Statistica Sinica **7**(1997), 1121-1133

ASYMPTOTICALLY EFFICIENT NONPARAMETRIC ESTIMATION WITH ADDITIONAL DICHOTOMOUS OBSERVATIONS

Yuly Koshevnik and William R. Schucany

Southern Methodist University

Abstract: Nonparametric estimation of a cumulative distribution function, F, is accomplished from data containing independent observations of two types. The first type of observation is simply a recorded value of a random variable X distributed according to F. The second type is incomplete (or censored) and contains partial information about X, namely only the indicator of the event $[X \leq d]$ is available. The value d belongs to a grid $\{d_1 \leq \cdots \leq d_r\}$, so the second type can be thought of as a stratified sample of dichotomous observations, each of them being represented as a pair containing a nonrandom d_j and realizations of the indicator $Y_j = \mathbf{I}[X \leq d_j]$.

Asymptotically efficient estimates are derived for a cumulative distribution function (CDF) and therefore, for a wide class of functionals that can be expressed via the CDF. Their limit distribution turns out to be normal, while this asymptotic normality can be established uniformly with respect to any precompact set of CDF's. This uniformity implies asymptotic efficiency of the proposed estimates.

Key words and phrases: Combining information, contingent evaluation, estimation under constraints, geometric approach, incomplete observations.

1. Introduction

In this paper we consider nonparametric estimation of a cumulative distribution function, F. The distinguishing feature of the problem considered here is that some of the observations are complete and some others are radically censored. The proposed estimators efficiently combine all of the available data. The problem was motivated by an econometric application in which individuals' "willingness to pay" were elicited in different ways. For example, one questionnaire may simply ask how much money one would pay for something, while another approach would set a fixed level, say d, and ask whether the respondent would be willing to pay that much. For more on the topic of contingent evaluation, see Desvousges et al. (1992).

To introduce some notation for such a data collection procedure, let Z be distributed according to F. For a certain number of observations, the value of Z is recorded. These records are denoted by X. In other subsets of given sizes,

the observations are simple dichotomous indicators of the event $[Z \leq d]$. For the sake of definiteness, consider a fixed grid of r different thresholds $d_1 < \cdots < d_r$, for which independent samples of size $m_1, \ldots, m_r$ respectively, are collected. The partial information in the indicators is denoted by $Y_j = \mathbf{I}[Z \leq d_j]$, taking the values 1 and 0 according to whether the event is true or false. At last, to describe the full data set, let

$$\mathbf{Z} = \Big\{ Z_{j,i} : \quad 1 \leq i \leq m_j; \quad 0 \leq j \leq r \Big\} \tag{1}$$

be independent real–valued random variables from the same CDF F. The completely observable sample is described by

$$\mathbf{X} = \Big\{ X_i = Z_{0,i} : 1 \leq i \leq m_0 \Big\} \tag{2}$$

and binary data represented as r samples,

$$\mathbf{Y}_j = \Big\{ Y_{j,i} = \mathbf{I}[Z_{j,i} \leq d_j] : \qquad 1 \leq i \leq m_j; \Big\} \quad (1 \leq j \leq r). \tag{3}$$

The first natural question to ask is: how should one estimate $F(t)$ having nothing more than the data in (2) and (3)? That the answer is nontrivial, may be appreciated by simply noting that the estimate of $F(d_1)$ from the X's will not usually agree with that from $\mathbf{Y}_1$; and furthermore some less direct information may even be found in the $\mathbf{Y}_2$, and so forth. The more general task is to estimate a median, another quantile, or a more sophisticated functional $\Lambda(F)$. Once an efficient estimate, $\hat{F}(t)$, has been derived, the plug–in rule generally provides satisfactory estimates of these functionals. Some notation for specific probabilities will be relevant. Set

$$F(d_j) = p_j = \mathbf{Pr}[Y_{j1} = 1], \qquad\qquad 1 \leq j \leq r. \tag{4}$$

To avoid pathological problems we assume that all of the $(r+1)$ cell probabilities defined by the grid, i.e. $p_1, p_2 - p_1, p_3 - p_2, \ldots, 1 - p_r$, are strictly positive. The results here will, as well, allow one to test the null hypothesis (4) with specified p_j values against a general alternative that F and $p = (p_1, \ldots, p_r)$ are quite arbitrary. However, our emphasis is on estimation. Another related issue arises in testing whether the X's and Y's are drawn from a common CDF, but we do not pursue that here.

There are at least two distinct limiting situations that may be meaningful. The first is the reasonable notion that all subsample sizes are of the same order of magnitude. By setting $N = \sum_{j=0}^{r} m_j$, the corresponding requirement is that the fractions converge to strictly positive (possibly unknown) numbers, namely

$$\frac{m_j}{N} \to \mu_j > 0 \qquad 0 \leq j \leq r. \tag{5}$$

In what follows, the limiting fraction, $\mu_0 = \lim_{n\to\infty} \frac{m_0}{n}$, will play a slightly more important role than the remaining fractions, $\mu_1, \ldots, \mu_r$. The problem is considered asymptotically as items in an array indexed by a vector subscript, $m = (m_0, \ldots, m_r)$, under assumption (5).

An alternative asymptotic condition reflects another realistic assumption that all the fractions $\{\frac{m_0}{m_j} : 1 \le j \le r\}$ converge to 0, as $N \to \infty$. Consequently, $\frac{m_0}{n} \to 0 = \mu_0$ and such a case can be considered from two different points of view. But in both cases, there will be a singularity phenomenon, as far as the estimates and their large sample behavior are concerned. Under this assumption, the best one can achieve estimating the cell probabilities comes from the Y_j observations, while the very small sample of completely recorded X values can be ignored. Selecting the sum, $\sum_{j=0}^{r} m_j$, as an analogue of the sample size for a large sample study, the whole information about F contained in the complete observations (2) is negligible. If the sample size of the complete data alone is considered to be increasing ($m_0 \to \infty$), i.e. m_0 plays the role of n, then the random function, $\sqrt{m_0}[\hat{F} - F]$ should be analyzed. The adjustment of the initially selected estimate, $\tilde{F}$, can be produced in the same manner, as if the probabilities assigned to the given partition were completely specified by (4). These probabilities can be estimated with a much higher level of accuracy than the usual $n^{-1/2}$. So, whether the probabilities (4) are given or replaced by their exceptionally precise estimates, asymptotically, this will imply the same large sample results for the statistic $\sqrt{m_0}[\hat{F} - F]$. The phenomenon of completely specified cell probabilities was considered by Pfanzagl (1982) as estimation under quantile constraints.

Therefore, this paper's emphasis is on the situation described by (5) as the more realistic limiting condition. An example with $r = 2$ is also presented in the final section.

Techniques and tools. Some necessary elements of the geometric approach presented in Koshevnik and Levit (1976), Pfanzagl (1982), and Millar (1983) are recalled in Section 2. Also, asymptotic normality and asymptotic efficiency results for the proposed estimates are formulated in this section. Section 3 contains proofs and necessary auxiliary results. Section 4 presents several examples, including those already mentioned, some concluding remarks and further developments.

It turns out that asymptotic normality holds *uniformly* in $F \in \mathcal{U}$, where $\mathcal{U}$ is a small but *fixed* neighborhood of the unknown true distribution. Uniform weak convergence was initially studied by E. Parzen (1954). Some further extensions concerning uniformity in nonparametric CDF estimation are recalled from Koshevnik (1982, 1984). These results have been extended and cover a stratified

sample case. Uniformity results combined with the description of lower bounds of risks, which is similar to Koshevnik and Levit (1980), lead to asymptotic efficiency for the proposed estimates.

2. Main Ideas and Results

First we need to consider a geometric interpretation of the proposed estimate for $F(t)$. Its asymptotic efficiency is implied by asymptotic normality, with the asymptotically minimal limiting variance of the process $\sqrt{N}(\hat{F} - F)$, that holds uniformly in $F \in \mathcal{U}$, where $\mathcal{U}$ is an arbitrarily small neighborhood of the unknown true distribution $F \in \mathcal{F}$. Some additional requirements will be formulated about this neighborhood in the next section.

2.1. Estimated orthogonal projection

To describe an estimate of $F(t)$ from all of the data (2) and (3), set $n = m_0$ and let $\tilde{F}$ be the usual empirical CDF based only on the n complete observations, (2), $\tilde{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}[X_i \le t]$. Recall that $p = (p_j : 1 \le j \le r)$ is a vector of probabilities directly related to the data set (3). For any j, there are at least two natural estimates for p_j, namely the value of the empirical CDF at d_j, $\tilde{F}(d_j)$ and $\tilde{p}_j = Y_j/m_j = \sum_{i=1}^{m_j} Y_{j,i}/m_j$, the empirical frequency of the event $[Z_{j,i} \le d_j]$. To make proper use of the additional observations (3) to improve $\tilde{F}(t)$, consider first the case with no relation between the unknown F and p. Then the $(r+1)$ data sets (2) and (3) should be processed separately to produce an empirical estimate, $\tilde{F}(t)$, for $F(t)$ and empirical probabilities, $\tilde{p}_j$, for each p_j. The constraints in (4) require that we do more. Under (4), introduce a family of linear combinations,

$$\hat{F}_a(t) = \tilde{F}(t) - \sum_{j=1}^{r} a_j(t)(\tilde{F}(d_j) - \tilde{p}_j), \tag{6}$$

indexed by an r–dimensional vector, $a = (a_j : 1 \le j \le r)$. If a vector a is chosen to minimize the variance of (6), it corresponds to the orthogonal projection of the function, $U(X) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}[X_i \le t]$, onto a subspace spanned by the random variables,

$$\Big\{V_j = V_j(X, Y_1, \ldots, Y_r) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}[X_i \le d_j] - \frac{Y_j}{m_j}) : 1 \le j \le r\Big\},$$

with respect to the Hilbert norm defined by a joint distribution F^N of all N variables (1). It should be clarified that this procedure generally differs from the one based on the exact conditional expectation of the empirical CDF, $\tilde{F}(t)$, given the data sets, (2) and (3). Such a procedure typically is performed to

adjust the initial unbiased estimate, due to the Rao–Blackwell theorem, but in the case under consideration, the linearized version of conditionals is exploited, with respect to the space of square integrable random variables. Asymptotically, under the conditions (5), this procedure leads to the equivalent estimate, but turns out to be less computationally intensive.

Each of the terms subtracted from the initial estimate $\tilde{F}(t)$ to obtain (6) has zero expectation, due to (4); therefore for every a, the estimate $\hat{F}_a(t)$ is unbiased for $F(t)$. To define an efficient estimate, the numbers $\{a_j : 1 \le j \le r\}$, must be chosen to minimize the variance of (6). As will be shown later, the asymptotic variance, as well as any other rather general risk, will be also minimized, under limiting conditions (5). Some additional notation is useful for the theorems that follow. The coefficients $a = (a_j)$ in (6) can be described as a solution of the linear system, which is formed by "normal equations", $Ca = D$, or in more detail,

$$\sum_{l=1}^{r} C_{jl} a_l = D_j \qquad (1 \le j \le r), \tag{7}$$

involving large sample covariance for any pair of distinct j and l,

$$C_{j,l} = \frac{1}{n}[F(\min(d_j, d_l) - F(d_j)F(d_l)], \tag{8}$$

for any j,

$$C_{j,j} = (\frac{1}{n} + \frac{1}{m_j})[F(d_j)(1 - F(d_j))], \tag{9}$$

and finally, the right–hand sides are

$$D_j = D_j(t) = \frac{1}{n}[F(\min(t, d_j)) - F(t)F(d_j)]. \tag{10}$$

These are all functionals of the unknown distribution F, so their natural estimates have $\tilde{F}$ replacing F in (8), (9) and (10). The same applies to the solution of (7). So, if F is replaced by $\tilde{F}$, then the corresponding value of $a_j(\tilde{F})$ will be denoted as $\tilde{a}_j$. Certainly, the asymptotic conditions (5) should be taken into account, so that the ideal (limiting) solution, $a = C^{-1}D$, will also depend on the limiting proportions, (5). Actually, it is here that unbiasedness and variance minimality are replaced by their asymptotic analogues.

Although all the elements of the normal equation depend on the sample sizes, $m_0, m_1, \ldots, m_r$, we omit the extra subscripts from the notation. The reduced asymptotic variance will appear in such a form that already assumes the limits were properly taken.

Straightforward calculations show that after the minimizing vector, $\hat{a}$, is found, the minimal asymptotic variance attainable by the estimate $F_{\hat{a}}$ is equal

to $\lim_{n\to\infty}(n \cdot \mathbf{Var}[\hat{F}(t)]) = \lim_{n\to\infty}\{(n \cdot \mathbf{Var}[\tilde{F}(t)]) - a\prime Ca\}$, where $a = C^{-1} \cdot D$, the $r \times r$–matrix C has elements, $(C_{j,l})$ defined by (8) and (9), and D is an $r \times 1$–vector, with the components given by (10). The asymptotic variance decrease, $a'D = a'Ca$, due to the system (7), is non–negative, since C is positive–definite. This is a standard reduction of the asymptotic variance due to the additional information contained in the dichotomous observations.

2.2. Theorems

The first theorem describes a limiting behavior of the proposed projection when the true solution to (7), the theoretical values $a_j = a_j(F)$ are used in (6). It is similar to a projection described in Koshevnik and Levit (1976) for a one sample study with moment constraints imposed on an underlying distribution.

Let $\mathbf{B}$ designate an F–Brownian bridge, i.e. a Gaussian process with zero mean and covariance function, $\mathbf{E}[\mathbf{B}(t)\mathbf{B}(s)] = F(\min(t,s)) - F(t) \cdot F(s)$. Further, let $B = (B_1, \ldots, B_r)$ be a Gaussian vector, having zero mean components, independent from each other and from the process $\mathbf{B}$, with the variances, $\mathbf{Var}(B_j) = p_j \cdot (1 - p_j)$, for $1 \le j \le r$. The limiting process $\mathbf{W}$ is defined by $\mathbf{W}(t) = \frac{1}{\sqrt{\mu_0}}[\mathbf{B}(t) - \sum_{j=1}^{r} a_j \mathbf{B}(d_j)] + \sum_{j=1}^{r} \frac{a_j}{\sqrt{\mu_j}} B_j$, while the fractions, $(\mu_j : 0 \le j \le r)$, are defined as limits in (5). Denote by $\mathbf{W}_0$ the process $\mathbf{W}_0(t) = \frac{1}{\sqrt{\mu_0}}[\mathbf{B}(t) - \sum_{j=1}^{r} a_j \mathbf{B}(d_j)]$. Generally, with the cell probabilities $\{p_j : 1 \le j \le r\}$ estimated from the samples (2) and (3), the term $\sum_{j=1}^{r} \frac{a_j}{\sqrt{\mu_j}} B_j$ represents the difference in the asymptotic variance, compared to that with known cell probabilities.

Notice that if all the sample sizes, $(m_j : 1 \le j \le r)$, increase more rapidly than $n = m_0$, then all the fractions, $n/m_j \to 0$, and using n, rather than N as an analogue of the sample size, we can see that $\sqrt{n}[\hat{F}(t) - F(t)] \xrightarrow{\mathcal{D}} \mathbf{W}_0$. It is this process that describes large sample behavior of the estimated distribution in the case of almost known cell probabilities. Since each of the cell probabilities is estimated from the essentially larger sample than the one of size n, the asymptotic accuracy will be in this case the same as if the cell probabilities were known and equal to the corresponding accurate estimates.

To derive asymptotic distribution results uniformly in the infinite–dimensional parameter, F, assume that F belongs a precompact set, $\mathcal{U}$, in the space $\mathbf{C} = \mathbf{C}[-\infty, \infty]$ of all continuous functions with finite limits as $t \to \pm\infty$. For such a set $\mathcal{U}$, as shown in Koshevnik (1982, 1984), weak convergence for the empirical CDF $\tilde{F}$ holds uniformly in $F \in \mathcal{U}$. Equivalently, as $n \to \infty$, empirical processes, $\mathbf{B}^n = \sqrt{n}[\tilde{F}(\cdot) - F(\cdot)]$, converge weakly to the corresponding (F)–Brownian bridge, $\mathbf{B}$, and furthermore, for any continuous and

bounded functional Γ on the space $\mathbf{C}$, convergence is uniform in $F \in \mathcal{U}$, i.e. $\lim_{n\to\infty} \sup_{F\in\mathcal{U}} |\mathbf{E}(\Gamma(\mathbf{B}^n)) - \mathbf{E}(\Gamma(\mathbf{B}))| = 0$.

Theorem 1. *Suppose that $\hat{F}$ is defined by* (6) *with coefficients defined by* (7). *Then, under* (5), *convergence in distribution,*

$$\sqrt{N}[\hat{F}_a(r) - F(t)] \xrightarrow{\mathcal{D}} \mathbf{W}(t), \tag{11}$$

holds uniformly in $F \in \mathcal{U}$.

The formulated result is not sufficient for statistical purposes. Asymptotic efficiency of the estimate $\hat{F}(t)$ cannot be claimed, since the coefficients (a_j) depend on the unknown distribution F. The next step is a plug–in rule that replaces F by its empirical analogue, $\tilde{F}$, and a_j by $\tilde{a}_j = a_j(\tilde{F})$. This will produce the asymptotically efficient estimate.

Theorem 2. *Suppose that coefficients, $a = \{a_j : 1 \le j \le k\}$, are replaced by their empirical versions, $\tilde{a} = \{\tilde{a}_j : 1 \le j \le r\}$. Also assume that asymptotic conditions* (5) *hold. Then the procedure* (6) *yields an estimate $\hat{F}_{\tilde{a}}(t)$ for $F(t)$, such that the asymptotic normality* (11) *also holds uniformly in $F \in \mathcal{U}$.*

Lower bounds of risks and limiting behavior of the proposed estimate can be derived easily. In particular, the following result can be derived similar to many other Information Inequalities (see Begun et al. (1983) for instance). Recall that a nonnegative loss function L defined on a Euclidean r–dimensional space, $\mathbf{R}^r$, is called lower semicontinuous (semicontinuous from below) and subconvex if a set $\{v \in \mathbf{R}^r$, such that $L(v) \le u\}$ is both closed and convex, for any positive u. We also assume that L is symmetric, i.e. $L(-u) = L(u)$, for any u. Only loss functions with these features are considered here. As far as a neighborhood $\mathcal{U}$ is concerned, the following assumption is needed.

The feature known as **Extensiveness of Neighborhood** is now given some explanation. This is the most significant part of asymptotic efficiency. If $\mathcal{U}$ denotes a neighborhood that contains F, then we assume that for any finite collection, $(h_k(\cdot) \in \mathbf{L}^2(F) : 1 \le k \le K)$, of functions with zero mean, and for every positive δ, there exists a regular parametric submodel, $\mathcal{U}_\delta \subset \mathcal{U}$, of probability distributions indexed by $c \in \mathbf{R}^K$ such that $|c| < \delta$, so that their densities with respect to $F = F_0$ can be represented as $1 + \sum_{1\le k\le K} c_k \cdot h_k^\delta(\cdot) + o(|c|)$, as $c \to 0$, where h_k^δ deviates from h_k by less than δ in the $\mathbf{L}^2(F)$–norm.

Theorem 3. *Let L be an arbitrary lower semicontinuous and semiconvex symmetric loss function. Suppose that the neighborhood, $\mathcal{U}$, has the extensiveness property. Then the following inequality holds for any estimator, V^*, of $V = \{F(d_j) : 1 \le j \le r\}$:*

$$\liminf \sup_{F\in\mathcal{U}} \mathbf{E}_F[L(\sqrt{N}(V^* - V))] \ge \sup_{F\in\mathcal{U}} \mathbf{E}[L(\mathbf{W}_F)], \tag{12}$$

where the $\liminf$ *is taken as* $N \to \infty$*, under* (5).

Theorem 4. *The estimate,* $\hat{F}_{\tilde{a}}(\cdot)$*, defined in Theorem* 2 *is asymptotically efficient for the entire function,* $F(\cdot)$*. Weak convergence of random functions* $\sqrt{n}[\hat{F}_{\tilde{a}}(\cdot) - F(\cdot)]$ *to a Gaussian process* $\mathbf{W}$ *holds uniformly in* $F \in \mathcal{U}$*.*

More general versions of Theorems 3 and 4 were presented in Koshevnik (1992, 1993). Some of their consequences are exploited here to demonstrate why the plug–in rule works for the problem under consideration. In particular, $\hat{F}_{\tilde{a}}(t)$ turns out to be asymptotically efficient for $F(t)$ at any fixed point t. Theorem 4 relates to estimation problems requiring the entire CDF, $F(\cdot)$, to be estimated as an element of the space $\mathbf{C}$. It implies that a wide class of functionals, including various M–functionals, can be efficiently estimated via the plug–in device. The notion of efficiency in this case means more than in Theorem 3, because $F(t)$ is not simply estimated at a fixed point t or a given grid of t values, but as the whole curve (see Millar (1983) and Koshevnik and Levit (1980) for more details).

Having constructed an asymptotically efficient estimate, $\hat{F}_{\tilde{a}}(\cdot)$, for $F(\cdot)$, one can again use the plug–in device and estimate any affine functional, i.e.

$$\Lambda(F) = \int \ell(x) dF(x), \tag{13}$$

by substituting $\hat{F}_{\tilde{a}}$ for F. Using arguments from Koshevnik and Levit (1976), for functionals that can be represented as a composition, $L(\Lambda(F))$, of a function L with an s–dimensional argument and a vector of s affine functionals, say $\Lambda(F) = (\Lambda_1, \ldots, \Lambda_s)$, estimation can be conducted in the same manner. General results from Koshevnik (1984) imply that both uniform weak convergence and lower bounds of risks are described in terms of the same limiting Gaussian vectors and processes, so that the desirable conclusions can be extended to a wider class of functionals.

3. Proofs

In this section some proofs are given and others outlined. For more details see Koshevnik (1984). In particular, uniform weak convergence for empirical CDF's is usually implied by the requirement that a set $\mathcal{U}$ of CDF's is precompact in $\mathbf{C}$. This suggests that we consider, as a suitable replacement for a small neighborhood in the family $\mathcal{F}$, only those which are open (with respect to an initially given topology) and at the same time precompact with respect to the topology in $\mathbf{C}$. It is not surprising for infinite dimensional parameter sets that a subset $\mathcal{U}$ fails to be both open and precompact with respect to the same topology.

3.1. Uniform weak convergence: some results

Let $\tilde{F}$ denote the empirical CDF based on n i.i.d. random variables with CDF F.

Lemma 1. *If $\mathcal{U}$ is precompact in $\mathbf{C}$, then weak convergence*

$$\sqrt{n}[\tilde{F}(\cdot) - F(\cdot)] \xrightarrow{\mathcal{D}} \mathbf{B}, \tag{14}$$

as $n \to \infty$ is uniform in $F \in \mathcal{U}$.

This is proved in Koshevnik (1982). It was shown there that the result remained valid for a multivariate distribution F. Applying Lemma 1 to observable data (2) and (3), the next result can be derived. Its proof is quite standard and omitted here.

Lemma 2. *If $N \to \infty$ and* (5) *holds, then the joint distribution of a function and a finite dimensional vector*

$$\sqrt{N}\{(\tilde{F}(\cdot) - F(\cdot)),\ (\tilde{p}_1 - p_1), \ldots, (\tilde{p}_r - p_r)\} \tag{15}$$

converges weakly to the distribution of

$$\Big\{\frac{1}{\mu_0}\mathbf{B}(\cdot), (\frac{1}{\mu_1}B_1, \ldots, \frac{1}{\mu_r}B_r)\Big\}. \tag{16}$$

Weak convergence holds uniformly in $F \in \mathcal{U}$, whenever $\mathcal{U}$ is precompact in $\mathbf{C}$.

Turn now to the orthogonal projections that play such an important role here. First, they appear in the lower bounds of risks, derived in Koshevnik and Levit (1976). Furthermore, Pfanzagl (1982), as well as Begun et al. (1983), also provide necessary explanations of this procedure. The second reason to illustrate their importance is that in this case the projection is performed empirically. The theoretical procedure adjusting the initial estimate, such as $\tilde{F}(t)$ for $F(t)$, essentially depends on a true distribution F itself, while after $\tilde{F}$ is repeatedly used to estimate the orthogonal projection matrix, the same limiting behavior can be derived for the proposed estimate. This phenomenon can be referred to as adaptiveness. Having known the vector $a = a(F)$, it is possible to improve the initial estimate. Otherwise, the estimate $\tilde{a}$, adapted to the observed data, replaces the unknown a, and the improvement is also implemented.

3.2. Lower bounds of risks

To avoid some technical difficulties, only Theorem 3 is proved here, rather than its natural extension covering the more general situation of estimating the whole CDF F. Consider a finite–dimensional vector representing the values taken

by F on a given grid that includes several t values. This includes the case when the estimand is a set of values taken by F on the grid $d = (d_1 < \cdots < d_r)$. The unknown CDF F is, therefore, replaced by a vector $V = (F(d_j) : 1 \le j \le r)$. The empirical frequencies $\{n_j = n\tilde{F}(d_j) : 1 \le j \le r\}$ and $\{Y_j : 1 \le j \le r\}$ together form a sufficient statistic for V. The initially nonparametric problem therefore becomes a parametric one, with a likelihood function,

$$L(p) = \text{Const } \cdot \Big[\prod_{1 \le j \le r} (p_j - p_{j-1})^{n_j - n_{j-1}} \Big] (1 - p_r)^{n - n_r} \Big[\prod_{1 \le j \le r} p_j^{Y_j} (1 - p_j)^{m_j - Y_j} \Big].$$

Here we assume, for the sake of brevity, $p_0 = 0$ and $n_0 = 0$. The interpretation of a maximum likelihood estimate (MLE) and its one–step approximation in terms of efficiency is based on a geometric approach and can be performed similar to Millar (1983). The $2r$–dimensional parameter relevant for an alternative hypothesis, under the assumption (4), turns into an r–dimensional one, so that a one–step approximation is efficient. Limiting behavior of the one–step maximum likelihood procedure is established with the same orthogonal projection as in Theorems 1, 2 and 4.

Proof of Theorem 4. This is implied by Lemma 2 and the following extension of the Continuous Mapping Theorem established in Koshevnik (1982). For the sake of simplicity, our goals are limited to finite dimensional vectors only. Finally, we address the proof that using empirical values, $\tilde{a}_j$, still yields asymptotically efficient estimates.

The following result seems to be rather obvious, and up to the technical details, its proof simply reproduces the routine arguments, quite common in the weak convergence theory. For the sake of convenience, we only formulate this result as

Lemma 3. *Suppose that random vectors V_θ^n converge weakly to V_θ, as $n \to \infty$ uniformly in $\theta \in \Theta$. If the functions $(K_\theta : \theta \in \Theta)$ mapping each v into another finite dimensional vector, $K_\theta v$, satisfy the Lipschitz condition, with the same constant C, i.e. $\rho_2[K_\theta(u), K_\theta(v)] \le C\rho_1[u, v]$ for any pair u, v, any $\theta \in \Theta$, where ρ_1 and ρ_2 denote distance functions in $\mathbf{V}$ and $\mathbf{Z}$, respectively, then transformed random variables $K_\theta(V_\theta^n)$ converge weakly to $K_\theta(V_\theta)$ uniformly in $\theta \in \Theta$, as $n \to \infty$.*

To prove Theorem 4, suppose first that the values $a_j(F)$ from (7) are all given. These values enable one to improve the initial estimates of $F(t)$, just as Theorem 1 suggests. However, only the estimates $\tilde{a}_j$ are available. Invoking Lemma 3, we can show that weak convergence in Theorem 1 holds for any fixed given set of coefficients $(a_j(F) : 1 \le j \le r)$, uniformly in $a \in A$, whatever a precompact set A is chosen. Therefore, using a consistent estimate $\tilde{a}$ replacing

a, the adaptive estimates under consideration converge to the same limit as for the theoretical values. In fact, this is just the uniform version of the well known Slutsky's theorem.

4. Examples and Further Developments

Consider a simple example that motivated our attention to the general problem. A hypothetical survey asked $n = 100$ respondents a direct question to obtain $X = (X_i : 1 \le i \le 100)$. In two additional surveys of $m_1 = m_2 = 100$, the respondents were summarized by Y_1 and Y_2, respectively.

4.1. Calculations for the case of a two–point grid

Suppose that the grid $0 < d_1 < d_2 < \infty$ is given. The primary concern is to estimate the two values $V = (F(d_1), F(d_2))$. In this case, Theorem 3 is not actually needed for efficiency, since everything can be reduced to the case of a multinomial distribution for frequencies from (2), $n_j = n\tilde{F}(d_j)$, and their analogues Y_1 and Y_2, calculated from the dichotomous data (3).

To estimate $F(d_1)$ and $F(d_2)$ we apply the general procedure. It is simplified under the assumption $r = 2$, as the system of equations can be easily solved in this case. With these simplifications, for $t = d_1$, the asymptotically efficient estimate of $F(d_1)$ from (6) is $\hat{F}_a(d_1) = \tilde{F}(d_1) - a_1(d_1)[\tilde{F}(d_1) - \frac{Y_1}{m_1}] - a_2(d_1)[\tilde{F}(d_2) - \frac{Y_2}{m_2}]$. A similar expression works to estimate $F(d_2)$. More generally, for an arbitrary t, not necessarily one of the grid values, the same idea is appropriate.

To find the coefficients a_1 and a_2, the linear system (7) must be solved. Coefficients C_{jl} in this system are simply $C_{jj} = (\frac{1}{n} + \frac{1}{m_j})F(d_j)(1 - F(d_j))$ for $j = 1, 2$, while the covariance coefficient is $C_{12} = \frac{1}{n}F(d_1)(1 - F(d_2))$. The right side term is a vector with components $D_j(t) = \frac{1}{n}(F(\min(t, d_j)) - F(t)F(d_j))$.

The exact solution $a = a(F) = (a_1, a_2)$ of the system can be explicitly written as a vector–functional. Then replacing all the unknown values by their empirical analogues, the estimates of the two coefficients are obtained. Hence the initial estimate $\tilde{F}(d_1)$ is adjusted up to $\hat{F}_{\tilde{a}}(d_1)$. Similarly, the value $F(t)$ at any t can be estimated via the same procedure.

Numerical Illustration. Suppose that our study produced the following initial estimates:

$$\tilde{F}(d_1) = .30; \tilde{F}(d_2) = .60; \quad \text{and} \quad \frac{Y_1}{m_1} = .35; \frac{Y_2}{m_2} = .70.$$

The coefficients C_{ij} in (8) and (9) have the values

$$C_{11} = 0.0042, C_{12} = C_{21} = 0.0012, \quad \text{and} \quad C_{22} = 0.0048;$$

while the D_j values from (10) for $t = d_1$ and $t = d_2$ are respectively $D_1(d_1) = 0.0021, D_2(d_1) = 0.0009$ and $D_1(d_2) = 0.0012, D_2(d_2) = 0.0024$. Solving the system (7) twice, for d_1 and d_2, we obtain the following data adaptive values of a_1 and a_2:

$$\text{at } d_1\text{: } \tilde{a}_1 = 0.481 \text{ and } \tilde{a}_2 = 0.067; \text{ at } d_2\text{: } \tilde{a}_1 = 0.154 \text{ and } \tilde{a}_2 = 0.462.$$

Using these values, the improved estimate for $F(d_1)$ is $\hat{F}_{\tilde{a}}(d_1) = 0.30 - 0.481(0.30 - 0.35) - 0.067(0.60 - 0.70) = 0.331$ and similarly for $F(d_2)$, $\hat{F}_{\tilde{a}}(d_2) = 0.60 - 0.154(0.3 - 0.35) - 0.462(0.6 - 0.7) = 0.654$.

4.2. Some comments on singular asymptotics

We have assumed mostly that the subsample sizes m_0, m_1, and m_2 are of the same magnitude. If the two sizes corresponding to incomplete observations substantially dominate the set of completely recorded data, then the problem is asymptotically equivalent to the case with precisely known cell probabilities. Indeed, using summaries drawn from incomplete data, say Y_1 and Y_2, we can ignore the possible improvement to estimates of the cell probabilities made by the data set X, since the fractions $\frac{m_0}{m_1}$ and $\frac{m_0}{m_2}$ both tend to zero.

Therefore, the incomplete data enable one to replace the initial estimation problem with unknown cell probabilities by the one where these probabilities are estimated with the higher accuracy level. Asymptotically, an expensive part of the survey, which is represented by (2), can be simply ignored. With the probabilities $\tilde{p}_1$ and $\tilde{p}_2$ estimated from the incomplete data, the estimate $\tilde{F}(t)$ can be improved, using the same method that is described in Pfanzagl (1982) with completely known cell probabilities p_1 and p_2.

This case appears to involve a singularity, since the significant improvement can be guaranteed for estimation of $F(t)$ only between the grid points, i.e. for $t < d_1$, $d_1 < t < d_2$, and $t > d_2$. As far as efficient estimation of $F(d_1)$ and $F(d_2)$ is concerned, this can be performed from the incomplete data (3) alone and asymptotically nothing else can work better.

4.3. Further developments

Certainly, the model considered here relates to the area of censored data. Connections between this case and more common censoring are investigated in Koshevnik (1993). In a model under random right censoring, a pair of random variables T = survival time and C = censoring time are replaced by $Y = \min(T, C)$ and an indicator of the event $(T \le C)$. The same approach leads in this case to a modification of the well known Kaplan–Meier estimate for $F(t)$.

Another extension is a biased sampling model, involving several constraints (possibly infinite dimensional) imposed on the underlying distributions of different strata. In the present case, the constraints (4) are simple. A similar procedure can be proposed in this case as well. However, the estimates for a finite–dimensional vector a of unknown coefficients require, as an intermediate step, estimation of an auxiliary infinite–dimensional parameter.

Acknowledgements

Research of the first author was supported by NSF Grant DMS–9311477. Some of the work was done while the second author was visiting the Department of Statistics, University of Oxford. We are grateful to the referees and editors for helpful comments that improved the earlier version of this paper.

References

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11**, 432-452.

Desvouges, W. H., Johnson, F. R., Dunford, R. W., Boyle, K. J., Hudson, S. P. and Wilson, K. N. (1992). *Measuring Nonuse Damages Using Contingent Evaluation: An Experimental Evaluation of Accuracy.* Research Triangle Institute Monograph 92-1.

Koshevnik, Y. (1982). *Limit Theorems of Nonparametric Statistics.* Sov. Inst. for Scientific and Technical Information. (*In Russian.*)

Koshevnik, Y. (1984). On some limit properties of nonparametric estimates of a distribution function. *Theory Probab. Appl.* **29**, 807-813.

Koshevnik, Y. (1993). Efficient estimation for restricted nonparametric problems. Technical Report SMU/DS/TR-269. Department of Statistical Science, Southern Methodist University, Dallas, Texas, 1994.

Koshevnik, Y. (1994). Nonparametric CDF estimation from stratified data. In *Computationally Intensive Statistical Methods,* **26** (Edited by: John Sall, Ann Lehman), 459-463.

Koshevnik, Y. and Levit, B. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21**, 738-753.

Koshevnik, Y. and Levit, B. (1980). Risk bounds in estimation of symmetrical distributions. *J. Sov. Math.* **21**, 65-75. *Translated in 1983.*

Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. *Lect. Notes in Math.* **976**, 75-265.

Parzen, E. (1954). On uniform convergence of families of sequences of random variables. *Univ. of California Publ. in Statist.* **2**, No. 2, 23-54.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* Lect. Notes in Statistics, **13.** Springer, New York.

MCI Telecommunications, Measures and Analysis Department, 2400 N. Glenville Road, Richardson, TX 75082, U.S.A.

E-mail: ykoshevnik@mcimail.com

Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, U.S.A.

E-mail: schucany@mail.smu.edu

(Received June 1995; accepted November 1996)