The Annals of Statistics
1985, Vol. 13, No. 1, 178–203

EMPIRICAL DISTRIBUTIONS IN SELECTION BIAS MODELS

BY Y. VARDI

AT&T Bell Laboratories

The following problem is treated: Given s not-necessarily-random samples from an unknown distribution F, and assuming that we know the sampling rule of each sample, is it possible to combine the samples in order to estimate F, and if so what is the natural way of doing it? More formally, this translates to the problem of determining whether there exists a nonparametric maximum likelihood estimate (NPMLE) of F on the basis of s samples from weighted versions of F, with known weight functions, and if it exists, how to construct it? We give a simple necessary and sufficient condition, which can be checked graphically, for the existence and uniqueness of the NPMLE and, under this condition, we describe a simple method for constructing it. The method is numerically efficient and mathematically interesting because it reduces the problem to one of solving $s - 1$ nonlinear equations with $s - 1$ unknowns, the unique solution of which is easily obtained by the iterative, Gauss-Seidel type, scheme described in the paper. Extensions for the case where the weight functions are not completely specified and for censored samples, applications, numerical examples, and statistical properties of the NPMLE, are discussed. In particular, we prove under this condition that the NPMLE is a sufficient statistic for F.

The technique has many potential applications, because it is not limited to the case where the sampled items are univariate. A FORTRAN program for the described algorithm is available from the author.

1. Introduction. Let $y_i \equiv (y_{i1}, \cdots, y_{in_i})$, $n_i \geq 1$, be a random sample from the cumulative distribution function (cdf)

$$F_i(t) = W_i(F)^{-1} \int_{-\infty}^{t} w_i(u)\, dF(u), \quad i = 1, \cdots, s, \tag{1.1}$$

where F is an unknown cdf, and

$$W_i(F) \equiv \int_{-\infty}^{\infty} w_i(u)\, dF(u), \quad i = 1, \cdots, s. \tag{1.2}$$

We assume that the weight functions, w_i, are known, nonnegative, real functions that satisfy $0 < W_i(F) < \infty$, $i = 1, \cdots, s$. The problem we consider is that of finding the nonparametric maximum likelihood estimator (NPMLE) of the cdf F on the basis of the data $y_1, \cdots, y_s$. To avoid uninteresting discussions, we assume throughout the paper that for each i, the set of all t's for which $w_i(t)$ is strictly positive has positive F measure, that $w_i(t) = 0$ for t's outside the support of F, and that the union of the supports of the F_i's is the support of F. (The support of a cdf F is the smallest closed set D for which $\int_D dF = 1$.) The case of

Received July 1983; revised May 1984.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62E99, 62M99, 62P10.

Key words and phrases. Nonparametric maximum likelihood, sample selection bias, weighted distributions.

a single sample from a weighted distribution is simple (see Section 8) and hence we also assume throughout that $s \geq 2$. The s samples are assumed independent.

When the cdf of a sample is a weighted version of F (i.e. it is of the form (1.1)), it can be thought of as if the sampled items were drawn from a population whose cdf is F, but the sampling mechanism is such that the probability of any individual to be included in the sample is proportional to $w(u)$, where u is the value (size, length, etc.) of that individual, and $w(\cdot)$ is the weight function. This dependency between the selection probability and the actual values of the sampled items (usually referred to as *selection bias*) makes the estimation of F, in a nonparametric setup, an interesting and nonstandard problem. The simpler problem of estimating F on the basis of a *single* sample from a weighted version of F has been treated by various authors for certain weight functions of interest, and Patil and Rao (1977) survey this literature. Nevertheless, nothing in the literature would answer the following simple question: Suppose three scientists, independently of each other, are recording measurements of a certain, uncontrolled, natural phenomenon whose cdf is F (to be estimated). The first scientist, because of limited experimental conditions, can observe the phenomenon only in the range 10 to 20. Outside this range the phenomenon, even if it occurred, would pass unnoticed. He reports his measurements to be 13, 15, 16, 18. The second scientist has slightly better equipment than the first one. In the range 10 to 20 he can always detect the phenomenon, but outside this range there is a 50 percent chance that an observation would pass unnoticed. He reports his measurements to be 9, 11, 17, 18. The third scientist can observe the phenomenon throughout its entire range, and his measurements are: 8, 11, 13, 16, 16, 17, 22. These sets of measurements are assumed to be statistically independent, and the question is then how to combine them in order to get a NPMLE, $\hat{F}$, of F (i.e. an equivalent of the empirical distribution function in regular sampling from F)? In the following section we show that a NPMLE of F need not always exist, or may exist but be nonunique, and we describe explicitly the types of weight functions and data sets for which a unique NPMLE does exist. We then give a simple method for deriving it, when it exists, and demonstrate how this works on the above and other examples. The method is efficient from a computational standpoint, and interesting from a mathematical standpoint, because it reduces the problem to one of solving $s - 1$ simple equations with $s - 1$ unknowns.

The material in the paper is organized as follows: In Theorems 1 and 1′ of Section 2 we give a necessary and sufficient condition for the existence of a unique NPMLE. Appendix A then replaces the algebraic condition of Theorem 1′ with an easy-to-verify graphical criterion. Theorem 2 of Section 3 reduces the problem of constructing the NPMLE (assuming it exists) to one of solving $s - 1$ equations with $s - 1$ unknowns. In Section 4 we give an algorithm for solving these $s - 1$ equations. The algorithm is summarized in (4.1). In Section 5 we look at some numerical examples (including the one described above). In Section 6 we show, under the existence and uniqueness condition, that *given the weight functions and the sample sizes*, the NPMLE is a sufficient statistic for F. Thus, anything that can be learned about F from the raw data can be learned from the NPMLE. In order to simplify the presentation we have assumed until Section 7 that the y_{ij}'s are univariate random variables; in Section 7 we point out that the methodology remains applicable when the y_{ij}'s are random elements from a general sample space (i.e., the y_{ij}'s could be random vectors, time series, etc.). In Section 8 we discuss an extension of the method to allow for censored samples, some more applications (including a multivariate one), and other related topics.

2. The likelihood function and the data.

Let $t_1 < t_2 < \cdots < t_h$ be the values occurring in the pooled sample $y_1 \cup \cdots \cup y_s$, arranged in increasing order ($h \le n_1 + \cdots + n_s$, because of possible ties), and let η_{ij} be the multiplicity of observations from y_i at t_j, $j = 1, \cdots, h$ and $i = 1, \cdots, s$. The total multiplicity of observations at t_j is denoted r_j, so that

$$r_j = \textstyle\sum_{i=1}^{s} \eta_{ij}, \quad \text{and} \quad n_i = \sum_{j=1}^{h} \eta_{ij}.$$

Throughout the paper the subscript F, say, in various probability computations indicates that the computation is done under the assumption that F is the true cdf. With this notation, the probability of our data is written as

$$\begin{aligned} P_F(\text{data}) &= P_F(y_1, \cdots, y_s) = P_F\{t_j, \eta_{1j}, \cdots, \eta_{sj};\, j = 1, \cdots, h\} \\ &= \textstyle\prod_{j=1}^{h} \left\{ \prod_{i=1}^{s} \left(\frac{w_i(t_j)dF(t_j)}{W_i(F)} \right)^{\eta_{ij}} \right\}. \end{aligned} \tag{2.1}$$

Clearly $P_F = 0$ if any t_j is a point of continuity of F, while $P_F > 0$ if $dF(t_j) > 0$, $1 \le j \le h$. Now, if F assigns a positive mass to any (Borel) set outside $\{t_1, \cdots, t_h\}$, then the cdf G, defined by

$$dG(t) = \begin{cases} dF(t)/(1 - \Delta) & t = t_1, \cdots, t_h \\ 0 & t \notin \{t_1, \cdots, t_h\} \end{cases}$$

where Δ is the total mass assigned by F to $R - \{t_1, \cdots, t_h\}$, satisfies $P_F(\text{data}) \le P_G(\text{data})$. 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. Put $p_1 \equiv dF(t_1), \cdots, p_h \equiv dF(t_h)$ and denote the likelihood function by

$$L(p) \equiv L(p \mid \text{data}) = P_p(y_1, \cdots, y_s);$$

then our problem becomes:

$$\text{maximize } L(p) = \textstyle\prod_{j=1}^{h} \left\{ \prod_{i=1}^{s} \left(\frac{w_{ij} p_j}{W_i(p)} \right)^{\eta_{ij}} \right\} \tag{2.2}$$

subject to

$$\textstyle\sum_j p_j = 1, \quad p_j > 0, \tag{2.3}$$

where we put $w_{ij} \equiv w_i(t_j)$, $i = 1, \cdots, s$, $j = 1, \cdots, h$, and $p \equiv (p_1, \cdots, p_h)$ so that

$$W_i(p) \equiv \textstyle\sum_j w_{ij} p_j, \quad i = 1, \cdots, s. \tag{2.4}$$

If the solution of (2.2) is denoted $\hat{p}$, then our estimate of F is, of course,

$$\hat{F}(t) = \textstyle\sum_{t_j \le t} \hat{p}_j$$

and it satisfies $P_G(\text{data}) \le P_{\hat{F}}(\text{data})$ for all cdf's G. We therefore call $\hat{F}$ a nonparametric maximum likelihood estimate (NPMLE) of F. (Also see Scholz, 1980.) Nevertheless, there exist weight functions and data sets for which the solution to the problem (2.2–2.3) is not unique or may not even exist. To see this, consider the following two simple examples.

EXAMPLE (*non-existence of the NPMLE*). Let here, and in the sequel, $I[\ ]$ denote the indicator function and suppose $s = 2$, $w_1(u) = I[4 \le u \le 9]$, $w_2(u) \equiv 1$, $y_1 = (6, 8)$, $y_2 = (1, 3)$; or in words: We have a sample of size two from the cdf F truncated to $[4, 9]$, with observed values 6 and 8, and a sample of size two from F, with observed values 1 and 3. The likelihood to be maximized, $L(p)$, satisfies

$$L(p) = p_1 p_2 \frac{p_3 p_4}{(p_3 + p_4)^2} < \frac{1}{16}$$

but

$$L(\tfrac{1}{2} - \varepsilon, \tfrac{1}{2} - \varepsilon, \varepsilon, \varepsilon) = (\tfrac{1}{2} - \varepsilon)^2 \tfrac{1}{4} \uparrow \tfrac{1}{16} \quad \text{as} \quad \varepsilon \downarrow 0,$$

so that indeed there does not exist an MLE. We note, however, that if the sample from F itself, y_2, had included an observed value from the truncation interval $[4, 9]$ then $L(p)$ would have possessed a maximum. For instance, suppose $y_2 = (1, 5)$ (instead of $(1, 3)$); then the likelihood to be maximized is

$$L(p) = p_1 p_2 \frac{p_3 p_4}{(p_2 + p_3 + p_4)^2}$$

and this function attains its maximum at $p = (\tfrac{1}{2}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6})$.

EXAMPLE (*nonuniqueness of the NPMLE*). Consider the case $s = 2$, $w_1(u) = I[u \le 20]$, $w_2(u) = I[u \ge 10]$, $y_1 = (6, 8)$, $y_2 = (26, 28)$. Then

$$L(p) = \frac{p_1 p_2 p_3 p_4}{(p_1 + p_2)^2 (p_3 + p_4)^2}$$

is maximized by any p of the form $p = (\alpha/2, \alpha/2, (1 - \alpha)/2, (1 - \alpha)/2)$, $0 < \alpha < 1$.

To give a necessary and sufficient condition for the maximization problem (2.2–2.3) to have a unique solution, let D_i be the set of t_j's for which $w_i(t_j)$ is positive:

$$D_i \equiv \{t_j;\ w_i(t_j) > 0,\ j = 1, 2, \cdots, h\}, \quad i = 1, \cdots, s. \tag{2.5}$$

The set of subscripts j such that $t_j \in D_i$ is denoted $\tilde{D}_i$:

$$\tilde{D}_i \equiv \{j;\ w_i(t_j) > 0,\ j = 1, 2, \cdots, h\}, \quad i = 1, \cdots, s.$$

Note that the D_i's are random sets because they depend on the data. Furthermore, if a point t_j belongs to D_i then necessarily $d\hat{F}_i(t_j) > 0$ (since $d\hat{F}(t_j) > 0$, always, and $w_i(t_j) > 0$ for $t_j \in D_i$) and so D_i should be thought of as the set of "*active points*" associated with the cdf F_i. With the above definition of D_i, we can rewrite the likelihood function $L(p)$ as

$$L(p) = \prod_{j=1}^{h} \left\{ \prod_{i=1}^{s} \left(\frac{w_{ij} p_j}{\sum_{k \in \tilde{D}_i} w_{ik} p_k} \right)^{\eta_{ij}} \right\}, \tag{2.6}$$

where all the terms w_{ik} appearing in the denominators are *strictly* positive. In order not to burden the paper with excessive notation, we let y_i denote both the ith sample $(y_{i1}, \cdots, y_{in_i})$ and the *set* $\{y_{i1}, \cdots, y_{in_i}\}$. It will always be clear from the context which definition is used.

THEOREM 1. *A necessary and sufficient condition for* (2.2–2.3) *to have a unique solution is that for each proper subset B of* $\{1, \cdots, s\}$, *the set of points $D_B \equiv \cup_{i \in B} D_i$ contains at least one observation from $\cup_{i \notin B} y_i$.*

PROOF. To prove the "necessary" part let p^* be the unique solution of (2.2–2.3), and suppose, by negation, that there exists a proper subset B of $\{1, \cdots, s\}$ such that all the points of $\cup_{i \in B} D_i$ belong to $\cup_{i \in B} y_i$. Since the reverse inclusion always holds, we get

$$D_B = \cup_{i \in B} y_i, \quad (\cup_{i \in B} y_i) \cap (\cup_{i \notin B} y_i) = \varnothing,$$

and since η_{ij} counts the multiplicity of observations from y_i at t_j we also have

$$\eta_{ij} = 0 \quad \text{for} \quad (i \in B, j \notin \tilde{D}_B) \quad \text{and for} \quad (i \notin B, j \in \tilde{D}_B).$$

Therefore, we can factor the likelihood function as follows:

$$L(p) = \left\{ \prod_{i \in B} \prod_{j \in \tilde{D}_B} \left(\frac{w_{ij} p_j}{W_i(p)} \right)^{\eta_{ij}} \right\} \left\{ \prod_{i \notin B} \prod_{j \notin \tilde{D}_B} \left(\frac{w_{ij} p_i}{W_i(p)} \right)^{\eta_{ij}} \right\}.$$

Now observe that since $D_B \cap (\cup_{i \notin B} y_i) = \varnothing$,

$$W_i(p) = \sum_{j \in \tilde{D}_B} w_{ij} p_j \quad \text{for} \quad i \in B$$

and so if we replace p^* with p^ε, defined by

$$p_j^\varepsilon = \begin{cases} \varepsilon p_j^* & \text{for} \quad j \in \tilde{D}_B \\ ((1 - \varepsilon\Delta)/(1 - \Delta)) p_j^* & \text{for} \quad j \notin \tilde{D}_B \end{cases}$$

where $1 > \Delta \equiv \sum_{j \in \tilde{D}_B} p_j^* > 0$, the first factor in the likelihood above will remain the same while the second factor will increase provided $\varepsilon > 0$ is small enough; i.e. for all sufficiently small $\varepsilon > 0$

$$\prod_{i \in B} \prod_{j \in \tilde{D}_B} \left(\frac{w_{ij} p_j^\varepsilon}{W_i(p^\varepsilon)} \right)^{\eta_{ij}} = \prod_{i \in B} \prod_{j \in \tilde{D}_B} \left(\frac{w_{ij} p_j^*}{W_i(p^*)} \right)^{\eta_{ij}}$$

$$\prod_{i \notin B} \prod_{j \notin \tilde{D}_B} \left(\frac{w_{ij} p_j^\varepsilon}{W_i(p^\varepsilon)} \right)^{\eta_{ij}} \geq \prod_{i \notin B} \prod_{j \notin \tilde{D}_B} \left(\frac{w_{ij} p_j^*}{W_i(p^*)} \right)^{\eta_{ij}}.$$

Combining these we get $L(p^{\varepsilon}) \geq L(p^*)$, contradicting the uniqueness and optimality of p^*. This proves the "necessary" part. To prove the "sufficient" part of the theorem we use two separate arguments; one to establish the existence of a solution and one to establish its uniqueness.

Argument I (existence). If the condition holds, and p^* is *any point of supremum* for the problem (2.2–2.3), then necessarily $p_j^* > 0$, $j = 1, \cdots, h$.

Argument II (uniqueness). If the solution of (2.2–2.3) is nonunique, then the condition is not satisfied.

To see why Argument I proves existence, note that it implies that for sufficiently small $\varepsilon > 0$, the supremum of $L(p)$ in the region $\sum p_j = 1$, $p_j > 0$, $j = 1, \cdots, h$, is the same as in the region $\sum p_j = 1, p_j \geq \varepsilon, j = 1, \cdots, h$, and since the latter region is compact and $L(p)$ is continuous the supremum is a maximum. The proof of Argument II is somewhat complicated and is deferred to Appendix B. We continue here with the *proof of Argument I*: Suppose, by negation, that the condition holds but some of the p_j^* are zero, and let B be the subset of $\{1, \cdots, s\}$ that registers all the samples which contain observations which were assigned zero mass by p^*. That is,

$$\begin{aligned} B &\equiv \cup_{j=1}^{h} \{\text{all } i\text{'s},\ 1 \leq i \leq s, \text{ for which } t_j \in y_i \text{ and } p_j^* = 0\} \\ &= \cup_{p_j^*=0} \{i;\ t_j \in y_i,\ 1 \leq i \leq s\}. \end{aligned} \tag{2.7}$$

Since we assumed that some of the p_j^*'s are zero, the set B is nonempty. We now state and prove the following:

PROPOSITION. *If $p_j^* = 0$ and $t_j \in y_i$, then $P_{p^*}(D_i) = 0$; consequently*

$$P_{p^*}(\cup_{i \in B} D_i) = 0. \tag{2.8}$$

PROOF OF THE PROPOSITION. If $t_j \in y_i$ then $\eta_{ij} > 0$ and so the term

$$p_j / \textstyle\sum_{k \in \tilde{D}_i} w_{ik} p_k$$

appears in the product of $L(p)$ in (2.6). Since p^* is a point of supremum of $L(p)$ and since $p_j^* = 0$, necessarily $\sum_{k \in D_i} w_{ik} p_k^* = 0$. But since $w_{ik} > 0$ for $k \in \tilde{D}_i$ we get that $\sum_{k \in \tilde{D}_i} P_k^* = P_{p^*}(D_i) = 0$, which proves the proposition.

Continuing with the proof of the theorem, suppose that B is a *proper* subset of $\{1, \cdots, s\}$. Then, since we assume that the condition holds, $\cup_{i \in B} D_i$ also includes observations from $\cup_{i \notin B} y_i$. In particular, there exist i' and j' such that

$$t_{j'} \in \cup_{i \in B} D_i, \tag{2.9a}$$

$$t_{j'} \in y_{i'}, \tag{2.9b}$$

$$i' \notin B. \tag{2.9c}$$

From (2.9a) and (2.8) we have

$$p_{j'}^* = 0.$$

But this, (2.9b), and the definition of B in (2.7), imply that $i' \in B$ which is in contradiction with (2.9c). Thus B could not be a *proper* subset. Suppose then that $B = \{1, \cdots, s\}$. It then follows from (2.8) that

$$\sum_{j=1}^{h} p_j^* = 0,$$

and this is, of course, also a contradiction because p^* is in the closure of the constraint region, and so

$$p^* \in \{p;\ \sum_{j=1}^{h} p_j = 1,\ p_j \geq 0\ j = 1, \cdots, h\}.$$

This finishes the proof of the theorem (except Argument II, above, which is proved in Appendix B). □

For any subset B of $\{1, \cdots, s\}$, let

$$D_B \equiv \cup_{i\in B} D_i \quad \text{and} \quad \tilde{D}_B \equiv \cup_{i\in B} \tilde{D}_i.$$

Since the number of observations that fell in D_B is

$$\sum_{j\in\tilde{D}_B} \sum_{i=1}^{s} \eta_{ij} = \sum_{j\in\tilde{D}_B} r_j = \sum_{j=1}^{h} r_j I[w_{ij} > 0 \text{ for some } i \in B],$$

and the number of sample observations that belong to $\cup_{i\in B}\, y_i$ is $\sum_{i\in B} n_i$, a more algebraic form of Theorem 1 is the following:

THEOREM 1′. *A necessary and sufficient condition for* (2.2–2.3) *to have a unique solution is that for each proper subset B of* $\{1, \cdots, s\}$,

$$(2.10) \qquad \sum_{j\in\tilde{D}_B} \sum_{i=1}^{s} \eta_{ij} > \sum_{i\in B} \sum_{j=1}^{h} \eta_{ij}$$

or, equivalently,

$$(2.10)' \qquad \sum_{j\in\tilde{D}_B} r_j \equiv \sum_{j=1}^{h} r_j I[w_{ij} > 0 \textit{ for some } i \in B] > \sum_{i\in B} n_i.$$

The reader is referred to Appendix A for an easy-to-verify, graphical criterion which is equivalent to condition (2.10).

The following corollary will be needed in the next section.

COROLLARY. *If* (2.10), *or equivalently the condition described in Theorem* 1, *holds for every proper subset B of* $\{1, \cdots, s\}$, *then for each i, $i = 1, \cdots, s$,*

$$(2.11) \qquad (1/n_i) \sum_{j=1}^{h} r_j I[w_{ij} > 0] > 1,$$

and

$$(2.12) \qquad (1/n_i) \sum_{j=1}^{h} r_j I[w_{ij} > 0,\ w_{kj} = 0 \textit{ for all } 1 \leq k \leq s,\ k \neq i] < 1.$$

PROOF. (2.11) follows from (2.10) by choosing $B = \{i\}$, and (2.12) follows from the fact that if equality holds in (2.12) then the sample y_i is entirely contained in $\cap_{k\neq i} D_k^c = (\cup_{k\neq i} D_k)^c$. Therefore it is impossible for $\cup_{k\neq i} D_k$ to include an observation from y_i, which is in contradiction with the assumption that the condition is satisfied. Since the left side of (2.12) could not possibly exceed 1, the result follows.

For the remainder of this paper we assume that the data and the weight functions are such that (2.10) *holds for every proper subset B of* $\{1, \cdots, s\}$, *so that a unique NPMLE exists.* We note, for instance, that if $w_i(t) > 0$ for all t in the support of F, $i = 1, \cdots, s$, as could be the case in some interesting applications, then the above assumption puts no restrictions on the data because then $D_i = \bigcup_{k=1}^{s} D_k$, $i = 1, \cdots, s$. Also in large samples the assumption above will typically be satisfied so long as the w_i's satisfy a certain overlapping requirement. It can be shown from the Lemma of Section 8(iv) that this requirement is that there does not exist a proper subset B of $\{1, \cdots, s\}$ such that

$$(\textstyle\bigcup_{i\in B} \text{Support of } F_i) \cap (\bigcup_{i\notin B} \text{Support of } F_i) = \varnothing.$$

Thus if, for instance, $w_1(t) > 0$ for all t's in the support of F, then as the sample sizes go to infinity, with probability one a unique NPMLE exists.

Finally we remark that if indeed we are in a situation where the NPMLE does not exist or it exists but it is not unique then, as demonstrated by the two examples preceding Theorem 1, we should think hard about what are the quantities that we are legitimately allowed to estimate with some degree of confidence. In this connection the reader is invited to interpret the estimates $p = (½ - \varepsilon, ½ - \varepsilon, \varepsilon, \varepsilon)$ and $p = (\alpha/2, \alpha/2, (1 - \alpha)/2, (1 - \alpha)/2)$ in these examples.

3. The NPMLE. First we note that $L(p)$ is homogeneous of degree zero and so, in order to simplify the mathematics, it is advantageous to replace the maximization problem (2.2–2.3) with a slightly modified problem, in which one of the samples, say the sth sample, plays a pivotal role. For $q = (q_1, \cdots, q_h)$, let

$$L^*(q) \equiv \textstyle\prod_{j=1}^{h} \left\{(w_{sj}q_j)^{\eta_{sj}} \prod_{i=1}^{s-1} \left(\dfrac{w_{ij}q_j}{\sum_{k=1}^{h} w_{ik}q_k}\right)^{\eta_{ij}}\right\}, \tag{3.1}$$

and consider the problem:

$$\text{maximize } L^*(q) \tag{3.2}$$

subject to

$$\textstyle\sum_{j=1}^{h} w_{sj}q_j = 1, \quad q_j > 0, \quad j = 1, \cdots, h. \tag{3.3}$$

LEMMA 1. *If $\hat p = (\hat p_1, \cdots, \hat p_h)$ is a solution of* (2.2–2.3) *then $\hat q = \hat p/\sum_{k=1}^{h} w_{sk}\hat p_k$ is a solution of* (3.2–3.3). *Conversely, if $\hat q = (\hat q_1, \cdots, \hat q_h)$ is a solution of* (3.2–3.3) *then $\hat p = \hat q/\sum_{j=1}^{h} \hat q_j$ is a solution of* (2.2–2.3).

The lemma is easily proved using the fact that $L(p)$ is homogeneous of degree zero, and the fact that in the region (3.3) L and L^* coincide. We omit the details.

We now proceed to solve (3.2–3.3) which, because of the lemma above, is equivalent to solving (2.2–2.3). *For the remainder of this paper, whenever the indices i and j appear without a specified range it is understood that they run from* 1 *to* $s - 1$ *and from* 1 *to* h, *respectively.*

Recall that

$$r_j = \textstyle\sum_{i=1}^{s} \eta_{ij}, \quad j = 1, \cdots, h, \tag{3.4}$$

and define for $A_1 > 0, \cdots, A_{s-1} > 0$

$$H_i(A_1, \cdots, A_{s-1}) = A_i^{-1} \textstyle\sum_j \left(\dfrac{r_j w_{ij}}{n_s w_{sj} + \sum_{k=1}^{s-1} n_k w_{kj} A_k^{-1}}\right), \quad i = 1, \cdots, s-1. \tag{3.5}$$

So that $H_1, \cdots, H_{s-1}$ are $s - 1$ functions from the positive orthant of R^{s-1} into the positive reals.

THEOREM 2. *Assume that* (2.10) *holds. Then the unique solution of* (3.2–3.3) *is*

$$\hat{q}_j = \frac{r_j}{n_s w_{sj} + \sum_{i=1}^{s-1} n_i w_{ij} \hat{V}_i^{-1}}, \quad j = 1, \cdots, h, \tag{3.6}$$

where $(\hat{V}_1, \cdots, \hat{V}_{s-1})$ *is the unique solution of the simultaneous equations*

$$H_i(A_1, \cdots, A_{s-1}) = 1, \quad i = 1, \cdots, s-1, \tag{3.7}$$

in the range $A_1 > 0, \cdots, A_{s-1} > 0$. *The unique solution* $\hat{p} \equiv (\hat{p}_1, \cdots, \hat{p}_h)$, *of* (2.2–2.3) *is then derived from* $\hat{q} \equiv (\hat{q}_1, \cdots, \hat{q}_h)$ *by setting*

$$\hat{p}_j \equiv \lambda \hat{q}_j, \quad j = 1, \cdots, h \tag{3.8}$$

where

$$\lambda \equiv 1/\textstyle\sum_{j=1}^{h} q_j. \tag{3.9}$$

Furthermore, we have

$$\lambda = \textstyle\sum_{j=1}^{h} w_{sj} \hat{p} = W_s(\hat{p}), \tag{3.10a}$$

$$W_i(\hat{p}) \equiv \textstyle\sum_{j=1}^{h} w_{ij} \hat{p}_j = \lambda \hat{V}_i, \quad i = 1, \cdots, s-1. \tag{3.10b}$$

PROOF. In Appendix B we show that (3.2–3.3) is equivalent to

$$\text{minimize}\left(\textstyle\prod_j q_j^{-r_j}\right)\left(\textstyle\prod_i u_i^{n_i}\right) \tag{3.11}$$

subject to

$$\begin{aligned} \textstyle\sum_j w_{sj} q_j &\le 1, \\ u_i^{-1} \textstyle\sum_j w_{ij} q_j &\le 1, \quad i = 1, \cdots, s-1, \\ u_i &> 0, \quad i = 1, \cdots, s-1, \\ q_j &> 0, \quad j = 1, \cdots, h. \end{aligned} \tag{3.12}$$

(Note that (3.11–3.12) is a standard form of a Geometric Programming problem—e.g. Zangwill, 1969—which may suggest an alternative solution to the one we describe here, in situations where software for such problems is available.) Upon substituting $q_j = e^{-\alpha_j}$ and $u_i = e^{\beta_i}$ (3.11–3.12) becomes

$$\text{minimize } \exp\{\textstyle\sum_j r_j\alpha_j + \sum_i n_i\beta_i\} \tag{3.13}$$

subject to

$$\begin{aligned} &\textstyle\sum_j w_{sj}e^{-\alpha_j} \leq 1, \\ &\textstyle\sum_j w_{ij}e^{-(\alpha_j+\beta_i)} \leq 1, \quad i = 1, \cdots, s-1, \end{aligned} \tag{3.14}$$

which is a problem of minimizing a convex function over a convex region. Now, because of the convexity, the Kuhn-Tucker (KT) conditions are necessary and sufficient for optimality in (3.13–3.14) (e.g. Zangwill, 1969, Theorem 2.19(e)) and by mapping these conditions back to the original variables we get that a necessary and sufficient condition for $\hat{q}$ to solve (3.2–3.3) is that

$$q_j(\partial l/\partial q_j)\,|_{\hat{q}} = r_j - \hat{q}_j(\textstyle\sum_i n_i w_{ij} W_i^{-1}(\hat{q}) + n_s w_{sj}) = 0, \quad j = 1, \cdots, h. \tag{3.15}$$

Here

$$l(q) \equiv \log L^*(q) - n_s(\textstyle\sum_j w_{sj}q_j - 1)$$

is the Lagrangian of $\log L^*(q)$, and n_s is the Lagrange multiplier, so that (3.15) are the KT conditions for (3.2–3.3). In particular, since (2.10) implies that the original problem has a unique solution, it follows that (3.15) has a unique solution. Continuing with the proof, we note (see the argument in (3.16–3.17) below) that if $(\hat{V}_1, \cdots, \hat{V}_{s-1})$ solves (3.7) and $\hat{q}$ is defined by (3.6), then $\hat{q}_j > 0$ and $\sum w_{ij}\hat{q}_j = \hat{V}_i$, and so $\hat{q}$ is a solution of (3.15); i.e., it is a point of maximum. This, however, does not establish yet that (3.7) has a unique solution. To prove this we first note that if $\hat{q}$ is the solution of (3.15) then $A_i = \sum_j w_{ij}\hat{q}_j$, $i = 1, \cdots, s-1$, is a solution of (3.7), and so it has at least one solution. To see that it has *exactly* one solution assume, by negation, that $A^* \equiv (A_1^*, \cdots, A_{s-1}^*)$ and $A^\# \equiv (A_1^\#, \cdots A_{s-1}^\#)$ are two *different* solutions of (3.7); then, from (3.7) and the definition of the H_i's,

$$q_j^* \equiv r_j/(n_s w_{sj} + \textstyle\sum_i n_i w_{ij} A_i^{*-1}), \quad j = 1, \cdots, h,$$

and

$$q_j^\# \equiv r_j/(n_s w_{sj} + \textstyle\sum_i n_i w_{ij} A_i^{\#-1}), \quad j = 1, \cdots, h,$$

satisfy $A_i^* = \sum_j w_{ij}q_j^*$ and $A_i^\# = \sum_j w_{ij}q_j^\#$, and since $A^* \neq A^\#$, we have $q^* \neq q^\#$. This, however, is a contradiction to the fact that (3.15) has a unique solution, because both q^* and $q^\#$ are solutions of (3.15), and so (3.6–3.7) is proved. The proof of (3.8) follows from Lemma 1, and the proof of (3.10) follows from the substitution (3.8–3.9) and the fact that if $(\hat{V}_1, \cdots, \hat{V}_{s-1})$ is a solution of (3.7) and $\hat{q}$ is defined by (3.6), then by multiplying (3.6) by w_{ij} and summing it over j we get, using (3.7),

$$\hat{V}_i = \textstyle\sum_j w_{ij}\hat{q}_j, \quad i = 1, \cdots, s-1, \tag{3.16}$$

and by multiplying (3.6) by the denominator of its right side, and summing it over j we get, using (3.16),

$$n_s \sum_{j=1}^{h} w_{sj}\hat{q}_j + \sum_{i=1}^{s-1} n_i = \sum_{j=1}^{h} r_j.$$

Since $\sum r_j = n_1 + \cdots + n_s$, indeed

$$\sum_{j=1}^{h} w_{sj}\hat{q}_j = 1. \tag{3.17}$$

(3.10a) now follows from (3.17), and (3.10b) follows from (3.16) and (3.17). This ends the proof of Theorem 2. □

To solve (3.7) one can use any method designed to solve a system of nonlinear equations, or, alternatively, any optimization method that would find the (unique) minimum of $\sum_i (H_i(A_1, \cdots, A_{s-1}) - 1)^2$ in the range $A_1 > 0, \cdots, A_{s-1} > 0$. Nevertheless, since the theorem suggests that all the work in finding $\hat{p}$ lies in solving these equations, it would be more efficient, computationally, to use a method that is tailored to the nice mathematical properties of the H_i's, rather than a general purpose nonlinear equations solver. To derive such a method we first note that for each i, H_i is monotone in each of its $s - 1$ variables. It is decreasing in A_i for fixed A_k $(k \neq i)$ and increasing in A_k $(k \neq i)$ when we hold all the variables but the kth fixed. Furthermore, from (2.11) and (2.12) we have

$$\begin{aligned} &H_i(A_1, \cdots, A_{i-1}, 0, A_{i+1}, \cdots, A_{s-1}) \\ &\qquad = (1/n_i) \sum_{j=1}^{h} r_j I[w_{ij} > 0] > 1, \\ &H_i(A_1, \cdots, A_{i-1}, \infty, A_{i+1}, \cdots, A_{s-1}) \\ &\qquad = (1/n_i) \sum_{j=1}^{h} r_j I[w_{ij} > 0,\ w_{kj} = 0 \text{ for all } 1 \leq k \leq s,\ k \neq i] < 1 \end{aligned} \tag{3.18}$$

and so, because of the monotonicity of H_i, for fixed but arbitrary a_k $(k \neq i)$, the equation

$$H_i(a_1, \cdots, a_{i-1}, A_i, a_{i+1}, \cdots, a_{s-1}) = 1 \tag{3.19}$$

has a unique solution in A_i, say a_i'.

Furthermore, in the special case where $w_{sj} > 0$ for $j = 1, \cdots, h$ then, regardless of the values of $a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_{s-1}$, we must have

$$0 < a_i' < A_i^{\max} \leq \bar{A}_i \equiv (\textstyle\sum_j r_j)\max\{w_{ih}/w_{s1}, \cdots, w_{ih}/w_{sh}\} \tag{3.20}$$

where $A_i^{\max}$ is the solution of the equation

$$H_i(\infty, \cdots, \infty, A_i, \infty, \cdots, \infty) = \sum_j \frac{r_j w_{ij}}{n_s w_{sj} A_i + n_i w_{ij}} = 1. \tag{3.21}$$

The relations (3.20) and (3.21) follow from the monotonicity of H_i and are easily understood from Figure 1. Note that the rightmost side of (3.20) is independent of the values of a_k, $k \neq i$, and so the same interval, $[0, \bar{A}_i]$, can be used in all the iterations (4.1), below. If, however, $w_{sj} = 0$ for some j's, then the interval within which we search for the solution a_i' should be redetermined in each iteration. This is so because when $w_{sj} = 0$ for some j's, a_i' may not be uniformly bounded in a_k, $k \neq i$. In this connection we call the reader's attention to the asymptotes

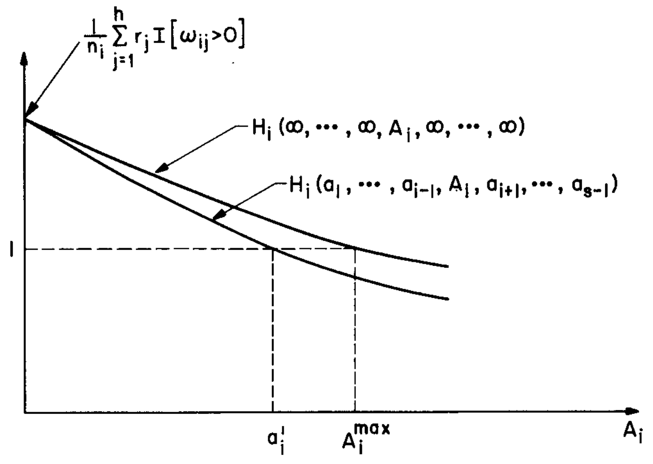
$\frac{1}{n_i}\sum_{j=1}^{h} r_j I[\omega_{ij}>0]$
$H_i(\infty,\cdots,\infty,A_i,\infty,\cdots,\infty)$
$H_i(a_1,\cdots,a_{i-1},A_i,a_{i+1},\cdots,a_{s-1})$
1
a_i^1
$A_i^{\max}$
A_i

FIG. 1. *A pictorial proof of* (3.20).

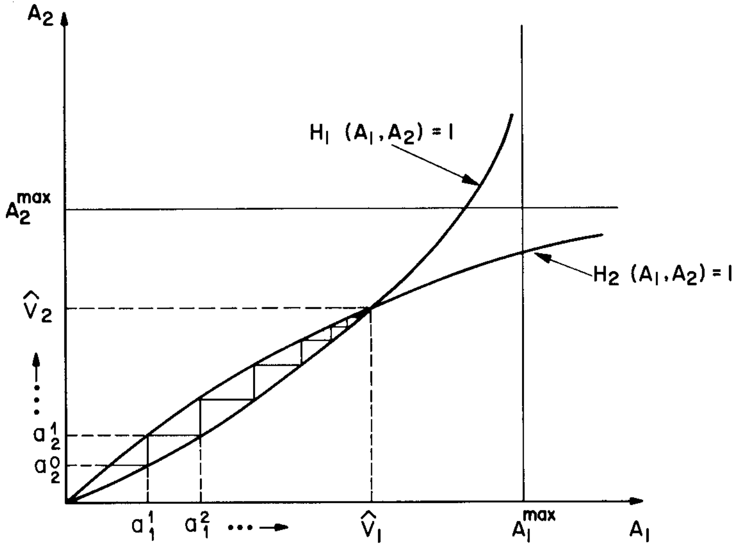

FIG. 2. *An illustration of the iterative scheme* (4.1) *for* $s = 3$. (*Note that if* $a_2^0 > \hat{V}_2$ *then the convergence would have been from above; that is* $a_2^k \downarrow \hat{V}_2$ *and* $a_1^k \downarrow \hat{V}_1$. *Also note that we assumed here that* $w_{sj} > 0$, $j = 1, \cdots, h$, *so that the graph of* $H_i = 1$ *has an asymptote at* $A_i^{\max}$, $i = 1, 2$.)

of the graphs $H_i = 1$ at $A_i^{\max}$, $i = 1, 2$, in Figure 2. Simple algebra would show that these asymptotes may not exist when $w_{sj} = 0$ for some j's.

The above analysis suggests the following:

4. Iterative method for solving (3.7). Choose (arbitrarily) positive real numbers $a_2^0, \cdots, a_{s-1}^0$ and solve

$$H_1(A_1, a_2^0, \cdots, a_{s-1}^0) = 1$$

for A_1. Denote the solution by a_1^1 and proceed to solve

$$H_2(a_1^1, A_2, a_3^0, \cdots, a_{s-1}^0) = 1$$

for A_2. Denote the solution by a_2^1 and continue in the same manner. In general,

$$\text{at the } (\alpha(s-1)+\beta)\text{th step,} \quad \alpha = 0, 1, \cdots \text{ and } \beta = 1, \cdots, s-1, \quad \text{solve}$$

$$H_\beta(a_1^{\alpha+1}, \cdots, a_{\beta-1}^{\alpha+1}, A_\beta, a_{\beta+1}^{\alpha}, \cdots, a_{s-1}^{\alpha}) = 1 \tag{4.1}$$

$$\text{for} \quad A_\beta, \quad \text{and denote the solution by} \quad a_\beta^{\alpha+1}.$$

This sequence of iterations is stopped when $a^{\alpha+1} \equiv (a_1^{\alpha+1}, \cdots, a_{s-1}^{\alpha+1})$ is sufficiently close to a^α, so that the desired accuracy has been achieved. We note that, because of the monotonicity of H_β, in each iteration equation (4.1) can be solved numerically by bisecting an interval $(0, \bar{A}_\beta)$ where $\bar{A}_\beta$ is chosen big enough to satisfy

$$H_\beta(a_1^{\alpha+1}, \cdots, a_{\beta-1}^{\alpha+1}, \bar{A}_\beta, a_{\beta+1}^{\alpha}, \cdots, a_{s-1}^{\alpha}) < 1. \tag{4.2}$$

(Recall that if $w_{sj} > 0$, $j = 1, \cdots, h$, then $\bar{A}_\beta$ can be defined as in (3.20).) Thus, from a computational standpoint the algorithm is very simple to implement. A FORTRAN program for finding the NPMLE as described in Theorem 2, based on the iterative scheme (4.1), is available from the author.

In discussing the structure of the algorithm, Mallows (1985) shows that (4.1) is an *alternating-maximization* scheme and so the likelihood function increases in each iteration. This procedure is also an example of a Gauss-Seidel method (e.g. Ortega and Rheinboldt, 1970), and for the case $s = 3$ it is depicted in Figure 2.

Note that the monotonicity of the H_i's and the fact that (3.7) has a unique solution, guarantee that the graphs of $H_i = 1$, $i = 1, 2$, are monotone increasing and intersect only once as depicted in the figure.

5. Numerical examples.

EXAMPLE 1 ($s = 3$, $w_s(u) \equiv 1$). This is the example of the introduction for which we have $w_1(u) = I[10 \le u \le 20]$, $w_2(u) = (1 + I[10 \le u \le 20])/2$, and $w_3(u) \equiv 1$. The data and the resulting NPMLE are summarized in Table 1.

The algorithm (4.1) was initialized at $a_2^0 = 15.0$ and converged through the following eight pairs of iterations: (1.10092, 15.00000) (1.10092, .98193) (.72244, .98193) (.72244, .85865) (.68620, .85865) (.68620, .84281) (.69108, .84281) (.68108, .84050) (.68033, .84050) (.68033, .84016) (.68021, .84016) (.68021, .84011) (.68020, .84011) (.68020, .84010) (.68019, .84010) (.68019, .84010) = $(\hat{V}_1, \hat{V}_2)$.

EXAMPLE 2 ($s = 4$, $w_s(u) \equiv 1$). In this example, we assume that in addition to the three scientists of Example 1, there is another scientist whose observations are 15, 19, 22, 22 and 25, and we assume that his measurements are size-biased, that is $w(u) = u$, $u \ge 0$. To save computations, it would be better to have $w_s(u) \equiv 1$, and so we labeled this scientist third, while the scientist who is labeled third in Example 1 is now labeled fourth; thus we have w_1 and w_2 as in Example 1, $w_3(u) = u$, $u \ge 0$, and $w_4(u) \equiv 1$. The algorithm (4.1) required 14 triple iterations

TABLE 1
Summary of data and the corresponding NPMLE. Entries in bold face are needed as input to the algorithm (4.1)

t_j	η_{1j}	η_{2j}	η_{3j}	$\mathbf{r}_j$	$\mathbf{w}_{1j}$	$\mathbf{w}_{2j}$	$\mathbf{w}_{3j}$	NPMLE
8.0	0	0	1	**1**	**0**	**½**	**1**	.10660
9.0	0	1	0	**1**	**0**	**½**	**1**	.10660
11.0	0	1	1	**2**	**1**	**1**	**1**	.11337
13.0	1	0	1	**2**	**1**	**1**	**1**	.11337
15.0	1	0	0	**1**	**1**	**1**	**1**	.05668
16.0	1	0	2	**3**	**1**	**1**	**1**	.17005
17.0	0	1	1	**2**	**1**	**1**	**1**	.11337
18.0	1	1	0	**2**	**1**	**1**	**1**	.11337
22.0	0	0	1	**1**	**0**	**½**	**1**	.10660
	$\mathbf{n}_1 = \mathbf{4}$	$\mathbf{n}_3 = \mathbf{4}$	$\mathbf{n}_3 = \mathbf{7}$					

TABLE 2
NPMLE for Example 2

t_j	$\hat{p}(t_j)$
8.0	.08323
9.0	.08111
11.0	.09015
13.0	.08768
15.0	.08533
16.0	.12631
17.0	.08311
18.0	.08204
19.0	.04050
22.0	.18289
25.0	.05766

TABLE 3
NPMLE for Example 3

t_j	$\hat{p}(t_j)$
9.0	.18654
11.0	.06006
13.0	.05798
15.0	.11207
16.0	.05511
17.0	.05422
18.0	.10671
19.0	.05251
22.0	.21624
25.0	.09856

to converge to $(\hat{V}_1, \hat{V}_2, \hat{V}_3) = (.59511, .79756, 15.95222)$ from the initial point $(a_2^0, a_3^0) = (20.0, 500.0)$ and the resulting NPMLE is given in Table 2.

EXAMPLE 3 ($s = 3$, $w_s(u) \not\equiv 1$). In this example we used the algorithm (4.1) with the data of the first, second and third scientists of Example 2 so that none of the weight functions is identically 1. The initial choice was $a_2^0 = 13$ and after 8 pairs of iterations it converged to $(\hat{V}_1, \hat{V}_2) = (a_1^8, a_2^8) = (.02983, .04482)$. This gave $\lambda = 16.71753$, so that $(\hat{W}_1, \hat{W}_2, \hat{W}_3) = \lambda(a_1^8, a_2^8, 1) = (.49866, .74933, 16.71753)$ and the NPMLE is given in Table 3.

6. Sufficiency of the NPMLE.

Aside from being an estimate of F, the main attraction of the empirical distribution function (EDF) as a data summary in random sampling from a cdf F, is that the pair (sample size, EDF) is a sufficient statistic. That is, all that can be learned about F from the raw data can be learned from the statistic (sample size, EDF). The following theorem shows that this property carries over to the NPMLE discussed in this paper.

THEOREM 3. *Suppose condition* (2.10) *is satisfied and let* $(t, \hat{p})$ *be the NPMLE described in Theorem* 2. *Then, given that we know the weight function for each of the s samples,* $(n_1, \cdots, n_s, (t, \hat{p}))$ *is a sufficient statistic for* F.

In other words, the theorem says that under condition (2.10) when we know the sampling rule (this is a different way of saying that we know the weight function), and size, for each of the s samples, the NPMLE $(t, \hat{p})$ is a sufficient statistic for F.

PROOF. Since $(t, \eta) \equiv ((t_1, \cdots, t_n), (\eta_{ij};\ i = 1, \cdots, s,\ j = 1, \cdots, h))$ is a sufficient statistic for F, and since $(n_1, \cdots, n_s, (t, \hat{p}))$ is a function of (t, η), the reader can easily check that the conditional probability $P_F[y_1, \cdots, y_s \mid (n_1, \cdots, n_s, (t, \hat{p}))$, condition (2.10) holds] is indepenent of F if for the random realization of the observed points, $t' = (t_1', \cdots, t_{h'}')$, and the random multiplicities matrix, $\eta' = \{\eta_{ij}';\ i = 1, \cdots, s, j = 1, \cdots, h'\}$, the conditional probability

$$P_F[t', \eta' \mid (n_1, \cdots, n_s, (t, \hat{p})), \text{condition (2.10) holds}] \tag{6.1}$$

is independent of F. Clearly, if t' and η' are such that condition (2.10) based on t' and η' is not satisfied, or if $t' \neq t$ or $\sum_j \eta_{ij}' \neq n_i$ for some $1 \leq i \leq s$, then (6.1) is zero, regardless of F. In the remaining case,

$$t' = t, \quad \textstyle\sum_{j=1}^h \eta_{ij}' = n_i, \quad i = 1, \cdots, s, \tag{6.2}$$

and η' is such that the NPMLE based on (t', η') is $(t, \hat{p})$. For this case we get, after some simple algebra,

$$(6.1) = \frac{(\prod_{i=1}^s \prod_{j=1}^h w_i(t_j)^{\eta_{ij}'}) \prod_{j=1}^h dF(t_j)^{r_j'}}{\sum_{\eta'' \text{as in (6.2)}} (\prod_{i=1}^s \prod_{j=1}^h w_i(t_j)^{\eta_{ij}''}) \prod_{j=1}^h dF(t_j)^{r_j''}}, \tag{6.3}$$

where,

$$r_j' = \textstyle\sum_{i=1}^s \eta_{ij}', \quad r_j'' = \sum_{i=1}^s \eta_{ij}''.$$

Now, we can easily see from Theorem 2 that all the η's that share the same NPMLE must have the same marginals r_j's and so we get from (6.3) that

$$(6.1) = \frac{\prod_{i=1}^s \prod_{j=1}^h w_i(t_j)^{\eta_{ij}'}}{\sum_{\eta'' \text{as in (6.2)}} \prod_{i=1}^s \prod_{j=1}^h w_i(t_j)^{\eta_{ij}''}}, \tag{6.4}$$

which is independent of F. This completes the proof. □

The following simple example captures the above ideas without obscuring them with the mathematical details of a formal proof.

EXAMPLE. Suppose, as in the first example in Section 2, that $s = 2$, $n_1 = n_2 = 2$, $w_1(u) = I[4 \leq u \leq 9]$, and $w_2(u) \equiv 1$. That is, we have a random sample of size 2 from the cdf F truncated to $[4, 9]$, and a random sample of size 2 from F.

Suppose further that the NPMLE exists and is given by $\hat{p}(1) = 1/2$, $\hat{p}(5) = \hat{p}(6) = \hat{p}(8) = 1/6$ (cf. end of the first example in Section 2). Then, the only possible data sets that could have resulted in the above NPMLE are (t, η') or (t, η'') or (t, η''') where $t = (1, 5, 6, 8)$ and

$$\eta' = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}, \quad \eta'' = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \quad \eta''' = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}. \tag{6.5}$$

Now, the matrix $w_i(t_j)$ is given by

$$\{w_i(t_j)\} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \tag{6.6}$$

and a substitution in (6.3) immediately verifies that the conditional probability for each of the possible data sets of (6.5) is $1/3$, independently of F.

Remark. The above example shows that the sufficient statistic $(n_1, \cdots, n_s, (t, \hat{p}))$ is a function of the sufficient statistic (t, η) *but not conversely*. The reader can compare the dimensions of the two sufficient statistics, $2h + s$ vs. $h(s + 1)$, to see that when $s \geq 2$ the use of $(n_1, \cdots, n_s, (t, \hat{p}))$ could result in a substantial reduction in dimension, by comparison to the use of (t, η).

7. General sample spaces. Since up until now we have never used the assumption that F is a univariate distribution in any important way (the ordering $t_1 < \cdots < t_h$ has been assumed only for convenience of notation), the reader could easily verify that the methodology herein, and in particular Theorems 1, 1′, 2, 3, the graphical criterion of Appendix A, the algorithm (4.1) all remain correct and applicable when the y_{ij}'s, $i = 1, \cdots, s$, $j = 1, \cdots, n_i$, are random elements in a *general* sample space, say Ω. In particular, if $\Omega = R^m$, the Euclidean m-space, then $y_{ij} \equiv (y_{ij}^{(1)}, \cdots, y_{ij}^{(m)})$, and the problem becomes: Given s random samples, with the ith sample, $y_{i1}, \cdots, y_{in_i}$ $(n_i \geq 1)$, being from the cdf

$$F_i(t^{(1)}, \cdots, t^{(m)}) \equiv W_i^{-1}(F) \int_{-\infty}^{t^{(1)}} \cdots \int_{-\infty}^{t^{(m)}} w_i(u^{(1)}, \cdots, u^{(m)})\, dF(u^{(1)}, \cdots, u^{(m)}),$$

find a NPMLE of the unknown, m dimensional, cdf F. Here, $w_i(u^{(1)}, \cdots, u^{(m)})$ are known, nonnegative, real, weight functions (which are strictly positive on a set of positive F measure) satisfying $0 < W_i(F) < \infty$.

The solution is, of course, the same as described in earlier sections, with $t_1, \cdots, t_h$ being $t_1 \equiv (t_1^{(1)}, \cdots, t_1^{(m)}), \cdots, t_h \equiv (t_h^{(1)}, \cdots, t_h^{(m)})$ and $p_1, \cdots, p_h$ being, as usual, $p(t_1), \cdots, p(t_h)$. The reader may find it instructive to go back to Table 1 of Section 5 and convince himself that the fact that the t_j's are real numbers is irrelevant to the derivation of $\hat{p}(t_j)$, and that the above statement is correct.

As will become apparent from the multivariate example of Section 8, the majority of applications of our methodology is expected to arise in multivariate situations.

8. More on applications and some additional remarks.

(i) *Other methods.* In order to find the NPMLE of F, our analysis suggests other possible approaches:

(a) Geometric programming techniques, which would usually involve formulating and solving the dual problem (Duffin, Peterson and Zener, 1967, Chapter III). See the remark following (3.12) in this connection.

(b) Try the following iterative scheme (suggested by (3.15)): If $p^{\text{old}} \equiv (p_1^{\text{old}}, \cdots, p_h^{\text{old}})$ is our current estimate of p then define

$$q_j = r_j/(n_s w_{sj} + \textstyle\sum_i n_i w_{ij} W_i^{-1}(p^{\text{old}})), \quad j = 1, \cdots, h,$$

and

$$p_j^{\text{new}} = q_j/\textstyle\sum_{k=1}^h q_k, \quad j = 1, \cdots, h.$$

(c) Try the following iterative scheme (suggested by (3.7)). Let $A^{\text{old}} \equiv (A_1^{\text{old}}, \cdots, A_{s-1}^{\text{old}})$ be the current estimate of $(V_1(p), \cdots, V_{s-1}(p))$, then define

$$A_i^{\text{new}} = \sum_j \frac{r_j w_{ij}}{n_s w_{sj} + \sum_{k=1}^{s-1} n_k w_{kj} (A_k^{\text{old}})^{-1}}, \quad i = 1, \cdots, s-1$$

(intermediate substitutions of A_k^{new} for A_k^{old} are also possible).

I have not tried to prove any convergence properties for (b) and (c) above because (4.1) seems simple, numerically efficient, and has the monotone convergence property demonstrated by C. L. Mallows (1985). As a remark, I would add that from a purely mathematical standpoint it is quite interesting to note that the h equations with h unknowns (3.15) plus the additional equation $\sum p_j = 1$ can be replaced with as few as $s - 1$ equations with $s - 1$ unknowns (3.7), and this is regardless of how large h is. Indeed the iterative scheme (4.1) is designed to take advantage of this fact, because in many applications one should expect to have $h \gg s$.

(ii) *NPMLE for* $s = 1$. In this case, it is easily verified that the NPMLE is $\hat{p}_j \propto \eta_{1j}/(n_1 w_{1j})$, $j = 1, \cdots, h$. Furthermore, denoting $n \equiv n_1$, $w \equiv w_1$, $W \equiv W_1(F)$, and assuming that F has a bounded density f, that w is positive on the support of F, and that W and

$$W_{-1} \equiv \int_{-\infty}^{\infty} w(y)^{-1} f(y)\, dy$$

are finite, we get the following results using standard limit theorems:

$$W(\hat{F}) = (n^{-1} \textstyle\sum_{j=1}^h w(y_j)^{-1})^{-1} \to W(F), \quad \text{w.p.1} \quad (F),$$

and

$$\sqrt{n}((\hat{W}/W) - 1) \to_d \textit{Normal}(0, W_{-1} W - 1).$$

When the observations are univariate we also have,

$$\sqrt{n}(\hat{F} - F) \rightarrow_d Z,$$

where Z is a pinned Gaussian process with mean zero and covariance function

$$c(s, t) = W_{-1} W \left\{ \frac{\int_{-\infty}^{s} w(y)^{-1} f(y)\, dy}{W_{-1}} (1 - F(t)) + F(s) \left(F(t) - \frac{\int_{-\infty}^{t} w(y)^{-1} f(y)\, dy}{W_{-1}} \right) \right\}, \qquad s \le t.$$

The method of proof is similar to that of Theorem 3.2 in Vardi (1982).

(iii) *Incorporating censored samples.* (For simplicity, we assume here that the observations are univariate.) Suppose that in addition to the original information $t_j, \eta_{1j}, \cdots, \eta_{sj}$, $j = 1, \cdots, h$, we are also given nonnegative integers, $\eta'_{1j}, \cdots, \eta'_{sj}$, where η'_{ij} is the multiplicity of observations from F_i for which it is only known that their values are *at least* t_j, $j = 1, \cdots, h$. (This is often called "arbitrary censoring" model.) The probability of the data is then

$$(8.1) \qquad P_F(\text{data}) = \prod_{j=1}^{h} \left\{ \prod_{i=1}^{s} \left(\frac{w_i(t_j)\, dF(t_j)}{W_i(F)} \right)^{\eta_{ij}} \left(\frac{\int_{t_j}^{\infty} w_i(u)\, dF(u)}{W_i(F)} \right)^{\eta'_{ij}} \right\}.$$

Unlike the situation in (2.1) it is not clear that the maximum of (8.1) can be attained with a cdf F that assigns positive mass only to the set $t_1, \cdots, t_h$. Nevertheless for any set of points $\tau \equiv \{\tau_1, \cdots, \tau_{\bar{h}}\}$ which includes $\{t_1, \cdots, t_h\}$, the cdf $\hat{F}$ (depends on τ) which maximizes (8.1), over all possible cdf's that assign positive mass only to τ, can be obtained using (4.1) and the EM algorithm (e.g. Dempster, Laird and Rubin, 1977) as follows: start with an initial estimate p^{old} satisfying $\sum_{j=1}^{\bar{h}} p^{\text{old}}(\tau_j) = 1, p^{\text{old}}(\tau_j) > 0, j = 1, \cdots, \bar{h}$. Proceed to compute (E-step) $\tilde{r}(\tau_j)$ = the expected conditional multiplicities of observations from the "complete data" (in the language of the EM paper) at the point τ_j, given our current estimate $p^{\text{old}}(\tau_j)$, $j = 1, \cdots, \bar{h}$, and the data. Then proceed to the M-step, in which you use (4.1) to solve

$$A_i^{-1} \sum_{j=1}^{\bar{h}} \frac{\tilde{r}(\tau_j) w_i(\tau_j)}{(n_s + n'_s) w_s(\tau_j) + \sum_{k=1}^{s-1} (n_k + n'_k) w_k(\tau_j) A_k^{-1}} = 1, \quad i = 1, \cdots, s - 1,$$

$(n'_k \equiv \sum_j \eta'_{kj}, k = 1, \cdots, s)$ for $A_1, \cdots, A_{s-1}$. Denote the solution by V^{new} and set

$$p^{\text{new}}(\tau_j) \propto \frac{\tilde{r}(\tau_j)}{(n_s + n'_s) w_s(\tau_j) + \sum_{k=1}^{s-1} (n_k + n'_k) w_k(\tau_j) (V_k^{\text{new}})^{-1}}, \quad j = 1, \cdots, \bar{h}.$$

The scheme is then iterated (until we get numerical convergence) by substituting p^{new} for p^{old} at the end of each iteration and returning to the E-step.

(iv) *On the asymptotics.* (For simplicity, we assume here that the observations are univariate.) The merit of an estimator is usually judged by its asymptotic behavior. In Vardi (1982) I prove that for the case $s = 2$, $w_1(u) = u$ (length biasing) and $w_2(u) \equiv 1$, the asymptotic behavior of the NPMLE is similar to that of the empirical distribution function of a random sample from a cdf. The same technique can be used to derive the asymptotic behavior of the NPLME when $s = 2$ and $w_1(\cdot)$ is any *arbitrary* weight function. In particular, under the assumptions that F is absolutely continuous with respect to the Lebesgue measure, and $N \equiv n_1 + n_2$ approaches infinity such that $N^{-1}n_2$ remains fixed and positive, we get that $(W_1(\hat{F}) - W_1(F))$ converges almost surely (F) to zero, $\sqrt{N}(W_1(\hat{F}) - W_1(F))$ converges weakly to a normal distribution with mean zero and variance σ^2, and $\sqrt{N}(\hat{F} - F)$ converges in distribution to a pinned Gaussian process with mean zero and covariance $c(s, t)$. The derivation of σ^2 and $c(s, t)$ is similar to the derivation of the variance and covariance terms in (3.3) and (3.6) of Vardi (1982).

The asymptotics for the case $s \geq 2$, and arbitrary weight functions, needs further study. One possible way to proceed is to adapt the method in Vardi (1982) to this situation as follows: Suppose F has a density f, the w_i's are such that the union of the supports of the F_i's is the support of F (otherwise the problem should be reformulated as estimating F restricted to the union of the supports of the F_i's) and the n_i's approach ∞, such that $\lambda_i \equiv n_i/(n_1 + \cdots + n_s) > 0$ remains fixed, $i = 1, \cdots, s$. Using the strong law of large numbers (SLLN), the limiting form of (3.7) becomes

$$H_i^*(A_1, \cdots, A_{s-1}) = 1, \quad i = 1, \cdots, s - 1, \tag{8.2}$$

where

$$H_i^*(A_1, \cdots, A_{s-1}) \equiv A_i^{-1} \int_{-\infty}^{\infty} w_i(y) \frac{\sum_{k=1}^{s} \lambda_k w_k(y) W_k(F)^{-1}}{\sum_{k=1}^{s} \lambda_k w_k(y) A_k^{-1}} f(y)\, dy; \tag{8.3}$$

here $A_s \equiv 1$ and $i = 1, \cdots, s - 1$. The almost sure convergence of $\hat{F}$ to F, and other standard asymptotic properties of $\hat{F}$ can then be deduced after proving the following hypothesis: (8.2) *has a unique solution whenever the* w_i*'s are such that in the limit, as the* n_i*'s* $\rightarrow \infty$, *condition* (2.10) *holds with probability one.* The reason why this hypothesis is relevant to the derivation of the asymptotic behavior of $\hat{F}$ is that,

$$A_i = W_i(F)/W_s(F), \quad i = 1, \cdots, s - 1, \tag{8.4}$$

is *always* a solution of (8.2), and so if the w_i's are such that (8.2) has a *unique* solution, the solution of (3.7) approaches (8.4) w.p.1 (F). This, plus a similar argument for the limiting version of (3.17), lead to the almost sure convergence of $W_i(\hat{F})$ to $W_i(F)$, $i = 1, \cdots, s$, and subsequently to the almost sure convergence of $\hat{F}(t)$ to $F(t)$, and other convergence results of the type given in Vardi (1982).

The above hypothesis can be further simplified. Define a graph G^* with s vertices, and an *edge* from vertex i to vertex i', $i \leftrightarrow i'$, if and only if,

$$\int_{-\infty}^{\infty} w_i(y) w_{i'}(y) f(y)\, dy > 0. \tag{8.5}$$

Since

$$n_{i'}^{-1} \textstyle\sum_j w_{ij}\eta_{i'j} \rightarrow W_{i'}(F)^{-1} \displaystyle\int_{-\infty}^{\infty} w_i(y)w_{i'}(y)f(y)\ dy, \quad \text{w.p.1} \quad (F),$$

an immediate application of the SLLN to the criterion of Appendix A gives the following.

LEMMA. *Under the assumption that the n_i's → ∞ and the λ_i's remain constant, condition* (2.10) *is satisfied with probability one* (F) *if, and only if, the graph G^* is connected* (*i.e. any two vertices are connected by a path*).

This reduces the hypothesis above to the following, equivalent hypothesis: (8.2) *has a unique solution whenever the w_i's are such that G^* is connected.*

We leave the problem of proving this hypothesis open. Once it is proved, however, the asymptotic properties of $\hat{F}$ could be derived along lines similar to Vardi (1982).

(v) *More on applications.* It is often the case that, due to technological advances, new measuring equipment have larger measuring range and so data collected using new equipment would be truncated into a bigger interval than data collected using old equipment. This gives rise to examples of the type described in the introduction, for which the conditions of Theorem 1 would typically hold, so that a NPMLE would exist. As an example of the above, Professor M. Alvo (University of Ottawa, personal communication) pointed out to me that in water quality studies, past experiments recorded the amount of phosphorus in water down to a level of .05 milligram/liter while more recent experiments measure it down to .03 milligram/liter. Also in experiments that involve human hearing (and other aspects of human perception) it is often natural to associate a weight function with each person that participates in the experiment, and so the data collected from such an experiment would be of the type we discussed.

In estimating heights on the basis of historical samples of military, naval and merchant marine, Wachter and Trussel (1982) noted that the samples suffer from undercounts of short people, and since "independence between selection probability and height or any other specifiable relationship between them cannot be assumed," the authors developed two original estimation methods, relying in part on the well accepted Gaussian model for heights, for ages after terminal heights are attained. It seems to me that by trying various plausible weight functions $w(\cdot)$, which represent the selection probability mechanism, one can apply the methodology I present in this paper to the heights data in a model free framework. This might be advantageous when incorporating data of adolescent heights for whom the Gaussian assumption may be violated (Section 5, Wachter and Trussell), and it also allows experimentation with different selection mechanisms for the various samples. This would be done, of course, by choosing the weight function $w(\cdot)$ of the military sample (say) to be different than the weight function of the merchant marine sample (say).

The following multivariate example, suggested to me by W. S. DeSarbo, demonstrates a class of applications frequently occurring in marketing research. Suppose we are interested in estimating the distribution of $Y^{(1)}$, the annual amount of money an individual spends on a particular product, say, and let $Y^{(2)}$ denote a "proxy" variable, such as the individual's annual income. Then, it is often the case that the data available to us are made up of several samples, and in each sample a different selection mechanism, which depends on the values of the proxy variables, is used. To be specific, suppose we have three samples ($s = 3$), and that in the first sample the selection rule was independent of the values of the attribute of interest, $Y^{(1)}$, or the values of the proxy variable $Y^{(2)}$; in the second sample the selection rule was such that individuals with annual income, $Y^{(2)}$, of less than \$10,000 were excluded from the sampled population; and in the third sample individuals with annual income, $Y^{(2)}$, of more than \$30,000 were excluded from the sampled population. The problem is to find the NPMLE of the cdf of $Y^{(1)}$. The way to solve this problem is the following: First, we identify the weight function associated with each sample; this gives $w_1(u^{(1)}, u^{(2)}) = 1$, $w_2(u^{(1)}, u^{(2)}) = I[u^{(2)} \geq 10{,}000]$, and $w_3(u^{(1)}, u^{(2)}) = I[u^{(2)} \leq 30{,}000]$. Second, we check (using the graphical criterion of Appendix A) whether there exists a unique NPMLE, $\hat{F}_{12}(u^{(1)}, u^{(2)})$, for the joint distribution of $(Y^{(1)}, Y^{(2)})$, and suppose the answer is positive. We then proceed to derive $\hat{F}_{1,2}$ using Theorem 2 and the algorithm (4.1). The NPMLE of the cdf of $Y^{(1)}$ is then given by $\hat{F}_1(u^{(1)}) \equiv \hat{F}_{1,2}(u^{(1)}, \infty)$.

The case of unknown weight functions. In many applications (e.g. Williams, 1978, Wachter and Trussel, 1982, and more) the *exact* sampling rules are unknown, which means that the weight functions are not specified in advance. In such situations, there are three main routes to proceed: (a) We can estimate the weight functions by conducting a separate study. (b) We can assume a parametric form for the weight functions and estimate the parameters from our data. (c) We can postulate a reasonable model for the true, unknown, sampling mechanism which incorporates a prior distribution on the weight functions. Of course, combinations of the above such as assuming a parametric form for the weight functions and then assuming a prior distribution for the parameters are also possible. Whichever of these approaches we take, our technique is needed as part of the estimation procedure. If (a) is chosen and $\hat{w} = (\hat{w}_1, \cdots, \hat{w}_s)$ are the estimated weight functions, we'll base our estimate of F on the estimated weight function, that is $\hat{F} \equiv \hat{F}_{\hat{w}}$. Of course, this estimate has greater variance than if the weight functions were known but that should only be expected. If (b) is chosen, so that the weight functions are $w_i(\theta, t)$ for some known form of w_i but unknown parameter θ, an iterative technique which incorporates the algorithm (4.1) can then be used to derive the joint maximum likelihood estimate, $(\hat{\theta}, \hat{F}_{\hat{\theta}})$, of θ and F. It's possible, however, that further smoothness assumptions will have to be imposed on F in order to make the combined estimation problem identifiable. (In connection with this parametric approach, note that the exponential family of distributions has the general form of weighted distributions.) If (c) is chosen and $G(w) \equiv G(w_1, \cdots, w_s)$ is the prior distribution over the weight functions, then

$$\hat{F}(t) = \int_w \hat{F}_w(t)\, dG(w), \tag{8.6}$$

where $\hat{F}_w$ is the NPMLE for *known* weight functions $w \equiv (w_1, \cdots, w_s)$, is the Bayes EDF based on our data. Note that since the $\hat{F}_w(t)$'s put zero mass outside of $t_1, \cdots, t_h$ irrespective of the weight functions $w = (w_1, \cdots, w_s)$, the Bayes EDF (8.6) is easily obtained by computing the probability mass that $\hat{F}_w$ assigns to $t_1, \cdots, t_h$ for each w and then averaging over all w's with the weights $dG(w)$.

Another important class of selection bias problems, for which our methodology can be adapted, arise in estimating the behavioral relationships in regression models based on several, nonrandomly selected, samples. See J. J. Heckman (1979) and the references thereof for a treatment of such a problem, in a parametric framework, under the normal distribution assumption.

More applications related to our methodology can be found, among other places, in Cox (1969), Tuma (1982), Turnbull (1976), Vardi (1982), Patil and Rao (1977), and in the interesting examples of selection bias described in Williams (1978).

Acknowledgement. I would like to thank F. K. Hwang for the graphical interpretation of condition (2.10), and J. A. Wellner for pointing out an error which led to a more precise statement of the results. Technical discussions with C. L. Mallows throughout this research, and some of the referee's suggestions improved this paper.

APPENDIX A

A Graphical Criterion for Checking Whether a Unique NPMLE Exists

The following criterion, suggested by F. K. Hwang, for testing the existence and uniqueness of a NPMLE replaces checking condition (2.10) for *each* subset B of $\{1, \cdots, s\}$ with checking whether a certain directed graph with s vertices is strongly connected (for which efficient algorithms exist). Thus, for large values of s this method is numerically efficient. For small values of s this method has the advantage that the strong connectivity of the graph could be checked by visually inspecting the graph.

Define a directed graph G with s vertices and a directed edge from vertex i to vertex i', $i \rightarrow i'$, if and only if, there exist a j such that $w_{ij} > 0$ and $\eta_{i'j} > 0$. (Note that this condition is equivalent to $\sum_j w_{ij}\eta_{i'j} > 0$.) G is called *strongly connected* if for any two vertices x and x' there exists a directed path from x to x', and a directed path from x' to x.

THEOREM. *Condition* (2.10) *holds for each proper subset B of* $\{1, \cdots, s\}$ *if, and only if, G is strongly connected.*

PROOF. Recall that for each subset B of $\{1, \cdots, s\}$, D_B is a subset of $\{t_1, \cdots, t_h\}$ which is defined by $D_B = \{t_j;\ w_{ij} > 0 \text{ for some } i \in B\}$, and suppose that G is strongly connected. Then for any subset B of vertices there always exists an edge from a vertex in B to a vertex not in B, say from x in B to x' not in B. The existence of the edge $x \longrightarrow x'$ implies the existence of a j such that $w_{xj} > 0$ and $\eta_{x'j} > 0$. Since $t_j \in D_B$, $\eta_{x'j}$ contributed to the left-hand side of (2.10), but not to the right-hand side (because $x' \notin B$). Since every η_{ij} in the right-hand side of (2.10) is also in the left-hand side, a strict inequality must hold and (2.10) is proved.

Conversely, suppose G is not strongly connected. Then there exists a subset B of vertices such that there exists no edge from B to outside B. This means that there exists no j such that $w_{ij} > 0$ and $\eta_{i'j} > 0$ for some $i \in B$ and $i' \notin B$. In other words, there exists no η_{ij} contributing to the left-hand side of (2.10) but not to the right-hand side. Hence in (2.10) we have equality rather than strict inequality. This completes the proof. □

REMARK. Since reversing the direction of all the edges does not effect the strong connectivity of a directed graph, replacing the above definition of $i \longrightarrow i'$ with "$i \longrightarrow i'$ iff $\sum_j w_{i'j}\eta_{ij} > 0$" would result in the same criterion.

APPENDIX B

Some Complementary Proofs

A proof of the equivalence between (3.2–3.3) *and* (3.11–3.12).

Clearly (3.2–3.3) is equivalent to

$$\text{(B.1)} \qquad \text{minimize } Q(q, u) \equiv \textstyle\prod_j q_j^{-r_j} \prod_i u_i^{n_i}$$

subject to

$$\text{(B.2)} \qquad \begin{aligned} \textstyle\sum_j w_{sj}q_j &= 1, \\ \textstyle\sum_j w_{ij}q_j &= u_i, \quad i = 1, \cdots, s-1, \\ u_i &> 0, \quad i = 1, \cdots, s-1, \\ q_j &> 0, \quad j = 1, \cdots, h. \end{aligned}$$

Since the constraint region (B.2) is strictly contained in (3.12), it is enough to show that if (q', u') satisfies (3.12) then there exists (q'', u'') satisfying (B.2) such that

$$Q(q'', u'') \le Q(q', u').$$

Let $\zeta = \sum_j w_{sj}q_j'$ and note that from (3.12) $0 < \zeta \le 1$. Define

$$q_j'' \equiv \zeta^{-1}q_j', \quad u_i'' \equiv \textstyle\sum_j w_{ij}q_j''.$$

Then (q'', u'') satisfies (B.2) and, again from (3.12),

$$u_i'' = \zeta^{-1} \textstyle\sum_j w_{ij}q_j' \le \zeta^{-1}u_i'.$$

Thus

$$Q(q'', u'') \le \zeta^{\sum_{j=1}^{h} r_j - \sum_{i=1}^{s-1} n_i} Q(q', u') = \zeta^{n_s} Q(q', u') \le Q(q', u'),$$

as desired. □

A Proof of "Argument II" (*cf. the proof of Theorem* 1), *which states that if the solution of* (2.2–2.3) *is nonunique, then condition* (2.10) *is not satisfied.*

First we note that since the problem (2.2–2.3) is equivalent to (3.2–3.3) which in turn is equivalent to (3.13–3.14), the assumption that (2.2–2.3) has two solutions, say p' and p'', $p' \ne p''$, implies that (3.13–3.14) also has two solutions, say (α', β') and (α'', β''), $(\alpha', \beta') \ne (\alpha'', \beta'')$. By taking log in (3.13), and using the equivalence of (2.2–2.3) and (3.13–3.14), this implies that the problem

$$\text{(B.3)} \qquad \text{minimize } g_0(\alpha, \beta) \equiv \textstyle\sum_{j=1}^{h} r_j \alpha_j + \sum_{i=1}^{s-1} n_i \beta_i$$

subject to

$$\text{(B.4)} \qquad \begin{aligned} g_s(\alpha) &\equiv \textstyle\sum_j w_{sj} e^{-\alpha_j} \le 1, \\ g_i(\alpha, \beta) &\equiv \textstyle\sum_j w_{ij} e^{-(\alpha_j + \beta_i)} \le 1, \quad i = 1, \cdots, s-1, \end{aligned}$$

has two different solutions, say (α', β') and (α'', β''). Since the constraint region (B.4) is convex and g_0 of (B.3) is linear, any convex combination of the two solutions, say,

$$(\alpha(\theta), \beta(\theta)) \equiv \theta(\alpha', \beta') + (1-\theta)(\alpha'', \beta''), \quad 0 \le \theta \le 1,$$

is also a solution of (B.3–B.4). Now, since $(\alpha(\theta), \beta(\theta))$ is a minimizer of (B.3–B.4) it follows [from the same argument given in the beginning of this Appendix, which established the equivalence of (B.1–B.2) and (3.11–3.12)] that (B.4) is satisfied with equalities. Namely,

$$\text{(B.5)} \qquad g_s(\alpha(\theta)) = 1, \quad 0 \le \theta \le 1,$$

$$\text{(B.6)} \qquad g_i(\alpha(\theta), \beta(\theta)) = 1, \quad 0 \le \theta \le 1, \quad i = 1, \cdots, s-1.$$

DEFINITION. For $i, i' \in \{1, \cdots, s\}$ we say that i *communicates* with i' iff either $\tilde{D}_i \cap \tilde{D}_{i'} \ne \varnothing$ or there exists a subset $\{i_1, \cdots, i_m\}$ of $\{1, \cdots, s\}$ such that

$$\tilde{D}_i \cap \tilde{D}_{i_1} \ne \varnothing, \ \tilde{D}_{i_1} \cap \tilde{D}_{i_2} \ne \varnothing, \ \cdots, \ \tilde{D}_{i_m} \cap \tilde{D}_{i'} \ne \varnothing.$$

CLAIM (to be proved below). If i communicates with s, then $\beta'_i = \beta''_i$ and $\alpha'_j = \alpha''_j$ for $j \in \tilde{D}_i$.

Since, by assumption, $(\alpha', \beta') \ne (\alpha'', \beta'')$ it follows from the above claim that not all i's communicate with s and so let B be the *largest subset of* $\{1, \cdots, s\}$ *such that none of its elements communicate with* s. From the definition of B, no observation from y_i, $i \notin B$, could belong to D_B and so

$$\textstyle\sum_{j \in \tilde{D}_B} r_j = \sum_{i \in B} n_i.$$

Thus condition (2.10) is not satisfied and the result follows. It remains therefore to prove the claim.

PROOF OF THE CLAIM. First note that from (B.5)

$$\textstyle\sum_{j\in\tilde{D}_s} w_{s_j} e^{-\alpha_j(\theta)} = 1, \quad 0 \le \theta \le 1.$$

Since e^x is *strictly* convex in x, $\alpha_j(\theta)$ must be constant for all $0 \le \theta \le 1$ and so

$$\alpha_j' = \alpha_j'' \quad \text{for} \quad j \in \tilde{D}_s. \tag{B.7}$$

Suppose now that i communicates *directly* with s, so that

$$\tilde{D}_i \cap \tilde{D}_s \neq \varnothing.$$

We get from (B.6)

$$\textstyle\sum_{j\in\tilde{D}_i\cap\tilde{D}_s} w_{ij}\exp[-(\alpha_j(\theta) + \beta_i(\theta))] + \sum_{j\in\tilde{D}_i\cap\tilde{D}_s^c} w_{ij}\exp[-(\alpha_j(\theta) + \beta_i(\theta))] = 1. \tag{B.8}$$

Combining (B.7) with (B.8) we get, after some simple algebra,

$$\begin{aligned}&\textstyle\sum_{j\in\tilde{D}_i\cap\tilde{D}_s} w_{ij}\exp[-\alpha_j' - \beta_i'' + (\beta_i'' - \beta_i')\theta]\\ &\quad + \textstyle\sum_{j\notin\tilde{D}_i\cap\tilde{D}_s^c} w_{ij}\exp[-\alpha_j'' - \beta_i'' + (\alpha_j'' - \alpha_j' + \beta_i'' - \beta_i')\theta] = 1,\\ &\hspace{22em} 0 \le \theta \le 1.\end{aligned} \tag{B.9}$$

Analytic continuation then implies that (B.9) holds for $-\infty < \theta < \infty$, which in turn implies that

$$\beta_i'' - \beta_i' = 0 \quad \text{and} \quad \alpha_j'' - \alpha_j' + \beta_i'' - \beta_i' = 0, \quad j \in \tilde{D}_i \cap \tilde{D}_s^c, \tag{B.10}$$

Combining (B.10) and (B.7) we conclude that if i communicates with s then

$$\beta_i' = \beta_i'' \tag{B.11}$$

and

$$\alpha_j' = \alpha_j'' \quad \text{for} \quad j \in \tilde{D}_i \cup \tilde{D}_s. \tag{B.12}$$

An induction proof, based on the above argument, shows that if i communicates with s via $\{i_1, \cdots, i_m\}$ then

$$\beta_i' = \beta_i'',\ \beta_{i_1}' = \beta_{i_1}'',\ \cdots,\ \beta_{i_m}' = \beta_{i_m}'' \tag{B.13}$$

and

$$\alpha_j' = \alpha_j'' \quad \text{for} \quad j \in (\cup_{k=1}^m \tilde{D}_{i_k}) \cup \tilde{D}_i \cup \tilde{D}_s. \tag{B.14}$$

This proves the claim and finishes the proof of "Argument II", which is part of the proof of Theorem 1. □

From the proof of Argument II we get the following.

COROLLARY. *Let G_{ww} be a graph with s vertices where $i \leftrightarrow i'$ iff $\sum_j w_{ij}w_{i'j} > 0$, $i, i' \in \{1, \cdots, s\}$, and suppose G_{ww} is connected. Then, if* (2.2–2.3) *has a solution it must be unique.*

PROOF. If G_{ww} is connected, all i's in $\{1, \cdots, s\}$ communicate with s, and so it follows from the above claim that the solution must be unique.

REFERENCES

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

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39** 1–37.

DUFFIN, D. J., PETERSON, E. L. and ZENER, C. (1967). *Geometric Programming—Theory and Application.* Wiley, New York.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

MALLOWS, C. L. (1985). Discussion of "Empirical Distributions in Selection Bias Models" by Y. Vardi. *Ann. Statist.* **13** 204–205.

ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic, 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.

SCHOLZ, F. N. (1980). Towards a unified definition of maximum likelihood. *Canad. J. Statist.* **8** 193–203.

TUMA, N. B. (1982). Comment on Wachter and Trussel. *J. Amer. Statist. Assoc.* **77** 297–301.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. B* **38** 290–295.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10** 616–620.

WACHTER, K. W. and TRUSSELL, J. (1982). Estimating historical heights. *J. Amer. Statist. Assoc.* **77** 279–293.

WILLIAMS, W. H. (1978). *A Sampler on Sampling.* Wiley, New York.

ZANGWILL, W. I. (1969). *Nonlinear Programming A Unified Approach.* Prentice Hall, New York.

AT&T BELL LABORATORIES
MURRAY HILL, NEW JERSEY 07974