The Annals of Statistics
1982, Vol. 10, No. 2, 616–620

NONPARAMETRIC ESTIMATION IN THE PRESENCE OF LENGTH BIAS

BY Y. VARDI

Bell Laboratories

We derive the nonparametric maximum likelihood estimate, $\hat{F}$ say, of a lifetime distribution F on the basis of two independent samples, one a sample of size m from F and the other a sample of size n from the length-biased distribution of F, i.e. from $G_F(x) = \int_0^x u\, dF(u)/\mu$, $\mu = \int_0^\infty x\, dF(x)$. We further show that $(m+n)^{1/2}(\hat{F} - F)$ converges weakly to a pinned Gaussian process with a simple covariance function, when $m + n \to \infty$ and $m/n \to$ constant. Potential applications are described.

1. Introduction. If items' lengths are distributed according to the cumulative distribution function (cdf) F, and if the probability of selecting any particular item is proportional to its length, then the lengths of sampled items are distributed according to the cdf

$$G_F(y) \equiv G(y) = \frac{1}{\mu}\int_0^y x\, dF(x), \quad y \geq 0. \tag{1.1}$$

Here $\mu = \int_0^\infty x\, dF(x)$, and we assume $\mu < \infty$. Such a G is usually called the *length-biased distribution* (of F) and it arises naturally in many fields. A good survey of real-life applications of length biased (and other weighted) distributions, which includes many references, is Patil and Rao (1977). Additional interesting applications are given by Patil and Rao (1978) and Coleman (1979).

In this paper we consider the problem of finding a nonparametric maximum likelihood estimate (NPMLE) for the cdf F on the basis of two independent samples: a sample $\{X_i; i = 1, \cdots, m\}$ from F and a sample $\{Y_i; i = 1, \cdots, n\}$ from the length-biased distribution G. There are two types of applications for such an estimator. As an illustration of the first type, consider m independent identically distributed stationary renewal processes, with a common underlying cdf F, and suppose that from each process we sample the inter-event interval that happens to include a fixed time point t (assumed to be independent of the process itself), and subsequent inter-event intervals. This sampling scheme (cf. the "inspection paradox" in Feller (1971) page 187) gives rise to a sample (of size m) from G and a sample from F. Our estimate provides the NPMLE of F on the basis of these two samples. To illustrate the second type of application, suppose that for each of the processes above we know only the times of "events" in the renewal process between t and $t + h$ where t and $h > 0$ are fixed and independent of the processes. The data will then be affected by length-biasing coupled with censoring, and our estimator is then needed in each iteration of an algorithm that derives the NPMLE for the common underlying cdf F on the basis of this data. The details of the algorithm are given in Vardi (1982).

After describing the estimator $\hat{F}$ in Section 2, in Section 3 we give a summary of its asymptotic properties, which are analogous to those of the empirical distribution function in the case of ordinary sampling from a distribution. In particular we show that if $(n + m) \to \infty$ and $n/m \to$ constant, then $(n + m)^{1/2}(\hat{F} - F)$ converges weakly to a pinned Gaussian process with a simple covariance function. This convergence result should simplify the

Received February 1981; revised November 1981.
AMS 1970 *subject classifications.* Primary 62G05; secondary 62D05, 60F05.
Key words and phrases. Empirical distribution function, biased sampling, maximum likelihood, weighted distribution.

derivation of the asymptotic distribution of various statistics based on $\hat{F}$, and is useful in setting asymptotic confidence intervals for parameters of interest.

In order that this paper not suffer from a form of length-biasing, only outlines of proofs are given (Section 4). Full details are given in Vardi (1981).

2. The estimate. Let $X_1, \cdots, X_m$ $(m \geq 1)$ be independent, identically distributed random variables with a common cdf F which satisfies $F(0) = 0$ and $\mu \equiv \int_0^\infty x\, dF(x) < \infty$, and let $Y_1, \cdots, Y_n$ $(n \geq 1)$ be independent, identically distributed random variables with the common cdf G of (1.1). We denote the values taken by the pooled sample, $\{X_i\} \cup \{Y_i\}$, ordered from smallest to largest, by $t_1 < \cdots < t_h$ $(h \leq n + m$ because of possible ties). We further denote by ξ_i and by η_i the multiplicity of the X's and the multiplicity of Y's at t_i, respectively. Of course if F has a density with respect to the Lebesgue measure, then with probability one $h = n + m$, $\xi_i + \eta_i = 1$, and our notation is somewhat redundant. In practice, however, F, and hence G, might very well be discrete.

The probability of the data at hand is given by

$$P_F\{t_i, \xi_i, \eta_i;\ i = 1, \cdots, h\} = \prod_{i=1}^h \{dF(t_i)\}^{\xi_i} \left\{ \frac{t_i\, dF(t_i)}{\int_0^\infty u\, dF(u)} \right\}^{\eta_i}. \tag{2.1}$$

Clearly $P_F = 0$ if any t_i is a point of continuity of F, while $P_F > 0$ if $dF(t_j) > 0$, $1 \leq j \leq h$. It is easy to see that if F is such that $dF(s) > 0$ for some $s \notin \{t_1, \cdots, t_h\}$ then there exists another cdf, say F_1, which is discrete with strictly positive jumps at each of the points $t_1, \cdots, t_h$, and only there, and F_1 satisfies

$$P_{F_1}\{t_i, \xi_i, \eta_i;\ i = 1, \cdots, h\} > P_F\{t_i, \xi_i, \eta_i;\ i = 1, \cdots h\}.$$

Thus, in order to find a cdf that maximizes (2.1), we can restrict our search to the class of discrete cdf's which have positive jumps at each of the points $t_1, \cdots, t_h$, and only there. This reduces our task to finding the values of $p_j = dF(t_j), j = 1, \cdots, h$, that

$$\text{maximize} \quad L(p_1, \cdots, p_h) = \prod_{i=1}^h p_i^{\xi_i} \left(\frac{t_i p_i}{\sum_{j=1}^h t_j p_j} \right)^{\eta_i} \tag{2.2}$$

subject to $\sum_{j=1}^h p_j = 1, \quad p_j > 0, \quad j = 1, \cdots, h.$

If we denote the solution of (2.2) by $\hat{p} = (\hat{p}_1, \cdots, \hat{p}_h)$ then our estimate for F is of course

$$\hat{F}(x) = \sum_{t_j \leq x} \hat{p}_j \tag{2.3}$$

and it satisfies $P_{\hat{F}}\{t_i, \xi_i, \eta_i;\ i = 1, \cdots, h\} \geq P_{F_1}\{t_i, \xi_i, \eta_i;\ i = 1, \cdots, h\}$ for all cdf's F_1. We therefore call $\hat{F}$ a nonparametric maximum likelihood estimate (NPMLE) of F.

THEOREM 2.1 *The unique solution of* (2.2) *is*

$$\hat{p}_k = \frac{(\xi_k + \eta_k)\hat{\mu}}{nt_k + m\hat{\mu}} \quad k = 1, \cdots, h, \tag{2.4}$$

where $\hat{\mu}$ *is the unique solution for* a *in the equation*

$$\sum_{k=1}^h \frac{(\xi_k + \eta_k)t_k}{nt_k + ma} = 1. \tag{2.5}$$

The proof is straightforward. It is instructive to look at some of the properties of $\{\hat{p}_k\}$. First we observe that multiplying (2.4) by t_k and summing over k gives (using (2.5)) $\hat{\mu} = \sum_{k=1}^h t_k \hat{p}_k$, as to be expected. Since $t_1 \leq \hat{\mu} \leq t_h$, and since the left hand side of (2.5) is monotone in the variable a, we can approximate $\hat{\mu}$, numerically, with accuracy of at least $(t_h - t_1)2^{-M}$ in M evaluations and comparisons of the left hand side of (2.5). This, in turn, gives approximately the same accuracy for the $\hat{p}_k$'s after substituting in (2.4). Thus very

little computational effort is required in order to compute $\hat{F}$ to any reasonable accuracy. To verify that the $\hat{p}_k$'s of (2.4) sum to one, we multiply (2.4) by $nt_k + m\hat{\mu}$ and sum over k. This gives $n + m = n + m\sum_{k=1}^{h} \hat{p}_k$, so that indeed $\sum_{k=1}^{h} \hat{p}_k = 1$. The underlying assumption in the theorem, and in the discussion so far, is that $m \geq 1$ and $n \geq 1$. Nevertheless, the theorem is correct if either m or n is zero. If $n = 0$, so that the Y-sample does not exist, then (2.5) implies that $\hat{\mu}$ is the mean of the X-sample and (2.4) implies that F is the empirical distribution function based on the X-sample. If $m = 0$ so that the X-sample does not exist, then (2.5) is vacuous and for $\sum_{k=1}^{h} \hat{p}_k = 1$ we get from (2.4) that $\hat{\mu}$ is the harmonic mean of the Y-sample, and that $\hat{p}_k \propto \eta_k/t_k$. This is the estimator proposed by Cox (1969) for the problem of estimating F on the basis of a length-biased sample.

The NPMLE of the length-biased cdf G is, of course,

$$\hat{G}(x) = \sum_{t_i \leq x} \hat{g}_i, \tag{2.6}$$

where

$$\hat{g}_i = \frac{t_i \hat{p}_i}{\hat{\mu}} = \frac{(\xi_i + \eta_i)t_i}{nt_i + m\hat{\mu}}. \tag{2.7}$$

3. Asymptotic properties. Suppose F is absolutely continuous with respect to Lebesgue measure, and let f and $\mu < \infty$ be its density and mean, respectively. We write $N = m + n$, $\lambda = m/N$ and we state the asymptotic properties of $\hat{F}$ (also of $\hat{G}$ and $\hat{\mu}$) under the assumption that $N \to \infty$ and $\lambda > 0$ remains fixed. Define

$$K(x) = \int_0^x \frac{y}{\lambda\mu + (1-\lambda)y} f(y)\, dy,$$
$$K = K(\infty). \tag{3.1}$$

Convergence in distribution, or weak convergence, for stochastic processes is denoted by $\to_d$.

THEOREM 3.1.

$$\hat{\mu} - \mu \to 0 \text{ a.e.} \tag{3.2}$$

$$\sqrt{N}\left(\frac{\hat{\mu}}{\mu} - 1\right) \to_d \mathcal{N}\left(0, \quad \frac{1-K}{K\lambda(1-\lambda)}\right). \tag{3.3}$$

Let μ_i denote the ith moment of F, and suppose that μ_{-1} and μ_2 exist. Then we have

$$\lim_{\lambda \to \lambda_0} \frac{\mu^2(1-K)}{K\lambda(1-\lambda)} = \begin{cases} \mu^2(\mu\mu_{-1} - 1) & \text{if } \lambda_0 = 0, \\ \mu_2 - \mu^2 = \mathrm{Var}(X) & \text{if } \lambda_0 = 1. \end{cases} \tag{3.4}$$

Note that (3.4) with $\lambda_0 = 0$ combined with (3.3) is the result stated in Cox (1969, formula (5.4)) for the corresponding problem on the basis of a sample from G alone, while (3.4) with $\lambda_0 = 1$ combined with (3.3) is the standard result of the central limit theorem for the corresponding problem on the basis of a single sample from F.

THEOREM 3.2. *Suppose f, the density of F, is bounded. Then we have* (i)

$$\sqrt{N}(\hat{F} - F) \to_d V_F, \tag{3.5}$$

where V_F is a pinned Gaussian process with the covariance function

$$\mathrm{Cov}\{V_F(s), V_F(t)\} = \frac{1}{\lambda}[F(s)\{1 - F(t)\} - (1-\lambda)K(s)\{1 - K(t)/K\}], \quad 0 \leq s \leq t. \tag{3.6}$$

(ii) $\sqrt{N}(\hat{G} - G) \to_d V_G$,

where V_G is a pinned Gaussian process with the covariance function

$$\text{(3.7)}\quad \operatorname{Cov}\{V_G(s), V_G(t)\} = \frac{1}{1-\lambda}[G(s)\{1-G(t)\} - \lambda K(s)\{1-K(t)/K\}] \quad 0 \le s \le t.$$

(iii) $\sqrt{N}(\hat{F}-F, \hat{G}-G) \to_d (V_F, V_G)$
and

$$\text{(3.8)}\qquad \operatorname{Cov}\{V_F(s), V_G(t)\} = \min\{K(s), K(t)\} - K(s)K(t)/K.$$

We note that (3.6) through (3.8) can be combined to give

$$\operatorname{Cov}\{\lambda V_F(s) + (1-\lambda)V_G(s), \lambda V_F(t) + (1-\lambda)V_G(t)\}$$
$$= \lambda F(s)\{1-F(t)\} + (1-\lambda)G(s)\{1-G(t)\}, \quad 0 \le s \le t,$$

as one would expect because of equation (2.4), which equates the empirical distribution function of the t_i's to $\lambda\, d\hat{F}(u) + (1-\lambda)\, d\hat{G}(u)$.

4. Outlines of proofs. Since F is absolutely continuous we have $h = N$ and $\xi_k + \eta_k = 1$ with probability one, so that we can replace (2.4) and (2.5) with

$$\text{(4.1)}\qquad \hat{p}_k = \frac{1}{N}\frac{\hat{\mu}}{(1-\lambda)t_k + \lambda\hat{\mu}},$$

where $\hat{\mu}$ is the solution of

$$\text{(4.2)}\qquad Q_N(a) \equiv \frac{1}{N}\sum_{k=1}^{N}\frac{t_k}{(1-\lambda)t_k + \lambda a} = 1.$$

Now (3.2) follows from (4.2) and the strong law of large numbers, upon observing that

$$Q_N(a) \to_{\text{a.e.}} Q(a) \equiv \frac{1}{\mu}\int_0^\infty \left\{\frac{(1-\lambda)x + \lambda\mu}{(1-\lambda)x + \lambda a}\right\} x f(x)\, dx \gtreqless 1 \quad \text{iff} \quad a \gtreqless \mu.$$

To prove (3.3) we note that

$$Q\left(\mu - \frac{1}{N}\right) + \varepsilon_N\left(\mu - \frac{1}{N}\right) \ge Q_N(\mu) \ge Q\left(\mu + \frac{1}{N}\right) + \varepsilon_N\left(\mu + \frac{1}{N}\right)$$

where $\varepsilon_N(a) = Q_N(a) - Q(a)$, $a > 0$. We then apply the Taylor expansion

$$Q_N^{-1}(Q_N(a) + \Delta) = a + \frac{\Delta}{Q_N'(a)} + O(\Delta^2),$$

to get that

$$\hat{\mu}_N - \mu = \frac{\mu}{\lambda K}\varepsilon_N(\mu) + O_p(N^{-1}).$$

The result now follows by approximating $\varepsilon_N(\mu)$ using the central limit theorem.

The proof of (3.4) is straightforward using the L'Hôpital rule.

The difficulty in studying $\hat{F}$, by comparison to the empirical distribution function in the case of random sampling from a distribution function, is that $\hat{F}(s)$ is not a sum of independent identically distributed random variables. This is, of course, because of the dependence on $\hat{\mu}$, which is itself a nonlinear function of the observations. To overcome this difficulty we introduce a pseudo-estimate of F, namely

$$\tilde{F}(s) = \frac{1}{N}\sum_{t_i \le s}\frac{\mu}{(1-\lambda)t_i + \lambda\mu}$$

which is $\hat{F}$ with μ in place of $\hat{\mu}$. Let

$$R_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{t_i - \mu}{(1-\lambda)t_i + \lambda\mu},$$

then we have

$$\sqrt{N}\{\hat{F}(s) - F(s)\} = \sqrt{N}\{\tilde{F}(s) - F(s)\} + (1-\lambda)\frac{K(s)}{K} R_N + O_p(N^{-1/2})$$

$$\equiv W(s) + O_p(N^{-1/2}).$$

A standard application of the Cramér-Wold device then shows that, for $n \geq 1$ and $0 < s_1 < \cdots < s_n < \infty$ arbitrary, $\sqrt{N}(\hat{F}(s_1) - F(s_1), \cdots, \hat{F}(s_n) - F(s_n))$ has as its limiting distribution the n dimensional $\mathcal{N}(0, C)$ with covariance matrix C given by

$$C_{ij} = \mathrm{Cov}\{W(s_i), W(s_j)\}.$$

The covariance function of the process $W(s)$ is

$$\mathrm{Cov}\{W(s), W(t)\} = N\mathrm{Cov}\{\tilde{F}(s), \tilde{F}(t)\} + \sqrt{N}(1-\lambda)\frac{K(t)}{K}\mathrm{Cov}\{\tilde{F}(s), R_N\}$$

$$+ \sqrt{N}(1-\lambda)\frac{K(s)}{K}\mathrm{Cov}\{\tilde{F}(t), R_N\} + (1-\lambda)^2\frac{K(s)K(t)}{K^2}\mathrm{Var}(R_n),$$

which equals the right hand side of (3.6). To complete the proof of Theorem 3.2(i) it remains to show that $\sqrt{N}(\hat{F} - F)$ is tight. This follows by showing that $\sqrt{N}(\hat{F} - \tilde{F})$ and $\sqrt{N}(\tilde{F} - F)$ are both tight. The actual tightness proof is somewhat long, however, and it relies on various tightness criteria that can be found in Billingsley (1968). This completes the outline of the proof of (3.5) and (3.6). The other parts of Theorem 3.2 are proved similarly.

Acknowledgement. I would like to thank Colin Mallows for helpful comments.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

COLEMAN, R. (1979). *An Introduction to Mathematical Stereology*. Memoirs No. 3, Dept. of Theoretical Statistics, Institute of Mathematics, Univ. of Aarhus, London.

COX, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, Johnson, N. L. and Smith, H. Jr. (eds.). Wiley-Interscience, New York 506–527.

FELLER, W. (1968). *Introduction to Probability Theory and its Applications*, Vol. 1, 3d ed. Wiley, New York.

FELLER, W. (1971). *Introduction to Probability Theory and its Applications*, Vol. 2, 2d ed. Wiley, New York.

PATIL, G. P. and RAO, C. R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, P. R. Krishnaiah, ed., North-Holland, Amsterdam. 383–405.

PATIL, G. P. and RAO, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34** 179–189.

VARDI, Y. (1981). Nonparametric estimation in the presence of length bias. *Bell Labs Tech. Memo.*

VARDI, Y. (1982). Nonparametric estimation in renewal processes. *Ann. Statist.* To appear.

BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974