pi}. \tag{2}$$

From the software output, we observe a sample of n nanotube measurements that has empirical distribution function $\tilde{F}_n(t)$. Let F_0 and F_1 be the distribution functions of the censored observations (nanotubes that cross the boundary) and uncensored observations, respectively.

From the mixture

$$F_L(t) = pF_0(t) + (1-p)F_1(t), \tag{3}$$

$\tilde{F}_n(t)$ is actually the nonparametric maximum likelihood estimator (NPMLE) for F_1 and not F_L. If the censoring probability p remains small, the bias from estimating $F_L(t)$ with $\tilde{F}_n(t)$ might be minuscule. For problems in which the nanotube length represents a more significant proportion of a, however, we will see the bias can be sizeable. From (3), the length bias ($w(x) = x$) affects both the censored population (F_0) and the observed population (F_1) as

$$dF_0(t) = \frac{t}{\mu_L} dF_L(t), \quad \text{and} \quad dF_1(t) = \frac{a\pi - 4\lambda t}{a\pi - 4\lambda\mu_L} dF_L(t), \tag{4}$$

where $0 \le t \le a$. While the censored observations are length biased, the observations lying strictly within the square are also biased in an opposite manner.

In the remainder of the paper, we restrict our consideration to distributions for which both F_0 and F_1 are proper distributions functions. From (4), we require that $F_L(t_0) = 1$ for some $t_0 < a\pi/4$.

2.3 Example

The following simple example illustrates how length bias can adversely affect the sample outcome if one does not compensate for it. Suppose the underlying distribution for (straight) nanotube lengths is Uniform$(0, 2\mu)$, observed on a unit grid (a=1). From $dF_1(t)$ in (4), the density function of the uncensored nanotubes is

$$f_1(x) = \left(\frac{1}{2\mu}\right) \frac{\pi - 4x}{\pi - 4\mu}, \quad 0 \le x \le 2\mu \tag{5}$$

where μ is the mean nanotube length. In this case, the mean for $F_1(t)$ is

$$\mu_1 = \mu \left(\frac{3\pi - 4\mu}{3\pi - 12\mu}\right).$$

Figure 4 shows the ratio mean for $F_L(t)$ over the mean of $F_1(t)$.

Once the mean nanotube length increases to 20% of the length of the grid, the increase in μ_1 will get out of hand quickly. The effect of length biasing will be further investigated for the nanotube data in the following section where we consider nonparametric estimation of the underlying length distribution.

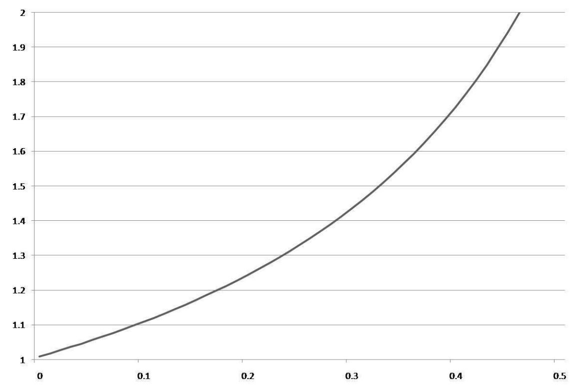

Figure 4: Ratio of μ_L over the mean of $F_1(t)$ (μ_1) for nanotube mean lengths up to 50% of grid size.

3 Estimating the length distribution

Although the bias for $dF_1(t)$ in (4) is not the commonly observed form of length bias, it presents nonetheless a straightforward selection-bias estimation problem that can be solved directly. If our observed sample is denoted by $x_1, \ldots, x_n$, then from Vardi (1985), the biased distribution can be expressed as some distribution function $G(t)$ such that

$$G(t) = \frac{\int_0^t w(u) dF_L(u)}{\int_0^\infty w(u) dF_L(u)}$$

where the weight function representing the bias is $w(x) = a\pi - 4\lambda x$. In this case, $G(t)$ is identifiable only for values of $t < a\pi/(4\lambda)$.

As a result, it can be shown from Theorem 1 and Theorem 2 in Vardi (1985) that the NPMLE of F_L is

$$\hat{F}_L(t) = \frac{\sum_{i=1}^n I(x_i \leq t)(a\pi - 4\lambda x_i)^{-1}}{\sum_{i=1}^n (a\pi - 4\lambda x_i)^{-1}}. \tag{6}$$

3.1 Properties of $\hat{F}_L$

Borrowing the method of proof from Theorem 3.2 in Vardi (1982), we can show the following result. We assume the coefficient representing nanotube curvature λ is fixed and known.

Theorem: Assume $F(t)$ is continuous for values $0 < t < a\pi/(4\lambda)$ and has bounded density with $F(0) = 0$ and $F\left(a\pi/(4\lambda)\right) = 1$. Then

$$\int_0^{a\pi/(4\lambda)} (a\pi - 4\lambda x)d\hat{F}_L(x) \longrightarrow \int_0^{a\pi/(4\lambda)} (a\pi - 4\lambda x)dF_L(x) \quad w.p.1$$

and

$$\sqrt{n}(\hat{F}_L(t) - F_L(t))$$

converges in distribution to a pinned Gaussian process with mean zero and covariance function $C(u, v)$, where, for $u \leq v \leq r$ and $\psi(t) = \int_0^t (a\pi - 4\lambda s)^{-1} dF_L(s)$,

$$C(u, v) = (a\pi - 4\lambda\mu_L)\left(\psi(u)(1 - F_L(v)) + F_L(u)\left(\psi(r)F_L(v) - \psi(v)\right)\right).$$

3.2 Nanotube data

In the analysis of the Wang, et al. (2006) data, we will use $\lambda = 1$ and refrain from making more elaborate assumptions about the apparent nanotube curvature displayed in the AFM image. Figure 5 shows the estimated survival function along with an approximate 95% confidence interval, based on the nonparametric percentile bootstrap procedure (with 1000 bootstrap samples) described in Chapter 15.2 of Kvam and Vidakovic (2007). Because of the large sample size, the coverage probability is not significantly improved using bias correction.

To see how the corrected nonparametric estimator compares with the uncorrected one, we can visually compare their density estimates. Figure 6 shows a smoothed version of the density associated with $\hat{F}_L(t)$ (dotted line) along side the biased estimator constructed from the unadjusted empirical density (solid line). The estimated mean and standard deviation using the standard estimator (530.5 nm, 261.4 nm) are not remarkably different from the mean and standard deviation of the data (510.4 nm, 253.2 nm). In this case, the mean nanotube length is on the order of 10 times smaller than the length of the side of the square from which they are sampled.

Obviously, if the nanotube lengths are more in proportion to the side of the square, the length bias can be much more dramatic. Figure 7 shows the original empirical density and estimators adjusted for two smaller (hypothetical) AFM image squares. The squares have sizes measuring 2500 and 2000 nm (roughly between four to five times bigger than the mean nanotube size). The estimated densities show how the NPMLE adjusts for more of the larger nanotubes being censored as the window size is reduced. For the 2500 nm^2 window, this is reflected in mass being transferred to the right, illustrating how substantial bias can be created by the missing censored observations. For the 2000 nm^2, the NPMLE

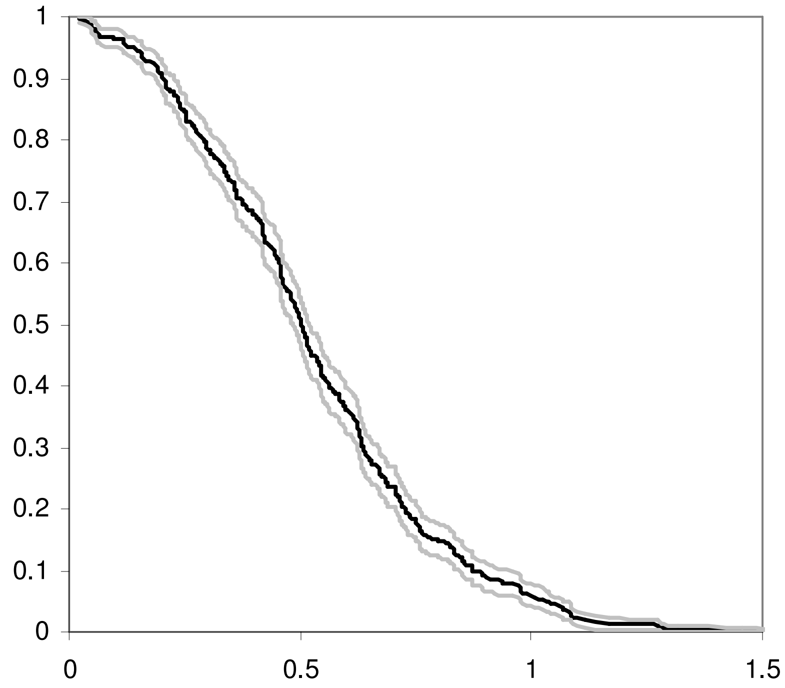

Figure 5: Nonparametric survival function (black line) and 95% confidence interval (gray lines) for carbon nanotube lengths.

has generated an extra mode for probability mass representing the lost observations due to censoring.

3.3 Effect of Shape

With sound knowledge about the curvature of nanotubes in the AFM image, models based on $\lambda < 1$ suggest that improvement in the estimation of the length distribution is possible. For the carbon nanotube data, however, changing the model by assigning $\lambda = 0.9$ will not change the NPMLE noticeably. For example, the estimated mean changes only 0.68%. Estimating the sample curvature has not been broached in the nanomanufacturing literature, although it posits an intriguing problem for future research. More importance has been placed on tube lengths, diameters and clusters than on how the nanotubes curve and bend.

Along with unknown curvature, heterogeneity within synthesized batches of carbon nanotubes has been carefully researched. See, for example, Lu and Bhattacharya (2005). Again, the randomness of the nanotube curvature has not received special attention beyond how the tubes will overlay with each other, which naturally affects electrical properties of nanodevices. Although the metric implied in Figure 3 is overly simplistic in many regards, here it will adequately serve the purpose of illustrating the effects of shape variability on the censoring, length-biasing and nonparametric estimation of nanotube length.

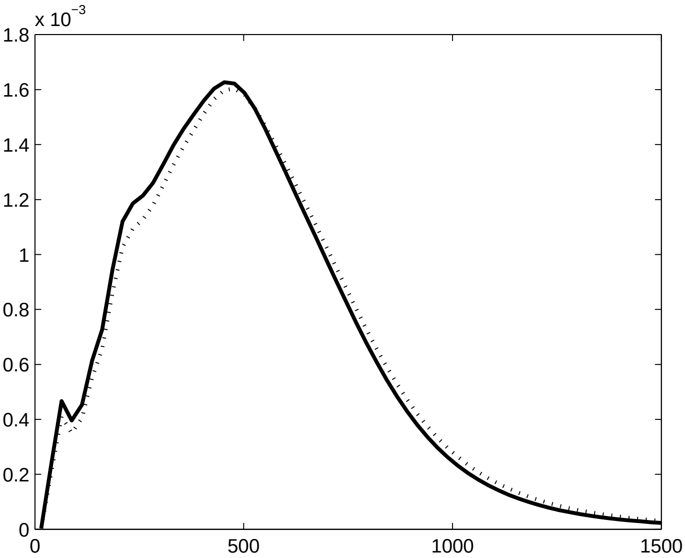

Figure 6: Empirical density (solid line) and adjusted density estimator (dotted line) for carbon nanotube lengths.

In this example, the carbon nanotubes are represented as bisected needles, and the angle of bend θ is governed by a uniform distribution, $\theta \sim U(0, \pi/4)$. The resulting distribution of λ (related to the arc since distribution) percolates uncertainty down through $\hat{F}_L$. In this example, the length distribution is $U(0, 2\mu)$, and F_1 is represented as a mixture with mixing density $g_\lambda(w) = 4\pi^{-1}(1 - w^2)^{-1/2}$, corresponding to $\lambda = (\cos(\theta) + 1)/2$. Through numerical integration, we can compare the variation between non-censored observations from the mixture. If $\mu \leq 0.25$, the standard deviation of the mixture sample is within 1% of the sample with fixed $\lambda = \pi/8$, which suggests that mild differences in nanotube curvature will not significantly affect the distribution of non-censored observations.

4 Discussion of Results

The problem of locating and measuring nano-sized matter has been greatly improved in recent years due to improving AFM technology and recognition software. However, if the sampling mechanism is unable to include observations crossing the boundary of the observation grid, the resulting length bias in the data set must be dealt with. If the bias is not addressed, the length distribution for nanotubes will be underestimated (along with the variability). In the example featured in Wang, et al. (2006), the consequences of the bias are relatively benign. However, other examinations of synthesized carbon nanotubes have included smaller regions, relative to the average size of the nanotubes. The length

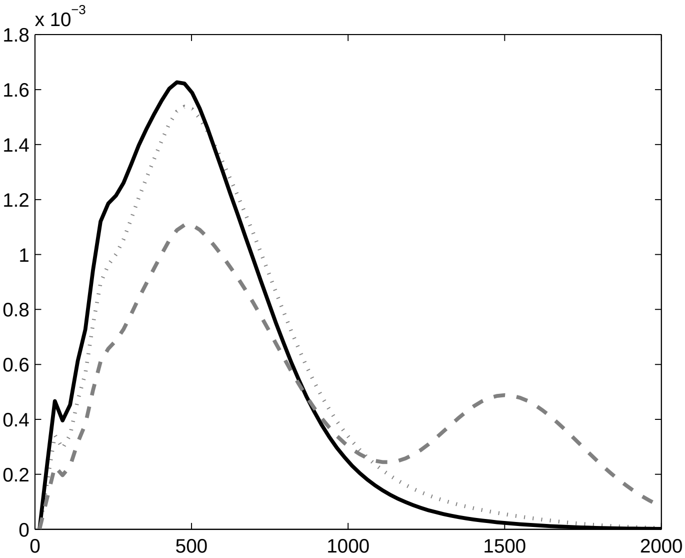

Figure 7: Empirical density (solid line) and adjusted density estimators based on different AFM image sizes. Original size is 5000 nm. Adjusted sizes are 2500 nm (dotted line) and 2000 nm (dashed line).

of synthesized nanotubes are increasing as the technology improves; Burke, et al. (2006) describe carbon nanotubes of 0.4 cm in length that have been developed in research labs. The analysis in this paper shows that the bias will have a substantial effect on the estimator of the length distribution if the mean nanotube length is as high as 20% of the size of the side of the square from which they are sampled.

Note that if censored observations were measured, the analysis would be straightforward using the Kaplan-Meier product limit estimator. In some cases, a nanotube located in the corner of a square can cross both grid boundaries, but in the nature of sampling nanotube lengths, the (right) censoring is the same.

In Wang, et al. (2006), the authors set to prove that the underlying length data can be characterized by the Weibull distribution. Indeed, the Weibull model has served as a precedent for similar distributional properties of fibrous materials. Based on the model for fiber composites in Fukuda and Kawata (1974) and Jayaraman and Kortschot (1996), Wang, et al. (2006) show the epoxy composite stiffness measurements are consistent with Weibull-distributed nanotube lengths. Although their probability plot of the original data in Figure 8 is not entirely convincing, one may conclude the Weibull presents "reasonable" distribution fit to the sample. Knowing the data are length biased, we can reconsider this goodness of fit for the NPMLE, $\hat{F}_L$. If we compare the original Weibull probability plot in Figure 8 to the corrected one in Figure 9, we see that the Weibull fit has actually improved slightly. This is confirmed by the decrease in the computed Kolmogorov-Smirnov test statistics. That is, the corrected estimator for the length-biased data offers stronger evidence to the conjecture made by Wang, et al. (2006).

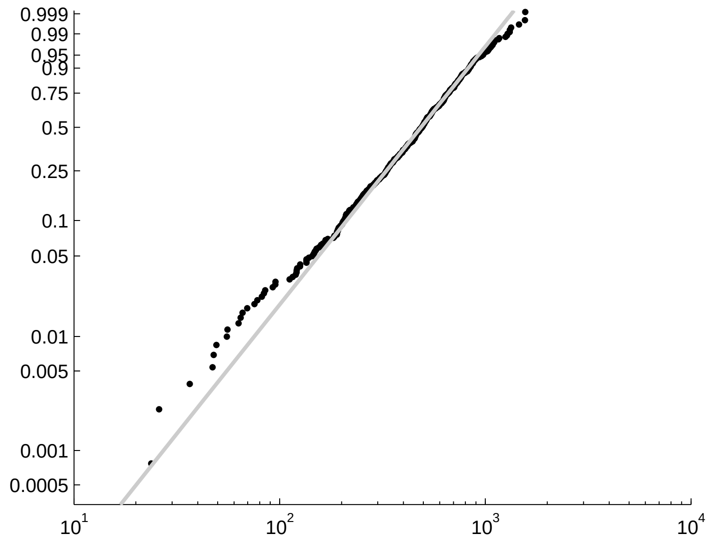

Figure 8: Original Weibull probability plot of the length-biased data for carbon nanotube lengths.

Acknowledgements

This research was supported by a National Science Foundation grant, CMMI-0700131. The author wishes to thank the associate editor and referees for their helpful comments that led to several improvements in the research paper.

References

[1] Asgharian, A., M'Lan, C. E. and Wolfson, D. B. (2002) "Length-biased sampling with right censoring: an unconditional approach", *Journal of the American Statistical Association*, Vol. 97, No. 457, 201 –209.

[2] Burke, P. J., Li, S., and Yu, J. (2006), "Quantitative Theory of Nanowire and Nanotube Antenna Performance", *IEEE Transactions on Nanotechnology*, Vol. 5, No. 4, 314 – 334.

[3] Fukuda, H. and Kawata, K. (1974) "On Young's modulus of short fiber composites", *Fibre Science and Technology*, Vol. 7, 207 – 212.

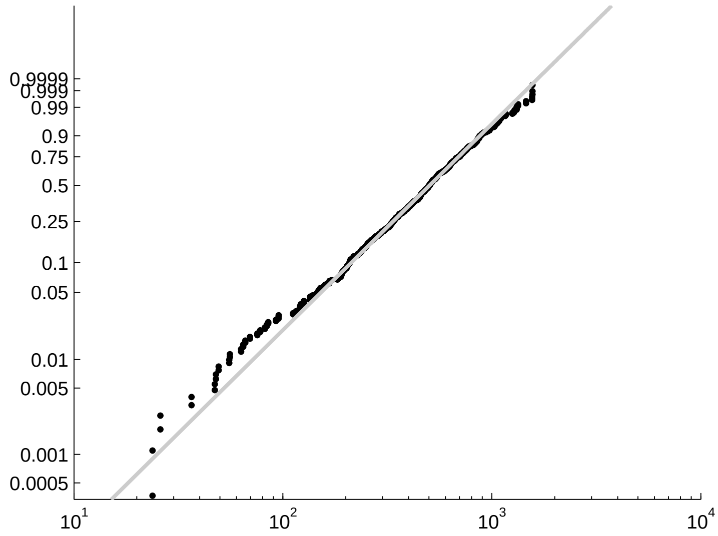

Figure 9: Weibull probability plot of the adjusted estimator of the carbon nanotube length distribution.

[4] Gnedenko, B. V. (1962) *The Theory of Probability*, MIS, Moscow.

[5] Horvath, L. (1985) "Estimation from a length-biased distribution", *Statistics and Decisions*, Vol. 3, 91 – 113.

[6] Jayaraman, K. and Kortschot, M. T. (1996) "Correction to the Fukuda-Kawata Young's modulus theory and the Fukuda-Chao strength theory for short fiber-reinforced composite materials", *Journal of Materials Science*, Vol. 31, 2059 – 2064.

[7] Kvam, P. H. and Vidakovic, B. (2007) *Nonparametric Statistics with Applications to Science and Engineering*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[8] Lu, Q. and Bhattacharya, B. (2005) "The Role Of Atomistic Simulations in Probing the Small-Scale Aspects of Fracture - A Case Study on a Single-Walled Carbon Nanotube." *Engineering Fracture Mechanics*, Elsevier, Vol. 72, No. 13, pp 2037-2071.

[9] Mantel, N. (1952) "An extension of the Buffon Needle Problem", *Annals of Mathematical Statistics*, Vol. 24, 674 – 677.

[10] Perlman, M. D. and Wichura, M. J. (1975) "Sharpening Buffon's needle", *The American Statistician*, Vol. 29, No. 4, 157 – 163.

[11] Rao, C. R. (1965) "On discrete distributions arising out of methods of ascertainment", in *Classical and Contagious Discrete Distributions*, ed. G. P. Patil, Pergamon Press and Statistical Publishing Society, Calcutta, 320 – 332.

[12] Ramaley, J. F. (1969) "Buffon's noodle problem", *American Mathematical Monthly*, Vol. 76, No. 8, 916 – 918.

[13] Sun, J. (2006) *The Statistical Analysis of Interval-censored Failure Time Data*, Springer Press, New York.

[14] Vardi, Y. (1982) "Nonparametric estimation in the presence of length bias", *The Annals of Statistics*, Vol. 10, 616 – 620.

[15] Vardi, Y. (1985) "Empirical distributions in selection bias models", *The Annals of Statistics*, Vol. 13, No. 1, 178 – 203.

[16] Wang, S., Liang, Z., Wang, B., Zhang, C. (2006) "Statistical characterization of single-wall carbon nanotube length distribution", *Nanotechnology*, Vol. 17, 634 – 639.

[17] Ziegler, K. J., Gu, Z., Peng, H., Flor, E. L., Hauge, R. H., and Smalley, R. E. (2005) "Controlled oxidative cutting of single-walled carbon SENTs", *Journal of the American Chemical Society*, Vol. 127, 1541 – 1547.