1989-06

Estimation of Multivariate Distributions Under Stochastic Ordering

Sampson, Allan R.; Whitaker, Lyn R.

Journal of the American Statistical Association, June 1989, Vol. 84, No. 406, Theory and Methods
http://hdl.handle.net/10945/41703

Estimation of Multivariate Distributions Under Stochastic Ordering

ALLAN R. SAMPSON and LYN R. WHITAKER*

Let F and G be the cdf's of two p-dimensional multivariate distributions, such that F is stochastically larger than G. A straightforward derivation is given of the generalized maximum likelihood estimators of F and G, based on random samples from each population. An algorithmic approach to computing these estimators is described and motivating numerical examples are discussed. The special case when F and G correspond to multivariate ordinal contingency tables is also presented. The relationship of these results to those of Robertson and Wright (1974) is considered.

KEY WORDS: Isotonic regression; Maximum likelihood estimation; Ordinal contingency tables; Two-sample problem.

1. INTRODUCTION

The problem of comparing two populations is basic, and widely encountered in statistics. Both parametric and nonparametric procedures are available for a variety of settings (see Hettmansperger 1984, chap. 3; Miller 1985, chap. 2). An appealing framework, often used to compare two univariate populations, assumes that one population is stochastically larger than the other; that is, larger values are more likely from one population than the other. Specifically, let X and Y be random variables from populations with respective cumulative distribution functions (cdf's) F and G. Then, F is said to be stochastically greater than G, denoted by $F \overset{st}{\geq} G$, or equivalently $X \overset{st}{\geq} Y$, if $\bar{F}(t) \geq \bar{G}(t)$ for all t. Stochastic ordering of univariate distributions extends to multivariate distributions. Let $\mathbf{X}$ and $\mathbf{Y}$ be p-dimensional random vectors with cdf's F and G, respectively. We say that F (or $\mathbf{X}$) is stochastically greater than G (or $\mathbf{Y}$), denoted by $F \overset{st}{\geq} G$ (or $\mathbf{X} \overset{st}{\geq} \mathbf{Y}$), if

$$E[h(\mathbf{X})] \geq E[h(\mathbf{Y})], \tag{1.1}$$

for all functions h that are nondecreasing in each argument and have finite expectations. Marshall and Olkin (1979, p. 483) showed that (1.1) is equivalent to the existence of random vectors $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ defined on the same probability space with $\Pr(\tilde{\mathbf{X}} \geq \tilde{\mathbf{Y}}) = 1$, where $\geq$ denotes componentwise ordering.

The multivariate version of stochastic ordering also conveys the intuition that larger values are more likely for one population than another. To gain further insight into multivariate stochastic ordering, consider the two-sample bivariate setting where experimental units are randomly assigned to either a control treatment (with cdf G) or an experimental treatment (with cdf F), and assume that larger bivariate values are associated with a better response. Based on each bivariate response (u, v), define a one-dimensional binary variable, $S(u, v)$, indicating whether response is satisfactory. We would expect experts to have different definitions for a satisfactory response, but most would agree that $S(u, v)$ must satisfy the criteria that if (u_1, v_1) is satisfactory and (u_2, v_2) is better, then (u_2, v_2) must be considered a satisfactory response; that is, if $S(u_1, v_1) = 1$ and $u_2 \geq u_1$ and $v_2 \geq v_1$, then $S(u_2, v_2) = 1$. With this in mind, $F \overset{st}{\geq} G$ has the interpretation that the probability of satisfactory response under the experimental treatment is larger than the probability of satisfactory response under the control, no matter how S is chosen.

Inference based on two-sample stochastically ordered univariate data has been studied extensively. The theory of isotonic regression has played a key role in deriving maximum likelihood estimators (MLE) and likelihood ratio tests (LRT) for two or more stochastically ordered populations. Brunk, Franck, Hanson, and Hogg (1966) obtained the MLE's of F and G under the assumption that $F \overset{st}{\geq} G$; see Barlow, Bartholomew, Bremner, and Brunk (1972) for a comprehensive review of the literature for this and related problems prior to 1972. Grove (1980) and Robertson and Wright (1981) obtained and studied the LRT for $H_0 : F = G$ against $H_a : F \overset{st}{\geq} G$, when F and G are discrete distributions on the same k points, or equivalently when F and G are multinomial distributions with the same k ordered categories. Lee (1987) obtained MLE's for multinomial distributions with fixed and random zeros. Franck (1984) considered the LRT for these hypotheses when the data come from arbitrary distributions, and obtained approximations to the distribution of the LRT. For censored multinomial data, Dykstra (1982) obtained the MLE's assuming $F \overset{st}{\geq} G$, and Dykstra, Madsen, and Fairbanks (1983) obtained the LRT of $H_0 : F = G$ versus $H_a : F \overset{st}{\geq} G$. Other research includes Robertson and Wegman (1978), Robertson and Wright (1982), and Dykstra and Robertson (1983).

Statistical inference in the multivariate setting has received much less attention. A notable exception is Robertson and Wright (1974), who for a general setting gave a theoretical derivation and consistency arguments for the MLE's of F and G under the assumption $F \overset{st}{\geq} G$. The

* Allan R. Sampson is Professor, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260. Lyn R. Whitaker is Assistant Professor, Department of Operations Research, U.S. Naval Postgraduate School, Monterey, CA 93940. Sampson's research was done in part at Carnegie-Mellon University, Department of Statistics, and sponsored by U.S. Air Force Office of Scientific Research Contract 84-0113. Whitaker's research was sponsored by U.S. Air Force Office of Scientific Research Contract 84-0159. The authors thank both referees for their thoughtful comments, which led to a substantially improved revision.

multivariate setting poses more difficulties than the univariate setting, because multivariate estimators cannot generally be recast as solutions to standard isotonic regression problems. In addition, once derived, numerical evaluation of these estimators is extremely difficult.

We obtain the MLE's of F and G in the multivariate case, using a different approach from Robertson and Wright (1974). We reduce the two-sample problem to two one-sample problems using a multivariate version of the techniques of Barlow and Brunk (1972). This reduction permits the use of existing algorithms, to obtain numerical approximations for the MLE's. In addition, we can solve several one-sample multivariate problems that are of interest in their own right.

In Section 2 we present specific formulations of several estimation problems that occur in practice. The basic theoretical results are given in Section 3. Computational procedures and numerical examples are discussed in Section 4. In Section 5, we include a brief discussion of related problems.

2. PROBLEM MOTIVATION AND FORMULATION

In this section, we discuss several settings for which estimation under stochastic ordering is applicable. In each setting we explicitly formulate the corresponding optimization problem whose solution is given in Section 3. To do so, we need an upper (lower) set. Let $\mathbf{u} = (u_1, \ldots, u_p)$ and $\mathbf{v} = (v_1, \ldots, v_p)$ be vectors in R^p; then, $U(L)$ is an upper (lower) set in R^p if $\mathbf{u} \in U$ ($\mathbf{u} \in L$) and $u_i \le v_i$ ($u_i \ge v_i$) for $i = 1, \ldots, p$ imply that $\mathbf{v} \in U$ ($\mathbf{v} \in L$). An alternate formulation of stochastic ordering (see Marshall and Olkin 1979, prop. 17.B2) equivalent to (1.1) is

$$\Pr(\mathbf{X} \in U(L)) \ge (\le) \Pr(\mathbf{Y} \in U(L)) \quad \text{for all } U \in \mathfrak{U} \text{ (for all } L \in \mathfrak{L}), \tag{2.1}$$

where $\mathfrak{U}(\mathfrak{L})$ is the class of all upper (lower) sets $U(L)$ in R^p.

For notational convenience, we restrict attention to bivariate data; however, our discussion clearly applies to higher-dimensional data.

2.1 Two-Sample Problems

Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are two discrete bivariate random vectors with common support on an $I \times J$ lattice. Equivalently, consider two ordinal contingency tables with I row categories and J column categories. Let $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$, where $p_{ij} = \Pr(\mathbf{X} = (i, j))$ and $q_{ij} = \Pr(\mathbf{Y} = (i, j))$ are the probabilities associated with the first and second tables, respectively. We are interested in the parameter space for which $\mathbf{X} \overset{st}{\ge} \mathbf{Y}$. Because P and Q have support in the $I \times J$ lattice, the constraint (2.1) becomes

$$\sum_{(i,j)\in U} p_{ij} \ge \sum_{(i,j)\in U} q_{ij} \quad \text{for all } U \in \mathfrak{U}_{I\times J}, \tag{2.2}$$

where $\mathfrak{U}_A \equiv \{U \cap A : U \in \mathfrak{U}\}$, for any subset A of R^p. Independent random samples of size m and n are chosen from P and Q, respectively, resulting in the observed contingency tables $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$, where a_{ij} and b_{ij} are the frequencies associated with the (i, j)th cell. Finding the MLE's of P and Q subject to (2.2) is equivalent to maximizing

$$\prod_{j=1}^{J} \prod_{i=1}^{I} p_{ij}^{a_{ij}} q_{ij}^{b_{ij}}, \tag{2.3}$$

subject to (2.2), $\sum_{i,j} p_{ij} = \sum_{i,j} q_{ij} = 1$, $p_{ij} \ge 0$, and $q_{ij} \ge 0$, and where $0^0 = 1$.

Interestingly, multivariate stochastic ordering has a property that is desirable when analyzing categorical data. If we collapse two or more adjacent columns or rows in both tables, $P \overset{st}{\ge} Q$ implies that the distributions for the collapsed tables will also be stochastically ordered.

In Section 4, we analyze the well-known British and Danish social-mobility data (see Bishop, Fienberg, and Holland 1975), under the assumption that the British population is stochastically larger than the Danish.

For the general two-sample problem, there is a variety of potential applications of stochastic ordering, such as in medicine and reliability engineering. Specifically, consider a reliability context where one is constructing a coherent system of components, where the components are operating in an environment that causes their lifetimes to be statistically dependent. Suppose that two manufacturers are available, and we want to determine whether the coherent system assembled from the first manufacturer's components is typically longer-lived than that from the second manufacturer. The lifetime of a coherent system is a nondecreasing function of the lifetimes of its component parts. Thus one approach is to test (in a common environment) m and n sets of components from each manufacturer, respectively. The stochastic ordering properties of the resulting empirical distributions could then be examined to determine which manufacturer is preferred, if either.

For these more general multivariate distributions, the previous formulation needs to be modified. Suppose that (X_1, X_2) and (Y_1, Y_2) are bivariate random vectors with respective cdf's F and G. Let (x_{1i}, x_{2i}), $i = 1, \ldots, m$, and (y_{1j}, y_{2j}), $j = 1, \ldots, n$, be the observed random samples from F and G, respectively. Using the notions of Kiefer and Wolfowitz (1956), we want to find the generalized maximum likelihood estimators (GMLE's) of F and G, subject to the constraint that $F \overset{st}{\ge} G$. [See Miller (1983) for a description of the GMLE approach.] This problem is equivalent to finding P and Q, which maximize (2.3) subject to the stated constraints (2.2), where p_{ij}, q_{ij}, a_{ij}, and b_{ij} are defined as follows. Let I be the number of distinct values of the pooled data, $x_{11}, \ldots, x_{1m}, y_{11}, \ldots, y_{1n}$, and let J be the number of distinct values of the pooled data, $x_{21}, \ldots, x_{2m}, y_{21}, \ldots, y_{2n}$. Let $u_1 < \cdots < u_I$ and $v_1 < \cdots < v_J$ be the distinct ordered values of the first and second coordinates of the pooled observations, respectively. Then, $p_{ij} = \Pr((X_1, X_2) = (u_i, v_j))$, $q_{ij} = \Pr((Y_1, Y_2) = (u_i, v_j))$, a_{ij} is the number of (x_{1k}, x_{2k})'s equal to (u_i, v_j), and b_{ij} is the number of (y_{1k}, y_{2k})'s equal to (u_i, v_j).

2.2 One-Sample Problems

Consider the case where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from a single discrete population, with joint probability mass function (pmf) P on an $I \times J$ lattice. In this one-sample setting, we are interested in $P \overset{st}{\geq} P_0$, where P_0 is a fixed pmf. This situation, although of interest in its own right, is important in our theoretical development. The MLE of P is found by maximizing $\prod_{i,j} p_{ij}^{a_{ij}}$ subject to $P \overset{st}{\geq} P_0$, $\sum_{i,j} p_{ij} = 1$, and $p_{ij} \geq 0$, and where $A = \{a_{ij}\}$ denotes the observed contingency table.

The choice of P_0 in this situation may represent certain structural conditions on P. Consider the simple comparative distribution, the bivariate uniform; that is, $P_0(\mathbf{x}) = (IJ)^{-1}$ for all $\mathbf{x}$ in the $I \times J$ lattice. A direct argument shows that the parameter space $P \overset{st}{\geq} P_0$ is equivalent to $(\text{vol}(U))^{-1}P(U) \geq (\text{vol}(U^c))^{-1}P(U^c)$ for all $U \in \mathfrak{U}_{I\times J}$, and where $\text{vol}(U)$, the volume of the upper set U, is the number of lattice points in U (see Sampson and Whitaker 1988, sec. 5). Following Robertson and Wright (1981, p. 1248) and considering this condition in higher dimensions, we call such a condition multivariate increasing on the average.

For general bivariate distributions, F, one can consider obtaining a GMLE for $F \overset{st}{\geq} F_0$ analogous to Section 2.1.

3. THEORETICAL RESULTS

To solve the most general two-sample problem, we build our results serially. We start by using isotonic regression techniques to solve the simplest one-sample problem. Much of the section extends this solution to the general one-sample problem in such a way that numerical techniques (discussed in Sec. 4) are feasible. In turn, the solution to the general one-sample problem provides the key and an almost immediate solution to the two-sample problem. The proofs for all of the main results of this section are found in the Appendix.

3.1 One Sample: Simplest Case

Let P be a discrete p-dimensional pmf with support $\mathfrak{X}$, and let P_0 be a known pmf whose support is a subset of $\mathfrak{X}$. We use the notation $P(\mathbf{y})$ and $P(S)$ to denote both the probability of the point $\mathbf{y}$ and the set S, respectively. Based on a random sample from P, let $a(\mathbf{x})$ and $\mathbf{x} \in \mathfrak{X}$ denote the number of observations with value $\mathbf{x}$, and for $S \subseteq \mathfrak{X}$ let $a(S)$ denote the number of observations in the set S.

For this simplest one-sample case, assume that $a(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathfrak{X}$, for example, when the data are in the form of a p-dimensional ordinal contingency table with no empty cells and where P_0 is some known pmf on the same table. To find the MLE $\hat{P}$ we maximize

$$L(P) = \prod_{\mathbf{x}\in\mathfrak{X}} P(\mathbf{x})^{a(\mathbf{x})} \tag{3.1}$$

among probability distribution P satisfying

$$P(U) \geq P_0(U) \quad \text{for all } U \in \mathfrak{U}. \tag{3.2}$$

Since both P and P_0 have support on $\mathfrak{X}$, and $\mathfrak{U}_{\mathfrak{X}}$ contains only a finite number of sets, clearly (3.2) can be replaced by the finite number of constraints:

$$P(U) \geq P_0(U) \quad \text{for all } U \in \mathfrak{U}_{\mathfrak{X}}. \tag{3.3}$$

An expression for the MLE of P ($\hat{P}$) is given in the following theorem, whose proof is based on isotonic regression techniques.

Theorem 3.1. Assume that $a(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathfrak{X}$. The MLE of P subject to $P \overset{st}{\geq} P_0$ is

$$\hat{P}(\mathbf{x}) = a(\mathbf{x}) \min_L \max_U \frac{P_0(L \cap U)}{a(L \cap U)}, \tag{3.4}$$

where the minimum is taken over $\{L \in \mathfrak{L}_{\mathfrak{X}} : \mathbf{x} \in L\}$ and the maximum is taken over $\{U \in \mathfrak{U}_{\mathfrak{X}} : \mathbf{x} \in U\}$.

Note that the optimization result of Theorem 3.1 holds for arbitrary positive weights $a(\mathbf{x})$, as well as positive integer weights $a(\mathbf{x})$. This is employed to prove the generalization of Theorem 3.1 to arbitrary distributions P and P_0, given in the next subsection.

3.2 One Sample: General Case

In the general one-sample case, suppose that the random sample comes from an arbitrary distribution P on R^p, where $P \overset{st}{\geq} P_0$ and P_0 is an arbitrary but known distribution on R^p. Again, let $\mathfrak{X}$ be the set of distinct observed values of a random sample from P, and let $a(\mathbf{x})$ be the number of observations with value $\mathbf{x}$. As before, we want to find $\hat{P}^*$, which maximizes (3.1) subject to (3.2).

We reformulate this constrained optimization problem by first reducing the set of feasible P to those P that concentrate as much mass as possible on $\mathfrak{X}$ while still preserving stochastic ordering. We then replace P_0 with a discrete version, P_0', so that the original constraints, $P \overset{st}{\geq} P_0$, are reexpressed as a finite number of constraints involving P_0', similar to (3.3). Finally, we use the results of Theorem 3.1 and a limiting argument to obtain the optimal solution. To illustrate the procedure for computing the solution to the one-sample problem, throughout this section we refer to the following example.

Example 3.1. Let P_0 be the bivariate uniform distribution on $[0, 1] \times [0, 1]$, and let $\mathfrak{X} = \{(.5, .8), (.6, .6), (.8, .7)\}$.

To maximize $L(P)$ under the constraint $P \overset{st}{\geq} P_0$, we need to identify the most mass that P can assign to $\mathfrak{X}$ and still satisfy $P \overset{st}{\geq} P_0$. The first step is to construct V, the smallest lower set containing $\mathfrak{X}$. Define $Q_{\mathbf{x}} = \{\mathbf{y} : y_i \leq x_i, i = 1, \ldots, p\}$. The set V is then easily found to be the finite union of certain lower rectangles:

$$V = \bigcup_{\mathbf{x}\in\mathfrak{X}} Q_{\mathbf{x}}. \tag{3.5}$$

Example 3.2. For Example 3.1,

$$V = (-\infty, .5] \times (-\infty, .8] \cup (-\infty, .6] \times (-\infty, .6] \cup (-\infty, .8] \times (-\infty, .7].$$

The following technical lemma allows us to partition the

constraints (3.2) into two sets of constraints: those involving subsets of V and similar constraints involving only subsets of V^c. It is used to prove Lemma 3.2, which states that we can restrict attention to P that assigns all of the mass in V to $\mathcal{X}$ and where $P(V) = P_0(V)$.

Lemma 3.1. Let P and P_0 be distributions on R^p, and W be any lower set such that $P(W) = P_0(W)$. Then, $P(U) \geq P_0(U)$ $[P(L) \leq P_0(L)]$ for all $U \in \mathcal{U}$ $(L \in \mathcal{L})$ is equivalent to

$$P(U) \geq P_0(U) \; [P(L) \leq P_0(L)] \quad \text{for all } U \in \mathcal{U}_W \; (L \in \mathcal{L}_W), \tag{3.6}$$

and

$$P(U) \geq P_0(U) \; [P(L) \leq P_0(L)] \quad \text{for all } U \in \mathcal{U}_{W^c} \; (L \in \mathcal{U}_{W^c}). \tag{3.7}$$

Lemma 3.2. Each maximizer $\hat{P}^*$ of $L(P)$ over $P \overset{\text{st}}{\geq} P_0$ must satisfy $\hat{P}^*(V) = P_0(V)$ and $\hat{P}^*(\mathcal{X}) = \hat{P}^*(V)$.

Note that as long as $P(V^c) = P_0(V^c)$ and $P(U) \geq P_0(U)$ for all $U \in \mathcal{U}_{V^c}$, the exact specification of P on V^c does not affect $L(P)$. Thus a viable choice for the optimal P is to consider only those P for which $P(\cdot) = P_0(\cdot)$ in the set V^c. From (3.7) we also notice that if $P_0(V^c) = 1$, then there is a trivial solution for the optimal P, because $P(V) = 0$ necessarily for such $P \overset{\text{st}}{\geq} P_0$. Without loss of generality in the remainder of the section, we take $P_0(V^c) < 1$ [equivalently, $P_0(V) > 0$]. In view of Lemmas 3.1 and 3.2, the original problem is reduced to finding $\hat{P}^*$ that maximizes $L(P)$ subject to

$$P(U) \geq P_0(U) \quad \text{for all } U \in \mathcal{U}_V, \tag{3.8}$$

and $P(\mathcal{X}) = P(V) = P_0(V)$.

The next step is to construct a discrete version P_0' of P_0 so that (3.8) can be reexpressed as a finite number of constraints. This involves shifting P_0 mass to points on a p-dimensional grid G in V, exactly containing $\mathcal{X}$. Specifically, for $i = 1, \ldots, p$, let $x_{i,(1)} < \cdots < x_{i,(D(i))}$ denote the $D(i)$ distinct ordered values of the ith components among $\mathbf{x} \in \mathcal{X}$. For any $\mathbf{x} \in \mathcal{X}$, we write $\mathbf{x} = (x_{1,(i_1)}, \ldots, x_{p,(i_p)})$ to indicate exactly where $\mathbf{x}$ lies in the $D(1) \times \cdots \times D(p)$ lattice constructed from $\mathcal{X}$. Then, $G = V \cap (X_{i=1}^p \{x_{i,(1)}, \ldots, x_{i,(D(i))}\})$. For P whose mass on V is concentrated on $\mathcal{X}$, $P(U) = P(U \cap G)$ for all $U \in \mathcal{U}_V$. It is also clear that $P(U) = P(W)$ for any $W \in \mathcal{U}_V$ such that $U \cap G = W \cap G$, from which it is easily seen that (3.8) is equivalent to

$$P(U) \geq \sup_W P_0(W) \quad \text{for all } U \in \mathcal{U}_G, \tag{3.9}$$

where the supremum is taken over all $\{W : W \in \mathcal{U}_V \text{ and } W \cap G = U\}$. For $U \in \mathcal{U}_G$, denote the right side of (3.9) by $P_0'(U)$.

The set W that optimizes the right side of (3.9) is the largest upper set in $\mathcal{U}_V$ that does not contain $G - U$, that is, $W = V - \bigcup_{\mathbf{x} \in G-U} \{Q_{\mathbf{x}}\}$. The set function P_0' shifts the mass of P_0 in the appropriate hyper-rectangle of G to the upper right-hand corner of the hyper-rectangle. Thus we can restate the optimization problem as follows: Find $\hat{P}^*$ that maximizes $L(P)$ subject to $P(U) \geq P_0'(U)$ for all $U \in \mathcal{U}_G$. Further, from the construction of P_0' it is straightforward to prove that $P_0'(\cdot)/P_0(V)$ is indeed a pmf on the set G.

Lemma 3.3. Let $P_0'(U)$ be defined by the right side of (3.9) for all $U \in \mathcal{U}_G$. Then, there exists a pmf $\tilde{P}_0$ on G such that $\tilde{P}_0(U) = P_0'(U)/P_0(V)$ for all $U \in \mathcal{U}_G$. Moreover, $\tilde{P}_0(\mathbf{x}) = [P_0(V)]^{-1} P_0(X_{j=1}^p (x_{j,(i_j-1)}, x_{j,(i_j)}))$ for $\mathbf{x} \in G$ and is 0 otherwise, where $\mathbf{x} = (x_{1(i_1)}, \ldots, x_{p(i_p)})$ and $x_{j(0)} = -\infty$, for $j = 1, \ldots, p$.

Example 3.3. Continuing Example 3.2, we let $G = \{(.5, .6), (.5, .7), (.5, .8), (.6, .6), (.6, .7), (.8, .6), (.8, .7)\}$. The grid points and the P_0 probability content of the appropriate hyper-rectangles are given in Figure 1. These determine $\tilde{P}_0(\mathbf{x})$ for $\mathbf{x} \in G$'s; for example, $\tilde{P}_0(.5, .6) = .30/.61$, $\tilde{P}_0(.8, .7) = .02/.61$, and so forth.

We have reduced the general one-sample problem to

$$\max \prod_{\mathbf{x} \in G} P(\mathbf{x})^{a(\mathbf{x})}, \tag{3.10}$$

subject to $P(U) \geq \tilde{P}_0(U)$ for $U \in \mathcal{U}_G$ and $P(G) = 1$. Note that we have reparameterized by dividing by $P_0(V)$, and thus the solution to (3.10) gives the conditional probability $\hat{P}^*(\cdot)/P_0(V)$ in the set V.

Theorem 3.1 cannot be applied directly to solve (3.10), because $a(\mathbf{x})$ may be 0 for some $\mathbf{x} \in G$. Nevertheless, with a simple limiting argument Theorem 3.1 can still be used to find the P that solves (3.10). For $\varepsilon > 0$, define $a_\varepsilon(\mathbf{x}) = a(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, and $a_\varepsilon(\mathbf{x}) = \varepsilon$ for $\mathbf{x} \in G - \mathcal{X}$. By Theorem 3.1, we can find $\hat{P}_\varepsilon$ that maximizes $\prod_{\mathbf{x} \in G} P(\mathbf{x})^{a_\varepsilon(\mathbf{x})}$ subject to $P(U) \geq \tilde{P}_0(U)$, for all $U \in \mathcal{U}_G$, and $P(G) = 1$. Further,

$$\hat{P}_\varepsilon(\mathbf{x}) = a_\varepsilon(\mathbf{x}) \min_L \max_U \frac{\tilde{P}_0(L \cap U)}{a_\varepsilon(L \cap U)}, \tag{3.11}$$

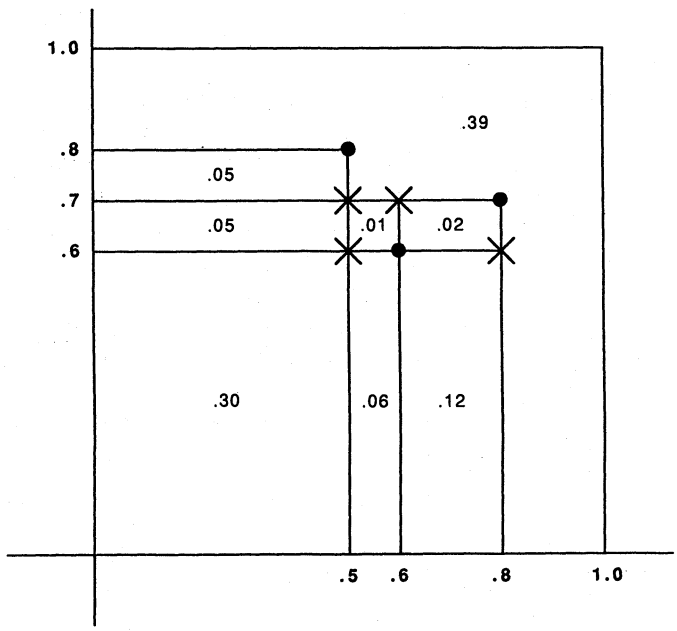

Figure 1. Hyper-Rectangles Used to Calculate P_0' and $\tilde{P}_0$ in Example 3.3: • indicates $\mathbf{x} \in \mathcal{X}$; × indicates $\mathbf{x} \in G - \mathcal{X}$.

where the minimum is taken over the set $\{L \in \mathcal{L}_G : \mathbf{x} \in L\}$ and the maximum is taken over $\{U \in \mathcal{U}_G : \mathbf{x} \in U\}$. By evaluating $\lim_{\varepsilon \to 0} \hat{P}_\varepsilon(\mathbf{x})$, we get the explicit solution to (3.10) that yields $\hat{P}^*$, given in the following theorem.

Theorem 3.2. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from P, and let $a(\mathbf{x})$ be the number of observations with value $\mathbf{x}$. A GMLE of P subject to $P \overset{\text{st}}{\geq} P_0$ is given by $\hat{P}^*$, where $\hat{P}^* = P_0$ on V^c, with V given by (3.5); $\hat{P}^*(\mathbf{x}) = 0$ for $\mathbf{x} \in V - \mathcal{X}$, and

$$\hat{P}^*(\mathbf{x}) = a(\mathbf{x}) \min_L \max_U \frac{P_0'(L \cap U)}{a(L \cap U)} \quad \text{for all } \mathbf{x} \in \mathcal{X}, \quad (3.12)$$

where the minimum is taken over the set $\{L \in \mathcal{L}_G : \mathbf{x} \in L\}$ and the maximum is taken over $\{U \in \mathcal{U}_G : \mathbf{x} \in U\}$.

Finding the GMLE for the case $P \overset{\text{st}}{\geq} P_0$ follows the same general argument leading to Theorem 3.2. Let $K = \cup\{L \in \mathcal{L} : L \cap \mathcal{X} = \phi\}$ and $H = K^c \cap (\bigcup_{i=1}^p \{x_{i,(1)}, \ldots, x_{i,(D(i))}\})$. For all $U \in \mathcal{U}_H$, define $Q_0(U) = \inf P_0(W)$, where the infimum is taken over $\{W \in \mathcal{U}_{K^c} : W \cap H = U\}$, and let $\tilde{Q}_0$ be the probability distribution determined by $\tilde{Q}_0(U) = Q_0(U)/P_0(K^c)$ for all $U \in \mathcal{U}_H$, where Q_0 is constructed from P_0 analogously to the construction of P_0'. Then, applying the reasoning that yields Theorem 3.2, we obtain the following corresponding theorem.

Theorem 3.3. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from P, and let $a(\mathbf{x})$ be the number of observations with the value $\mathbf{x}$. A GMLE of P subject to $P \overset{\text{st}}{\geq} P_0$ is given by $\hat{P}^*(\mathbf{x})$, where $\hat{P}^* = P_0$ on K, $\hat{P}^*(\mathbf{x}) = 0$ for $\mathbf{x} \in K^c - \mathcal{X}$, and $\hat{P}^*(\mathbf{x}) = a(\mathbf{x}) \min_U \max_L \{Q_0(L \cap U)/a(L \cap U)\}$ for all $\mathbf{x} \in \mathcal{X}$, where the minimum is taken over the set $\{U \in \mathcal{U}_H : \mathbf{x} \in U\}$ and the maximum is taken over $\{L \in \mathcal{L}_H : \mathbf{x} \in L\}$.

3.3 The Two-Sample Case

We approach the general two-sample problem by recasting it as two one-sample optimization problems, and then apply the results of Section 3.2. This approach has been used successfully in the univariate setting; for example, see Barlow and Brunk [1972, eq. (5.8)] and Dykstra (1982).

Following the notation of Section 3.2, we use $\mathcal{X}$ to denote the distinct values of the pooled samples from P and Q, and we let $a(\mathbf{x})$ and $b(\mathbf{x})$ be the number of observations sampled with the value $\mathbf{x}$ from P and Q, respectively. The appropriate function to be maximized for this two-sample problem is

$$L(P, Q) = \prod_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x})^{a(\mathbf{x})} \prod_{\mathbf{x} \in \mathcal{X}} Q(\mathbf{x})^{b(\mathbf{x})}, \quad (3.13)$$

among probability distributions $P \overset{\text{st}}{\geq} Q$.

By fixing Q, applying Lemma 3.2, fixing P, and applying the analog of Lemma 3.2 when the stochastic ordering is reversed, we can show that the constraints $P \overset{\text{st}}{\geq} Q$ can be replaced by $P(\mathcal{X}) = Q(\mathcal{X}) = 1$ and $P(U) \geq Q(U)$ for all $U \in \mathcal{U}_\mathcal{X}$.

The solution to (3.13) subject to the constraint $P \overset{\text{st}}{\geq} Q$ must also satisfy the multivariate analog of Barlow and Brunk [1972, eq. (5.8)]:

$$P(U) \geq \frac{a(U) + b(U)}{a(\mathcal{X}) + b(\mathcal{X})} \geq Q(U) \quad \text{for all } U \in \mathcal{U}_\mathcal{X},$$

where $(a(\cdot) + b(\cdot))/(a(\mathcal{X}) + b(\mathcal{X}))$ is the pooled estimator of P and Q under the assumption that $P = Q$. Finally, the two-sample problem can be reformulated as the following one-sample problems: max $\prod_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x})^{a(\mathbf{x})}$ subject to $P(U) \geq a(U) + b(U)/(a(\mathcal{X}) + b(\mathcal{X}))$ for all $U \in \mathcal{U}_\mathcal{X}$, and max $\prod_{\mathbf{x} \in \mathcal{X}} Q(\mathbf{x})^{b(\mathbf{x})}$ subject to $a(U) + b(U)/(a(\mathcal{X}) + b(\mathcal{X})) \geq Q(U)$ for all $U \in \mathcal{U}_\mathcal{X}$. The solutions are given by Theorems 3.2 and 3.3, respectively.

4. COMPUTATION AND NUMERICAL EXAMPLES

Our aim is to demonstrate that $\hat{P}$ in (3.4) and $\hat{P}^*$ in (3.12) can be computed using existing isotonic regression algorithms. First, consider the simplest one-sample problem discussed in Section 3.1. Direct computation of $\hat{P}$ from (3.4) for any reasonable-sized data set is impractical, because of the large numbers of sets in $\mathcal{U}_\mathcal{X}$ and $\mathcal{L}_\mathcal{X}$ (see Sampson and Whitaker 1988). Nevertheless, viewing $\hat{P}(\mathbf{x})/a(\mathbf{x})$ as the least squares isotonic regression of $P_0(\mathbf{x})/a(\mathbf{x})$ with weights $a(\mathbf{x})$ and with respect to the component-wise partial order on $\mathcal{X}$, we could employ the minimum lower sets algorithms described by Barlow et al. (1972, sec. 2.3) or the algorithm of Gebhardt (1970). But as pointed out by Dykstra and Robertson (1982), any of these algorithms either take an extremely large amount of computer time or involve intricate branching logic that is difficult to program.

Dykstra and Robertson (1982) proposed an efficient iterative algorithm for approximating the isotonic regression on finite subsets of R^p with respect to component-wise partial ordering (see also Dykstra 1983). Bril, Dykstra, Pillers, and Robertson (1984) provided FORTRAN code implementing the Dykstra and Robertson algorithm, for when $\mathcal{X}$ is a two-dimensional lattice. If $\mathcal{X}$ is a subset of a two-dimensional lattice G, this program approximates the desired optimization problem by defining $a(\mathbf{x}) = 10^{-5}$ for $\mathbf{x} \in G - \mathcal{X}$, and then proceeds to a solution. To handle our p-dimensional situations, the authors have developed a FORTRAN 77 program (available for distribution) implementing the Dykstra and Robertson algorithm.

The general one-sample problem presents much the same difficulties. In addition, this problem cannot be recast as a standard isotonic regression problem because, as discussed in Section 3.2, $a(\mathbf{x}) = 0$ does not necessarily imply $P_0(\mathbf{x}) = 0$. We now illustrate the use of (3.11) to solve this problem by letting $a(\mathbf{x}) = \varepsilon$ for all $x \in G - \mathcal{X}$ and then letting ε become small. The approximating $\hat{P}_\varepsilon$'s are obtained using existing isotonic regression algorithms.

Example 4.1. Let $\mathcal{X} = \{(.1, .1), (.1, .4), (.3, .1), (.4, .5), (.7, .4), (.7, .7), (.9, .4), (1.0, .7), (1, 1)\}$ be a simple hypothetical sample assumed to be drawn for a single bivariate population. Based on this data we calculate the GMLE $\hat{P}^*$ of P under the assumption that $P \overset{\text{st}}{\geq} P_0$, where P_0 is the bivariate uniform distribution on $[0, 1] \times [0, 1]$.

The estimate $\hat{P}^*$ is approximated by $\hat{P}_\varepsilon$ for several small values of ε [note that $P_0(V) = 1$, and thus $\hat{P}^* = \lim_{\varepsilon \to 0} \hat{P}_\varepsilon^*$]. In Table 1 we present $P_0'(\mathbf{x})$ and $a(\mathbf{x})$ for all $\mathbf{x} \in G$. $\hat{P}_\varepsilon(\mathbf{x})/a_\varepsilon(\mathbf{x})$ is the isotonic regression of $P_0'(x)/a_\varepsilon(\mathbf{x})$ with the weights $a_\varepsilon(\mathbf{x})$, and can be approximated using the Dykstra and Robertson algorithm. Table 1 also gives $\hat{P}_\varepsilon$ for $\varepsilon = .1, .05, .001$, and .0001, and the GMLE $\hat{P}^*$. Even though the hypothetical sample was chosen so that the unrestricted GMLE would be far from multivariately increasing on the average, it took a very short time (on an IBM PC/AT clone) to actually compute each of the $\hat{P}_\varepsilon$'s.

A possible alternative approach to computing $\hat{P}^*$ for the general one-sample case can be seen from the following easily provable lemma.

Lemma 4.1. Let $\rho_\varepsilon(\mathbf{x}) = P_\varepsilon(\mathbf{x})/a_\varepsilon(\mathbf{x})$ for all $\mathbf{x} \in G$. Let $\hat{\rho}^*(\mathbf{x}) = \lim \hat{\rho}_\varepsilon(\mathbf{x})$ as $\varepsilon \to 0$, where $\hat{\rho}_\varepsilon(\mathbf{x})$ is the isotonic regression of $\rho_\varepsilon(\mathbf{x})$, with weights $a_\varepsilon(\mathbf{x})$ and with respect to the component-wise partial ordering. Then, there exists ε_0 such that for all $0 < \varepsilon < \varepsilon_0$, the sets where $\hat{\rho}_\varepsilon$ is constant are the same as those where $\hat{\rho}^*$ is constant.

Lemma 4.1 states that for each one-sample problem there is a sufficiently small ε_0, so for $\varepsilon < \varepsilon_0$ the level sets (see Barlow et al. 1972, def. 2.5) for $\hat{\rho}_\varepsilon(\mathbf{x})$ are the same as those for $\hat{\rho}^*(\mathbf{x})$. Thus if ε_0 could be identified, $\hat{\rho}^*$ can be computed exactly from the level sets of $\hat{\rho}_\varepsilon$. For example, if A is a level set of $\hat{\rho}_\varepsilon$ $(\varepsilon < \varepsilon_0)$, then $\hat{P}^*(\mathbf{x}) = a(\mathbf{x})P_0'(A)/a(A)$ for $\mathbf{x} \in A$. It is interesting to note that in Example 4.1 the level sets for $\hat{P}_\varepsilon/a_\varepsilon$ are the same for $\varepsilon = .05, .001$, and .0001, and can be used to compute $\hat{P}^*$.

To illustrate the applicability of our approach to a two-sample problem, we estimate the underlying distributions P and Q, respectively, for the British and Danish social-mobility data, assuming (as suggested by the data) that $P \overset{st}{\geq} Q$. Table 2 gives both of these MLE's, and the usual (unrestricted) MLE's. Because neither the British nor the Danish contingency table has empty cells, $\hat{P}^*$ and $\hat{Q}^*$ are found by a single application of Dykstra and Robertson's (1982) approximation algorithm.

5. DISCUSSION

Since we obtained the MLE's under stochastic ordering, we can construct the likelihood ratio test (LRT) of $H_0 : F = G$ against $H_1 : F \overset{st}{\geq} G$. Analogous to the argument of Robertson and Wright (1981) and Grove (1980), it can be shown that the LRT statistic has a $\bar{\chi}^2$ distribution asymptotically, which is a mixture of χ^2 distributions. Not only is the mixing distribution extremely difficult to compute under H_0, the distribution itself depends on the specific values of $F = G$. In the univariate case, Grove obtained

Table 1. $a(\mathbf{x})$, $P_0'(\mathbf{x})$, $\hat{P}_\varepsilon(\mathbf{x})$ for $\varepsilon = .1, .05, .001, .0001$, and $\hat{P}^*(\mathbf{x})$ for $\mathbf{x}$ in $\mathcal{X}$, From Example 4.1

$\mathbf{x}$	$a(\mathbf{x})$	$P_0'(\mathbf{x})$	$\hat{P}_{.1}$	$\hat{P}_{.05}$	$\hat{P}_{.001}$	$\hat{P}_{.0001}$	$\hat{P}^*$
(.1, .1)	1	.01	.0100	.0100	.0100	.0100	.01
(.1, .3)	0	.02	.0027	.0014	.0000	.0000	.00
(.1, .4)	1	.01	.0273	.0286	.0300	.0300	.03
(.1, .5)	0	.01	.0054	.0030	.0001	.0000	.00
(.1, .7)	0	.02	.0100	.0060	.0001	.0000	.00
(.1, 1.0)	0	.03	.0200	.0120	.0003	.0000	.00
(.3, .1)	1	.02	.0200	.0200	.0200	.0200	.02
(.3, .3)	0	.04	.0054	.0030	.0001	.0000	.00
(.3, .4)	0	.02	.0054	.0030	.0001	.0000	.00
(.3, .5)	0	.02	.0054	.0030	.0001	.0000	.00
(.3, .7)	0	.04	.0100	.0060	.0001	.0000	.00
(.3, 1.0)	0	.06	.0200	.0120	.0003	.0000	.00
(.4, .1)	0	.01	.0054	.0030	.0001	.0000	.00
(.4, .3)	1	.02	.0538	.0609	.0699	.0700	.07
(.4, .4)	0	.01	.0054	.0030	.0001	.0000	.00
(.4, .5)	1	.01	.0538	.0609	.0697	.0700	.07
(.4, .7)	0	.02	.0100	.0060	.0001	.0000	.00
(.4, 1.0)	0	.03	.0200	.0120	.0003	.0000	.00
(.5, .1)	0	.03	.0083	.0045	.0001	.0000	.00
(.5, .3)	0	.06	.0083	.0045	.0001	.0000	.00
(.5, .4)	1	.03	.0833	.0909	.0998	.1046	.10
(.5, .5)	0	.03	.0100	.0060	.0001	.0000	.00
(.5, .7)	1	.06	.1000	.1200	.1493	.1499	.15
(.5, 1.0)	0	.09	.0200	.0120	.0003	.0000	.00
(.7, .1)	0	.02	.0083	.0045	.0001	.0000	.00
(.7, .3)	0	.04	.0083	.0045	.0001	.0000	.00
(.7, .4)	1	.02	.0833	.0909	.0998	.0954	.10
(.7, .5)	0	.02	.0100	.0060	.0001	.0000	.00
(.7, .7)	0	.04	.0100	.0060	.0001	.0000	.00
(.7, 1.0)	0	.06	.0200	.0120	.0003	.0000	.00
(1.0, .1)	0	.01	.0100	.0060	.0001	.0000	.00
(1.0, .3)	0	.02	.0100	.0060	.0001	.0000	.00
(1.0, .4)	0	.01	.0100	.0060	.0001	.0000	.00
(1.0, .5)	0	.01	.0100	.0060	.0001	.0000	.00
(1.0, .7)	1	.02	.1000	.1200	.1493	.1499	.15
(1.0, 1.0)	1	.03	.2000	.2400	.2985	.2999	.30

Table 2. Social-Mobility Data: British Restricted (Unrestricted) MLE's and Danish Restricted (Unrestricted) MLE's

Father's status	Son's status				
	1	2	3	4	5
1	.012 (.014) .012 (.008)	.011 (.013) .010 (.007)	.002 (.002) .007 (.007)	.005 (.005) .002 (.002)	.002 (.002) .001 (.008)
2	.008 (.008) .011 (.010)	.048 (.050) .047 (.044)	.024 (.024) .047 (.046)	.043 (.044) .026 (.025)	.015 (.016) .009 (.009)
3	.003 (.003) .010 (.010)	.021 (.022) .038 (.035)	.032 (.031) .118 (.121)	.065 (.064) .089 (.091)	.028 (.027) .039 (.040)
4	.004 (.004) .004 (.003)	.041 (.043) .022 (.021)	.054 (.053) .071 (.073)	.208 (.204) .142 (.146)	.130 (.128) .081 (.083)
5	.001 (.001) .003 (.003)	.011 (.012) .004 (.003)	.021 (.021) .028 (.029)	.093 (.091) .082 (.084)	.120 (.117) .100 (.103)

an approximation that Robertson and Wright discussed further. In this multivariate case, however, it appears that obtaining such approximations is impractical.

To obtain the MLE's of $F_1 \overset{st}{\geq} F_2 \overset{st}{\geq} \cdots \overset{st}{\geq} F_k$ for the multiple-sample multivariate problem, an iterative algorithm can be developed based on the Feltz and Dykstra (1985) multiple-sample univariate solution.

Finally, we note that further study of the numerical aspects of the algorithms discussed here is important, especially to carry out effective simulations.

APPENDIX: PROOFS

Theorem 3.1. Let $\rho(\mathbf{x}) = P(\mathbf{x})/a(\mathbf{x})$ and $\rho_0(\mathbf{x}) = P_0(\mathbf{x})/a(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$. Reparameterizing the objective function (3.1) and the constraints (3.3), we recast the optimization problem as finding the solution to

$$\min - \sum_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x})\ln(\rho(\mathbf{x})), \tag{A.1}$$

subject to the constraints

$$\sum_{\mathbf{x}\in U} a(\mathbf{x})\rho(\mathbf{x}) \geq \sum_{\mathbf{x}\in U} a(\mathbf{x})\rho_0(\mathbf{x}) \quad \text{for all } U \in \mathcal{U}_{\mathcal{X}}, \tag{A.2}$$

$$\sum_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x})\rho(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x})\rho_0(\mathbf{x}) = 1, \tag{A.3}$$

and

$$\rho(\mathbf{x}) \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{A.4}$$

Let K be the class of all nondecreasing functions $f(\mathbf{x})$ on $\mathcal{X}$, and define its dual cone: $K^{a*} = \{g(\mathbf{x}), \mathbf{x} \in \mathcal{X} : \sum_{\mathbf{x}\in\mathcal{X}} a(\mathbf{x})g(\mathbf{x})f(\mathbf{x}) \leq 0, \text{ for all } f \in K\}$. The constraints (A.2) and (A.3) imply that $(\rho_0 - \rho) \in K^{a*}$. Barlow and Brunk (1972, theorem 3.4) showed that the unique minimum of (A.1), subject to $(\rho_0 - \rho) \in K^{a*}$, is

$$\hat{\rho}(\mathbf{x}) = \min_L \max_U \frac{\sum_{\mathbf{v}\in L\cap U} a(\mathbf{v})\rho_0(\mathbf{v})}{\sum_{\mathbf{v}\in L\cap U} a(\mathbf{v})} \geq 0, \tag{A.5}$$

where the minimum is taken over $\{L \in \mathcal{L}_{\mathcal{X}} : \mathbf{x} \in L\}$ and the maximum is taken over $\{U \in \mathcal{U}_{\mathcal{X}} : \mathbf{x} \in U\}$. Furthermore, Barlow et al. (1972, theorem 1.4) showed that $\hat{\rho}(\mathbf{x})$ [defined by (A.5)] also satisfies (A.2) and (A.3), so $\hat{P}(\mathbf{x}) = a(\mathbf{x})\hat{\rho}(\mathbf{x})$, and the result follows.

Lemma 3.1. Let $P(W) = P_0(W)$. Suppose that P and P_0 satisfy (3.6) and (3.7). Then, if $U \in \mathcal{U}$, $P(U) = P(U \cap W) + P(U \cap W^c) \geq P_0(U \cap W) + P_0(U \cap W^c) = P_0(U)$; hence, $P(U) \geq P_0(U)$ for all $U \in \mathcal{U}$. Conversely, if $P(U) \geq P_0(U)$ for all $U \in \mathcal{U}$, then (3.7) holds, because $U \cap W^c \in \mathcal{U}$. To show (3.6), note that

$$\begin{aligned} P(U \cap W) &= P(W) - P(U^c \cap W) \\ &= P_0(W) - 1 + P(U \cup W^c) \\ &\geq P_0(V) - 1 + P_0(U \cup W^c) \\ &= P_0(U \cap W). \end{aligned}$$

The result for lower sets follows similarly.

Lemma 3.2. To show $\hat{P}^*(\mathcal{X}) = \hat{P}^*(V)$, begin by defining $\mathcal{X}^* \subseteq \mathcal{X}$ as the upper extreme points of $\mathcal{X}$, that is, $\mathcal{X}^* = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} \geq \mathbf{x}, \mathbf{y} \neq \mathbf{x} \text{ implies } \mathbf{y} \in V^c\}$. Note that $V = \cup_{\mathbf{x}\in\mathcal{X}^*} Q_{\mathbf{x}}$. Suppose that $\hat{P}^*(V) > \hat{P}^*(\mathcal{X})$. Then, there exists $\mathbf{x}^* \in \mathcal{X}^*$ such that $\hat{P}^*(Q_{\mathbf{x}^*}) - \hat{P}^*(Q_{\mathbf{x}^*} \cap \mathcal{X}) > 0$. Construct a pmf Π by shifting the excess mass $\hat{P}^*(Q_{\mathbf{x}^*}) - \hat{P}^*(Q_{\mathbf{x}^*} \cap \mathcal{X})$ onto $\mathbf{x}^*$; that is, define $\Pi = \hat{P}^*$ on $Q_{\mathbf{x}^*}^c$, $\Pi(\mathbf{x}^*) = \hat{P}^*(\mathbf{x}^*) + \hat{P}^*(Q_{\mathbf{x}^*}) - \hat{P}^*(Q_{\mathbf{x}^*} \cap \mathcal{X})$ and $\Pi(\mathbf{x}) = \hat{P}^*(\mathbf{x})$ for $\mathbf{x} \in (Q_{\mathbf{x}^*} - \{\mathbf{x}^*\}) \cap \mathcal{X}$, and $\Pi = 0$ otherwise. Then, $L(\Pi) > L(\hat{P}^*)$ and $\Pi \overset{st}{\geq} \hat{P}^* \overset{st}{\geq} P_0$, yielding a contradiction.

The proof that $\hat{P}^*(V) = P_0(V)$ uses two basic steps that allow us to add mass to the extreme points of $\mathcal{X}$ and yet not violate the stochastic ordering constraint.

Step 1. Suppose that M is a subprobability measure satisfying $M(L) \leq P_0(L)$ for all lower sets $L \subseteq V$. We observe that if for every $\mathbf{x}^* \in \mathcal{X}^*$ there exists a lower set $L_{\mathbf{x}^*}$ with $Q_{\mathbf{x}^*} \subseteq L_{\mathbf{x}^*} \subseteq V$ and $M(L_{\mathbf{x}^*}) = P_0(L_{\mathbf{x}^*})$, then $M(V) = P_0(V)$. To see this, write $\mathcal{X}^*$ as $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*\}$, and note that $P_0(V) \geq M(V) = M(\cup_{i=1}^k L_{\mathbf{x}_i^*}) = \sum_{i=1}^k M(L_{\mathbf{x}_i^*}) - \sum_{i=1}^{k-1} M((\cup_{j=i+1}^k L_{\mathbf{x}_j^*}) \cap L_{\mathbf{x}_i^*}) \geq P_0(V)$.

Step 2. We now suppose that $\hat{P}^*(V) < P_0(V)$; Step 1 shows that mass can again be shifted to obtain a contradiction.

Step 1 (with $M = \hat{P}^*$) guarantees the existence of a $Q_{\mathbf{x}_i^*}$ such that there exists a lower set L, $Q_{\mathbf{x}_i^*} \subseteq L \subseteq V$, such that $P_0(L) - \hat{P}^*(L) > 0$. Without loss of generality, call such a set $Q_{\mathbf{x}_1^*}$, and let

$$a_1 = \min_{\{L\in\mathcal{U}:\, Q_{\mathbf{x}_1^*}\subseteq L\subseteq V\}} (P_0(L) - \hat{P}^*(L)). \tag{A.6}$$

Because $\hat{P}^*(X) = \hat{P}^*(V)$, the right side of (A.6) can be found by taking the minimum over the finite number of lower sets L in $\{\cap_{\mathbf{x}\in A} Q_{\mathbf{x}} : \mathbf{x}_1^* \in A \subseteq \mathcal{X}\}$; thus $a_1 > 0$. Now, define the subprobability measure Π_1', with full mass on $\mathcal{X}$, by $\Pi_1'(\mathbf{x}) = \hat{P}^*(\mathbf{x})$, if $\mathbf{x} \in \mathcal{X} - \{\mathbf{x}_1^*\}$, and $\Pi_1'(\mathbf{x}_1^*) = \hat{P}^*(\mathbf{x}_1^*) + a_1$. It then follows that $\Pi_1'(L) \leq P_0(L)$ for all lower sets $L \subseteq V$, and that there exists a lower set L_1 such that $Q_{\mathbf{x}_1^*} \subseteq L_1 \subseteq V$ and $\Pi_1'(L_1) = P_0(L_1)$.

Let $a_2 = \min(P_0(L) - \Pi_1'(L))$, where the minimum is taken over lower sets $Q_{\mathbf{x}_i^*} \subseteq L \subseteq V$. If there does exist a lower set L

with $Q_{\mathbf{x}_2^*} \subseteq L \subseteq V$ such that $\Pi_1'(L) = P_0(L)$, then $a_2 = 0$, and we define $\Pi_2' = \Pi_1'$. Otherwise, $a_2 > 0$; we define Π_2' by $\Pi_2'(\mathbf{x}_2^*) = \Pi_1'(\mathbf{x}_2^*)$ and $\Pi_2' = \Pi_1'$ otherwise. In either case ($a_2 = 0$ or $a_2 > 0$), observe that $\Pi_2'(L) \le P_0(L)$ for all lower sets $L \subseteq V$ and that there exists lower sets L_1 and L_2, where $Q_{\mathbf{x}_1^*} \subseteq L_1 \subseteq V$ and $Q_{\mathbf{x}_2^*} \subseteq L_2 \subseteq V$ such that $\Pi_2'(L_1) = P_0(L_1)$ and $\Pi_2'(L_2) = P_0(L_2)$. We repeat this procedure until all k extreme points have been examined. By Step 1, the resulting Π_k' satisfies $\Pi_k'(V) = P_0(V)$ and $\Pi_k'(L) \le P_0(L)$ for all lower sets $L \subseteq V$. Now, define $\tilde{\Pi} = \Pi_k'$ on V and $\tilde{\Pi} = P_0$ on V^c. Note that $\tilde{\Pi}$ is a pmf, $L(\tilde{\Pi}) > L(\hat{P}^*)$, and by Lemma 3.1 $\tilde{\Pi} \overset{\text{st}}{\geq} P_0$, which contradicts the fact that $\hat{P}^*$ is the MLE.

Theorem 3.2. For $\mathbf{x} \in \mathcal{X}$, let $\hat{P}^*$ be given by (3.12), $\hat{P}^*(\mathbf{x}) = 0$ for $\mathbf{x} \in V - \mathcal{X}$, and $\hat{P}^* = P_0$ for the set V^c. It is first shown that $\hat{P}_\varepsilon(\mathbf{x}) \to \hat{P}^*(\mathbf{x})/P_0(V)$ as $\varepsilon \to 0$ for all $\mathbf{x} \in G$. [Note that the right side of (3.12) is well defined, because $\mathbf{x} \in L \cap U$ and $\mathbf{x} \in \mathcal{X}$ imply that $a(L \cap U) \ge a(\mathbf{x}) > 0$.] It is clear that $\hat{P}_\varepsilon(x) \to \hat{P}^*(\mathbf{x})/P_0(V)$ as $\varepsilon \to 0$ for all $\mathbf{x} \in \mathcal{X}$. To see that $\hat{P}_\varepsilon(\mathbf{x}) \to 0$ as $\varepsilon \to 0$ for $\mathbf{x} \in G - \mathcal{X}$, we show that $\lim_{\varepsilon \to 0} \max_U \min_L [\tilde{P}_0(L \cap U)/a_\varepsilon(L \cap U)] < \infty$, where the maximum is understood to be taken over $\{U \in \mathcal{U}_G : \mathbf{x} \in U\}$; similarly, the minimum is understood to be over $\{L \in \mathcal{L}_G : \mathbf{x} \in L\}$. Note that for convenience max and min have been interchanged (see Barlow et al. 1972, theorem 2.8). Because $G \in \mathcal{L}_G$,

$$\max_U \min_L \frac{\tilde{P}_0(L \cap U)}{a_\varepsilon(L \cap U)} \le \max_U \frac{\tilde{P}_0(U \cap G)}{a_\varepsilon(U \cap G)} = \max_U \frac{\tilde{P}_0(U)}{a_\varepsilon(U)}. \tag{A.7}$$

The limit as $\varepsilon \to 0$ of the right side of (A.7) is $\max_U \tilde{P}_0(U)/a(U)$, which is finite, because by construction of V and G, $a(U) > 0$ for all $U \in \mathcal{U}_G$. Therefore, $\hat{P}_\varepsilon(\mathbf{x}) \to 0$ as $\varepsilon \to 0$ for $\mathbf{x} \in G - \mathcal{X}$.

Next, we verify that $\hat{P}^* \overset{\text{st}}{\geq} P_0$. Because $P_0(V)\hat{P}_\varepsilon \to \hat{P}^*$ for $\mathbf{x} \in G$, we have $\hat{P}^*(U) \ge \tilde{P}_0(U)P_0(V)$ for all $U \in \mathcal{U}_G$, and $\hat{P}^*(G) = \tilde{P}_0(G)P_0(V)$. This implies that $\hat{P}^*(U) \ge P_0(U)$ for all $U \in \mathcal{U}_V$, and $\hat{P}^*(V) = P_0(V)$. This, along with $\hat{P}^* = P_0$ on V^c, implies that $\hat{P}^* \overset{\text{st}}{\geq} P_0$, by Lemma 3.1.

It remains to show that $\hat{P}^*$ maximizes $L(P)$ among $P \overset{\text{st}}{\geq} P_0$. Suppose that $L(Q) > L(\hat{P}^*)$ and $Q \overset{\text{st}}{\geq} P_0$. From the discussion in Section 3.2 and the definition of $\hat{P}_\varepsilon$, it is clear that for $\varepsilon > 0$

$$\prod_{\mathbf{x} \in G} Q(\mathbf{x})^{a_\varepsilon(\mathbf{x})} \le \prod_{\mathbf{x} \in G} (\hat{P}_\varepsilon(\mathbf{x})P_0(V))^{a_\varepsilon(\mathbf{x})}. \tag{A.8}$$

Take the limit as $\varepsilon \to 0$ of both sides of (A.8) to yield $L(Q) \le L(\hat{P}^*)$, which contradicts the fact that $L(Q) > L(\hat{P}^*)$.

[*Received January 1987. Revised June 1988.*]

REFERENCES

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference Under Order Restrictions,* New York: John Wiley.

Barlow, R. E., and Brunk, H. D. (1972), "The Isotonic Regression Problem and Its Dual," *Journal of the American Statistical Association,* 67, 140–147.

Bishop, Y. M., Fienberg, S. W., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice,* Cambridge, MA: MIT Press.

Bril, G., Dykstra, R., Pillers, C., and Robertson, T. (1984), "Isotonic Regression in Two Independent Variables," *Applied Statistics,* 33, 352–357.

Brunk, H. D., Franck, W. E., Hanson, D. L., and Hogg, R. V. (1966), "Maximum Likelihood Estimation of the Distribution of Two Stochastically Ordered Random Variables," *Journal of the American Statistical Association,* 61, 1067–1080.

Dykstra, R. L. (1982), "Maximum Likelihood Estimation of the Survival Functions of Stochastically Ordered Random Variables," *Journal of the American Statistical Association,* 77, 621–628.

——— (1983), "An Algorithm for Restricted Least Squares Regression," *Journal of the American Statistical Association,* 78, 837–842.

Dykstra, R. L., Madsen, R. W., and Fairbanks, K. (1983), "A Nonparametric Likelihood Ratio Test," *Journal of Statistical Computation and Simulation,* 18, 247–264.

Dykstra, R. L., and Robertson, T. (1982), "An Algorithm for Isotonic Regression for Two or More Independent Variables," *The Annals of Statistics,* 10, 708–716.

——— (1983), "On Testing Monotone Tendencies," *Journal of the American Statistical Association,* 78, 342–350.

Feltz, C. J., and Dykstra, R. L. (1985), "Maximum Likelihood Estimation of the Survival Functions of N Stochastically Ordered Random Variables," *Journal of the American Statistical Association,* 80, 1012–1019.

Franck, W. E. (1984), "A Likelihood Ratio Test for Stochastic Ordering," *Journal of the American Statistical Association,* 79, 686–691.

Gebhardt, F. (1970), "An Algorithm for Monotone Regression With One or More Independent Variables," *Biometrika,* 57, 263–271.

Grove, D. M. (1980), "A Test of Independence Against a Class of Ordered Alternatives in a 2 × C Contingency Table," *Journal of the American Statistical Association,* 75, 454–459.

Hettmansperger, T. P. (1984), *Statistical Inference Based on Ranks,* New York: John Wiley.

Kiefer, J., and Wolfowitz, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics,* 27, 887–906.

Lee, C. C. (1987), "Maximum Likelihood Estimates for Stochastically Ordered Multinomial Populations With Fixed and Random Zeros," in *Foundations of Statistical Inference,* Toronto: D. Reidel, pp. 189–197.

Marshall, A., and Olkin, I. (1979), *Inequalities: Theory of Majorization and Its Applications,* New York: Academic Press.

Miller, R. (1983), *Survival Analysis,* New York: John Wiley.

——— (1985), *Beyond ANOVA: Basic Applied Statistics,* New York: John Wiley.

Robertson, T., and Wegman, E. (1978), "Likelihood Ratio Tests for Order Restrictions in Exponential Families," *The Annals of Statistics,* 6, 485–505.

Robertson, T., and Wright, F. T. (1974), "On the Maximum Likelihood Estimation of Stochastically Ordered Random Variates," *The Annals of Statistics,* 2, 528–534.

——— (1981), "Likelihood Ratio Tests for and Against Stochastic Ordering Between Multinomial Populations," *The Annals of Statistics,* 9, 1248–1257.

——— (1982), "On Measuring the Conformity of a Parameter Set to a Trend, With Applications," *The Annals of Statistics,* 10, 1234–1245.

Sampson, A. R., and Whitaker, L. R. (1988), "Positive Dependence, Upper Sets, and Multidimensional Partitions," *Mathematics of Operations Research,* 13, 254–264.