ON UNBIASED ESTIMATION FOR RANDOMIZED RESPONSE MODELS

By

Pranab Kumar Sen

Department of Biostatistics
University of North Carolina, Chapel Hill, N. C.

Institute of Statistics Mimeo Series No. 867

APRIL 1973

ON UNBIASED ESTIMATION FOR RANDOMIZED RESPONSE MODELS[1]

PRANAB KUMAR SEN

University of North Carolina, Chapel Hill

ABSTRACT

For quantitative randomized response models, optimal unbiased estimation of regular functionals of distribution functions is considered. In this context, distribution theory of Hoeffding's U-statistics and von Mises' differentiable statistical functions is extended to randomized response models. Estimation of the basic distributions is also considered.

1. INTRODUCTION

In practical surveys, particularly, involving sensitive questions, the randomized response technique of reducing respondent bias has been found to be quite effective. Since the randomization affords protection to the respondent in answering the question without revealing his personal situation, potential embarrassments and stigma have been removed, and therefore, the primary reason for either a refusal or an evasive answer does not exist. An extensive amount of work in this area has been carried out by the North Carolina group; we may refer to Greenberg et al (1970, 1971) where other references are also cited.

In a quantitive randomized response model utilizing two questions, the respondent selects at random one of the two questions in such a way that the interviewer does not know which question is being answered. However, the probabilities of selecting the two questions are set beforehand. Thus, the response distribution is a mixture of two basic distributions where the mixing

---

1) Work sponsored by the Aerospace Research Laboratories, Air Force Systems Command Contract F33615-71-C-1927. Reproduction in whole or in part permitted for any purpose of the U.S. Government.

coefficients are known. Mostly, the current literature deals with the situation where the functional forms of the basic distributions are assumed to be specified, and one is interested in the set of parameters (algebraic constants) associated with these distributions. For such problems, standard statistical procedures for mixture of distributions are usually adapted without much problem. For a broad class of problems, one may have very little knowledge on the forms of the underlying distributions, and it may be more reasonable to assume that the basic distributions belong to some broad class of distributions. In this framework, estimable parameters are defined as suitable functionals of the distribution functions. Our first objective is to sketch this formulation of estimable parameters for randomized response models and to provide unbiased estimators of these parameters. In this context, the theory of unbiased estimation of regular functionals of distribution functions, studied in detail by Halmos (1946), Hoeffding (1948) and others, is extended here to randomized response models. These developments along with the distribution theory of the related von Mises' (1947) differentiable statistical functions are treated in Section 3. The theory is illustrated with the aid of some examples.

The characterization of randomized response distributions as mixtures of two basic distribution provides clue for the estimation of the latter in terms of the former, in a reasonably simple manner [See Section 4]. These estimates are consistent and unbiased. On the other hand, unlike the usual estimates of distribution functions, these are (i) not necessarily non-decreasing everywhere, and (ii) do not necessarily assume values in the closed interval [0,1]. These drawbacks call for certain modifications which are considered in Section 4. Certain confidence bounds for the basic distributions are also provided.

The case of more than two questions in the basic model is treated briefly in the last section.

2. PRELIMINARY NOTIONS

Consider a randomized response model where a respondent selects at random either of the two questions A and B, where A is usually sensitive. Suppose, we have two mutually independent and non-overlapping random samples of sizes n_1 and n_2. In the ith sample, a respondent selects the two questions A and B with respective probabilities p_i and q_i $(=1-p_i)$, for i=1,2. The response is assumed to be quantitative in nature, and the distribution functions (df) of the response (assumed to be a stochastic k-vector for some $k\geq 1$) for the two questions A and B are denoted by $F_1(x)$ and $F_2(x)$, respectively, which are both defined on the common Euclidean space R^k. The random variables associated with the n_i observations in the ith sample are denoted by $X_{i1},\ldots,X_{in_i}$, so that these are independent and identically distributed (iid) with a df denoted by $G_i(x)$, for i=1,2. Then, we have

$$G_i(x) = p_iF_1(x) + q_iF_2(x),\ x\in R^k,\ i=1,2. \tag{2·1}$$

We assume that $0<p_2<p_1<1$, so that $\begin{vmatrix} p_1 & q_1 \\ p_2 & q_2 \end{vmatrix} = p_1-p_2>0$. For $p_1=p_2$, $G_1\equiv G_2$, so that F_1,F_2 are not expressible in terms of G_1,G_2. Equation (2.1) leads us to

$$F_1(x)=\{q_2G_1(x)-q_1G_2(x)\}/(p_1-p_2) \text{ and } F_2(x)=\{p_1G_2(x)-p_2G_1(x)\}/(p_1-p_2). \tag{2·2}$$

The last equation leads us to estimate F_1,F_2 as well as their parameters from $X_{11},\ldots,X_{2n_2}$. We shall deal with this problem in the next two sections. In passing, we may remark that whereas G_1 and G_2 are convex combination of F_1 and F_2, the converse is not true [as the coefficients in (2.2) are not all non-negative]. This creates certain problems which will be discussed in Section 4.

3. UNBIASED ESTIMATION OF PARAMETERS OF F_1, F_2.

We assume that both F_1 and F_2 belong to a family of df's, $\mathcal{F}$, which contains all convex combination of F_1, F_2, that is, if $F_1 \in \mathcal{F}$, $F_2 \in \mathcal{F}$, then

$$\alpha F_1 + (1-\alpha)F_2 \in \mathcal{F} \text{ for every } 0 \leq \alpha \leq 1. \tag{3·1}$$

For example, if $\mathcal{F}$ is the family of all continuous df's or distributions with finite moments up to the pth order for some $p>0$, then (3·1) holds. Consider a functional $\theta(F)$ of F defined on $\mathcal{F}$, such that for a sample $Z_1,\ldots,Z_n$ of size n from the df F, there exists a statistic $\phi(Z_1,\ldots,Z_n)$ for which

$$\theta(F) = E\phi(Z_1,\ldots,Z_n), \ \forall \ F \ \in \mathcal{F}. \tag{3·2}$$

Following Hoeffding (1948), $\theta(F)$ is then said to be regular over $\mathcal{F}$. Let $m(\geq 1)$ be the smallest n for which (3·2) holds. Then, $\phi(Z_1,\ldots Z_m)$ is called the kernel of $\theta(F)$ and m, the degree of $\theta(F)$ over $\mathcal{F}$. Without any loss of generality, we may assume that $\phi(Z_1,\ldots,Z_m)$ is symmetric in its m arguments, so that by (3·2),

$$\theta(F) = \int_{E^m}\cdots\int \phi(z_1,\ldots,z_m)dF(z_1)\ldots dF(z_m), \ \forall \ F \in \mathcal{F} \tag{3·3}$$

In the randomized response model, $\{Z_i, i\geq 1\}$ are not observable. Based on the observable random variables $\{X_{ij}, 1\leq j\leq n_i, i=1,2\}$, one is interested in estimating $\theta(F_1)$ and $\theta(F_2)$, defined by (3·3) with $F=F_1$ and F_2. We shall consider the case of $\theta(F_1)$ only, as the other case follows on parallel lines.

THEOREM 3·1. If (3·1) holds and $\theta(F)$ is regular over $\mathcal{F}$ with degree m (≥ 1), then for the randomized response model, $\theta(F_1)$ is estimable for all $n_1 \geq m, n_2 \geq m$.

Proof. We have to show that under (3·1) and (3·3), for every $F_1, F_2 \in \mathcal{F}$, there exists a kernel $\psi(X_{11},\ldots,X_{1m},X_{21},\ldots,X_{2m})$ which unbiasedly estimates $\theta(F_1)$. Let us denote by

$$\theta_s(F_1,F_2) = \int_{R^m}\cdots\int \phi(z_1,\ldots,z_m)\prod_{j=1}^{m-s}dF_1(z_j)\prod_{\ell=m-s+1}^{m} dF_2(z_\ell), \tag{3·4}$$

for $0\leq s\leq m$, where for s=0 or m, one of the product terms equals to one. Since, by (3·1) and (3·3), $F_1, F_2 \in \mathcal{F} \Rightarrow G_1, G_2 \in \mathcal{F}$, and

$$\theta(\alpha F_1+(1-\alpha)F_2) \text{ exists for every } 0\leq\alpha\leq 1,\ F_1,F_2\in\mathcal{F}, \tag{3·5}$$

we conclude that $\theta_s(F_1,F_2)$ exists for all $0\leq s\leq m$ and $F_1,F_2\in\mathcal{F}$. Hence

$$\theta_s(G_1,G_2) \text{ exists for every } 0\leq s\leq m,\ F_1,F_2\in\mathcal{F}. \tag{3·6}$$

Then, by using (2·2) and (3·3), we have

$$\theta(F_1) = \int\cdots\int_{E^{km}}\phi(z_1,\ldots,z_m)\prod_{j=1}^{m}\{(p_1-p_2)^{-1}\{q_2dG_1(z_j)-q_1dG_2(z_j)\}$$

$$= (p_1-p_2)^{-m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s q_2^{m-s}q_1^s\int\cdots\int_{E^{km}}\phi(z_1,\ldots,z_m)\prod_{j=1}^{m-s}dG_1(z_j)\prod_{\ell=m-s+1}^{m}dG_2(z_\ell)$$

$$= (p_1-p_2)^{-m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s q_2^{m-s}q_1^s\theta_s(G_1,G_2). \tag{3·7}$$

From the well-known results on generalized U-statistics [viz., Puri and Sen (1971, Section 3·2)], it follows that $\theta_s(G_1,G_2)$ is estimable, and

$$E\phi(X_{11},\ldots,X_{1m-s},\ X_{21},\ldots,X_{2s}) = \theta_s(G_1,G_2), \tag{3·8}$$

for every $0\leq s\leq m$ and $G_1,G_2\in\mathcal{G}$ (i.e., $F_1\in\mathcal{F},F_2\in\mathcal{F}$). Thus, if we let

$$\psi(X_{11},\ldots,X_{1m},X_{21},\ldots,X_{2m}) =$$

$$(p_1-p_2)^{-m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s q_1^s q_2^{m-s}\phi(X_{11},\ldots,X_{1m-s},X_{21},\ldots,X_{2s}), \tag{3.9}$$

it follows from (3·7), (3·8) and (3·9) that

$$E\,\psi(X_{11},\ldots,X_{1m},X_{21},\ldots,X_{2m}) = \theta(F_1) \text{ for every } G_1,G_2\in\mathcal{G}. \tag{3·10}$$

Hence the proof of the theorem is complete.

The generalized U-statistic corresponding to $\theta_s(G_1,G_2)$ is

$$U_s(n_1,n_2) = \binom{n_1}{m-s}^{-1}\binom{n_2}{s}^{-1}\Sigma_s^*\phi(X_{1j_1},\ldots,X_{1i_{m-s}},X_{2j_1},\ldots,X_{2j_s}), \tag{3·11}$$

where the summation Σ_s^* extends over all possible $1\leq i_1<\ldots<i_{m-s}\leq n_1$ and $1\leq j_1<\ldots<j_s\leq n_2$, for $s=0,1,\ldots,m$; (for $s=0$ or m, one of the two sets is null). As an unbiased estimator of $\theta_s(G_1,G_2)$, $U_s(n_1,n_2)$ possesses certain optimal properties. In particular, $U_s(n_1,n_2)$ is symmetric in $X_{11},\ldots,X_{1n_1}$ as well as in $X_{21},\ldots,X_{2n_2}$, and hence, a function of the two sample order statistics. Thus, if these order statistics are complete, then $U_s(n_1,n_2)$ is the minimum

variance unbiased (MVU) estimator of $\theta_s(G_1,G_2)$, for each s(=0,1,...,m). Also, jointly, the vector $[U_o(n_1,n_2),\ldots,U_m(n_1,n_2)]$ is the minimum concentration ellipsoid unbiased estimator of $[\theta_o(G_1,G_2),\ldots,\theta_m(G_1,G_2)]$, so that if we define

$$U(n_1,n_2) = (p_1-p_2)^{-m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s q_1^s q_2^{m-s} U_s(n_1,n_2), \tag{3·12}$$

it follows from (3·9) through (3·12) that under the completeness of the two sample order statistics [viz., Fraser (1953)], $U(n_1,n_2)$ is the MVU estimator of $\theta(F_1)$.

Let us denote the two empirical df's by

$$G_{n_i}(x) = n_i^{-1}\sum_{j=1}^{n_i} u(x-X_{ij}),\ x\varepsilon R^k,\ \text{for } i=1,2, \tag{3·13}$$

where u(t) is 1 iff all the k components of t are non-negative, and 0, otherwise. Then, following von Mises (1947), we define a (generalized) differentiable statistical function for $\theta_s(G_1,G_2)$ by

$$V_s(n_1,n_2) = \int\cdots\int_{E^{km}}\phi(z_1,\ldots,z_m)\prod_{j=1}^{m-s} dG_{n_1}(z_j)\prod_{\ell=m-s+1}^{m} dG_{n_2}(z_\ell)$$

$$= n_1^{-(m-s)} n_2^{-s}\sum_{i_1=1}^{n_1}\cdots\sum_{i_{m-s}=1}^{n_1}\sum_{j_1=1}^{n_2}\cdots\sum_{j_s=1}^{n_2}\phi(X_{1i_1},\ldots,X_{1i_{m-s}},X_{2j_1},\ldots,X_{2j_s}), \tag{3·14}$$

for s = 0,1,...,m. then, an alternative estimator of $\theta(F_1)$ is

$$V(n_1n_2) = (p_1-p_2)^{-m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s q_1^s q_2^{m-s} V_s(n_1,n_2) \tag{3·15}$$

Whereas $U(n_1,n_2)$ is an unbiased estimator of $\theta(F_1)$, $V(n_1,n_2)$ is not, in general, a strictly unbiased estimator. By virtue of (2·2), we may estimate the df F_1 by using the G_{n_i} in (3·13) by

$$\hat{F}_1(x;n_1,n_2) = \{q_2 G_{n_1}(x) - q_1 G_{n_2}(x)\}/(p_1-p_2), \tag{3·16}$$

which unbiasedly estimates $F_1(x)$ for every $x\varepsilon E^k$ and $n_1,n_2\geq 1$. Then, $V(n_1,n_2)$ can also be written as

$$\int\cdots\int_{E^{km}}\phi(z_1,\ldots,z_m)d\hat{F}_1(z_1;n_1,n_2)\ldots d\hat{F}_1(z_m;n_1,n_2) \tag{3·17}$$

The expression (3·17) corresponds to the form in von Mises (1947), where the empirical df is based on the basic random variables. Using the results in Puri and Sen (1971, pp.64-66) for the individual $U_s(n_1,n_2) - V_s(n_1,n_2)$, $0\leq s\leq m$, we obtain by (3·12) and (3·15), that if $n_1/(n_1+n_2)$ is bounded away from zero and one, when $n_1+n_2=n \longrightarrow \infty$, then

$$n^{\frac{1}{2}}[U(n_1,n_2)-V(n_1,n_2)] \longrightarrow 0, \text{ in probability} \qquad (3\cdot18)$$

Thus, for large n_1,n_2, the two estimators $U(n_1,n_2)$ and $V(n_1,n_2)$ share the common properties. As such, we shall not discuss the case of $V(n_1,n_2)$.

Let us now denote by

$$\zeta_{cd}(s,s';F_1,F_2) = \mathrm{Cov}\{\phi(X_{1i_1},\ldots,X_{1i_{m-s}},X_{2j_1},\ldots,X_{2j_s}), \qquad (3\cdot19)$$

$$\phi(X_{1i'_1},\ldots,X_{1i'_{m-s'}},X_{2j'_1},\ldots,X_{2j'_{s'}})\},$$

when $(i_1,\ldots,i_{m-s})$ and $(i'_1,\ldots,i'_{m-s'})$ have exactly c (≥ 0) indices in common, and $(j_1,\ldots,j_s)$ and $(j'_1,\ldots j'_{s'})$ have d (≥ 0) in common, for $0\leq c\leq (m-s,m-s')$, $0\leq d\leq (s,s')$ and $0\leq s,s'\leq m$. Then, it is well-known [cf. Puri and Sen (1971, Section 3·2)] that for $n_1,n_2\geq m$,

$$\mathrm{Cov}[U_s(n_1,n_2),U_{s'}(n_1,n_2)]$$

$$= \binom{n_1}{m-s}^{-1}\binom{n_2}{s}^{-1}\sum_{c=0}^{m-s}\sum_{d=0}^{s}\binom{m-s'}{c}\binom{s'}{d}\binom{n_1-m+s'}{m-s-c}\binom{n_2-s'}{s-d}\zeta_{cd}(s,s';F_1,F_2)$$

$$= \frac{(m-s)(m-s')}{n_1}\zeta_{10}(s,s';F_1,F_2) + \frac{ss'}{n_2}\zeta_{01}(s,s';F_1,F_2) + O(n^{-2}). \qquad (3\cdot20)$$

Using than (3·12) and (3·20), we have

$$\mathrm{Var}[U(n_1,n_2)] = (p_1-p_2)^{-2m}\sum_{s=0}^{m}\sum_{s'=0}^{m}\binom{m}{s}\binom{m}{s'}(-1)^{s+s'}q_1^{s+s'}\cdot$$

$$q_2^{2m-s-s'}\mathrm{Cov}[U_s(n_1,n_2),U_{s'}(n_1,n_2)]$$

$$= (p_1-p_2)^{-2m}m^2\sum_{s=0}^{m}\sum_{s'=0}^{m}(-1)^{s+s'}q_1^{s+s'}q_2^{2m-s-s'}\left\{\binom{m-1}{s}\binom{m-1}{s'}\frac{1}{n_1}\zeta_{10}(s_1s';F_1,F_2)\right.$$

$$+ \binom{m-1}{s-1}\binom{m-1}{s'-1}\frac{1}{n_2}\zeta_{01}(s,s';F_1,F_2)\Big\} + O(n^{-2}). \tag{3·21}$$

Moreover, from the well-known results on the asymptotic normality of generalized U-statistics [viz., Puri and Sen (1971, Section 3·2)], it follows that asymptotically $n^{\frac{1}{2}}[\{U_o(n_1,n_2)-\theta_o(G_1,G_2)\},\ldots,\{U_m(n_1,n_2)-\theta_m(G_1,G_2)\}]$ has a (m+1)-variate normal distribution with null mean vector and dispersion matrix whose elements are n times those in (3·20). Thus, if we set

$$\lim_{n\to\infty} n_1/n = \lambda : 0<\lambda<1, \tag{3·22}$$

we conclude that asymptotically (as $n\longrightarrow \infty$),

$$\mathcal{L}(n^{\frac{1}{2}}[U(n_1,n_2)-\theta(F_1)]) \longrightarrow \mathcal{N}(0,\gamma^2), \tag{3·23}$$

where by (3·21) and (3·22),

$$\gamma^2 = m^2(p_1-p_2)^{-2m}\sum_{s=0}^{m}\sum_{s'=0}^{m}(-1)^{s+s'}q_1^{s+s'}q_2^{2m-s-s'}\Big\{\frac{1}{\lambda}\binom{m-1}{s}\binom{m-1}{s'}\zeta_{10}(s,s';F_1,F_2)$$
$$+ \frac{1}{1-\lambda}\binom{m-1}{s-1}\binom{m-1}{s'-1}\zeta_{01}(s,s';F_1,F_2)\Big\}. \tag{3·24}$$

In the above development, it is assumed that

$$\max_{0\le s\le m} E\phi^2(X_{11},\ldots,X_{1s},X_{21},\ldots,X_{2m-s}) < \infty \text{ and } \gamma^2 > 0. \tag{3·25}$$

We may remark that all the $\zeta_{cd}(s_1,s_2;F_1,F_2)$, $0\le s_1,s_2\le m$, are regular functionals of G_1 and G_2 ($\varepsilon\mathfrak{F}$), and hence, are estimable. Sen (1960) has obtained some simple estimates of these functionals for the conventional one and two sample problems. His estimators remain good for the randomized response model too. So, substituting these estimates in (3·24), one gets an estimator of γ^2, which we denote by $\hat{\gamma}^2_{n_1n_2}$. Thus, noting that $\hat{\gamma}_{n_1n_2}\longrightarrow\gamma$, in probability as $n_1,n_2\longrightarrow\infty$ and (3·23) holds, one obtains by using the well-known Slutzky theorem that as $n_1,n_2\longrightarrow\infty$,

$$\mathcal{L}(n^{\frac{1}{2}}[U(n_1,n_2)-\theta(F_1)]/\hat{\gamma}_{n_1n_2}) \longrightarrow \mathcal{N}(0,1). \tag{3·26}$$

The last result is useful in providing a large sample test or confidence bound for $\theta(F_1)$.

We consider now some illustrative examples. Let $\mathcal{F} = \{F: \int_{-\infty}^{\infty} x^2 dF(x) < \infty\}$, so that $F_1, F_2 \in \mathcal{F}$ implies that (3·1) holds. We desire to estimate

$$\sigma_1^2 = \sigma^2(F_1) = \int_{-\infty}^{\infty} x^2 dF_1(x) - \left(\int_{-\infty}^{\infty} x dF_1(x)\right)^2, \tag{3·27}$$

when $\{X_{ij}, 1 \le j \le n_i, i=1,2\}$ are observed. If we let $\phi(z_1,z_2) = \frac{1}{2}(z_1-z_2)^2$, then, in (3·3), $\theta(F) = \sigma^2(F)$, so that $\sigma^2(F_1)$ is estimable and m=2. Then, by (3·4), $\theta(G_1,G_1) = \sigma^2(G_1)$, $\theta(G_2,G_2) = \sigma^2(G_2)$ and $\theta(G_1,G_2) = \frac{1}{2}[\sigma^2(G_1) + \sigma^2(G_2) + \{\mu(G_1)-\mu(G_2)\}^2]$, where $\mu(G_i) = \int_{-\infty}^{\infty} x dG_i(x)$, i=1,2. Hence, by (3·11), we obtain that $U_o(n_1,n_2) = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{ij}-\bar{X}_i)^2$, $U_2(n_1,n_2) = \frac{1}{n_2-1}\sum_{j=1}^{n_2}[X_{2j}-\bar{X}_2]^2$ and $U_1(n_1,n_2) = \frac{1}{n_1 n_2}\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2}[X_{ij}-X_{2j'}]^2$, where $\bar{X}_i = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$, i=1,2. Consequently, by (3·12),

$$U(n_1,n_2) = (p_1-p_2)^{-2}\{q_2^2 U_o(n_1,n_2) - 2q_1q_2U_1(n_1,n_2) + q_1^2U_2(n_1,n_2)\} \tag{3·28}$$

is the MVU estimator of $\sigma^2(F_1)$. Secondly, consider the case of bivariate df's, and define

$$\theta(F) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x,y)dF(x,y), \quad F \in \mathcal{F}, \text{ all continuous df's.} \tag{3·29}$$

The sample measures for (3·29), known as the Kendall tau, can be easily computed by using the kernel $\phi(\underset{\sim}{z}_1, \underset{\sim}{z}_2) = 1$ or 0 according as $\underset{\sim}{z}_1 \le \underset{\sim}{z}_2$ or not (where $\underset{\sim}{a} \le \underset{\sim}{b}$ means $a_j \le b_j$, j=1,2). These are $U_o(n_1,n_2) = \binom{n_1}{2}^{-1}\sum_{1\le i<j\le n_1}\phi(X_{1i},X_{ij})$, $U_1(n_1,n_2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\phi(X_{1i},X_{2j})$ and $U(n_1,n_2) = \binom{n_2}{2}^{-1}\sum_{1\le i<j\le n_2}\phi(X_{2i},X_{2j})$. So that

$$U(n_1,n_2) = (p_1-p_2)^{-2}\{q_2^2U_o(n_1,n_2) - 2q_1q_2U_1(n_1,n_2) + q_1^2U_2(n_1,n_2)\} \tag{3·30}$$

is the MVU estimator of $\theta(F_1)$. In either of these examples, $V(n_1,n_2)$, defined by (3·15), will be different from $U(n_1,n_2)$, and is biased.

4. ESTIMATION OF THE BASIC DISTRIBUTIONS

In many practical situations, a complete knowledge of the distributions F_1 and F_2 is deemed to perform a more detailed study of the data. For example, in the North Carolina Abortion study [1], Greenberg et al have estimated the

average number of abortions per female in the child bearing age group, stratified by various socio-economic factors. Instead of estimating these averages, one may be more interested in comparing the distributions of the number of abortions per woman over the different strata. This may reveal the proportion of women having at least one abortion as well as similar other characteristics of the distributions. In order to increase the scope of our inference procedures, in such a case, we may be interested in providing distribution-free estimates of F_1 and F_2 which do not require any stringent assumption on the form of F_1 and F_2.

It is known that the empirical df's G_{n_1} and G_{n_2}, defined by (3·13), are unbiased estimates of G_1 and G_2, and they are sufficient statistics. Thus, as in (3·16), we have the estimates of F_1 and F_2 as

$$\hat{F}_1(x;n_1,n_2) = \frac{q_2 G_{n_1}(x) - q_1 G_{n_2}(x)}{p_1 - p_2}, \quad \hat{F}_2(x;n_1,n_2) = \frac{p_1 G_{n_2}(x) - p_2 G_{n_1}(x)}{p_1 - p_2}, \tag{4·1}$$

for $x \in R^k$ and $n_1, n_2 \geq 1$. Since G_{n_i} unbiased estimates G_i, $i=1,2$, by (4·1) and (2·2)

$$E\hat{F}_i(x;n_1,n_2) = F_i(x), \quad x \in R^k, \quad i=1,2. \tag{4·2}$$

Also, by the Glivenko-Cantelli theorem, as $n_i \longrightarrow \infty$,

$$\sup_x \left| G_{n_i}(x) - G_i(x) \right| \longrightarrow 0 \text{ almost surely, } i=1,2. \tag{4·3}$$

Thus, by (4·1), (4·3) and (2·2),

$$\sup_x \left| \hat{F}_i(x;n_1,n_2) - F_i(x) \right| \longrightarrow 0 \text{ almost surely, } i=1,2 \tag{4·4}$$

Consequently, the derived empirical df's $\hat{F}_1$ and $\hat{F}_2$ are unbiased and strongly consistent estimates of F_1 and F_2, respectively. On the other hand, by (4·1), $\hat{F}_1$ and $\hat{F}_2$ are not convex combination of G_{n_1} and G_{n_2}. This introduces the following undesirable properties of $\hat{F}_1$ and $\hat{F}_2$. First, $\hat{F}_1$ and $\hat{F}_2$ are <u>not</u> non-decreasing everywhere. Second, $\hat{F}_1$ and $\hat{F}_2$ can assume negative values, and

third, they can also assume values greater than one. To illustrate these, we consider the case of k=1 (i.e., X_{ij} real valued) and denote the ordered random variables of the ith sample by $X_{i,1} \leq \ldots \leq X_{i,n_i}$ for i=1,2. Also let $X_{i,0} = -\infty$ and $X_{i,n_i+1} = +\infty$, for i=1,2. Then, by (3·13),

$$G_{n_i}(x) = (j-1)/n_i \text{ for } x_{i,j-1} \leq x < X_{i,j}, 1 \leq j \leq n_i+1, i=1,2. \tag{4·5}$$

Thus, by looking at (4·1) and (4·5), we obtain that

$$[X_{1,1} > X_{2,1}] \Longrightarrow \hat{F}_1(x;n_1,n_2) < 0, \ \forall\, X_{2,1} \leq x < X_{1,1}, \tag{4·6}$$

$$[X_{1,1} < X_{2,1}] \Longrightarrow \hat{F}_2(x;n_1,n_2) < 0, \ \forall\, X_{1,1} \leq x < X_{2,1}, \tag{4·7}$$

so that $\hat{F}_1, \hat{F}_2$ can be negative. Similarly,

$$[X_{1,n_1} < X_{2,n_2}] \Longrightarrow \hat{F}_1(x;n_1,n_2) > 1, \ \forall\, X_{1,n_1} \leq x < X_{2,n_2} \tag{4·8}$$

$$[X_{2,n_2} < X_{1,n_1}] \Longrightarrow \hat{F}_2(x;n_1,n_2) > 1, \ \forall\, X_{2,n_2} \leq x < X_{1,n_1} \tag{4·9}$$

so that $\hat{F}_1, \hat{F}_2$ can be greater than one. Again, by (4·1) and (4·5),

$$d\hat{F}_1(x;n_1,n_2) = \begin{cases} q_2/n_1(p_1-p_2), & \text{if } x=X_{1,i}, 1 \leq i \leq n_1, \\ -q_1/n_2(p_1-p_2), & \text{if } x=X_{2,i}, 1 \leq i \leq n_2, \\ 0, & \text{otherwise;} \end{cases} \tag{4·10}$$

$$d\hat{F}_2(x;n_1,n_2) = \begin{cases} p_1/n_2(p_1-p_2), & \text{if } x=X_{2,i}, 1 \leq i \leq n_2, \\ -p_2/n_1(p_1-p_2), & \text{if } x=X_{1,i}, 1 \leq i \leq n_1, \\ 0, & \text{otherwise;} \end{cases} \tag{4·11}$$

so that $\hat{F}_1$ and $\hat{F}_2$ can not be non-decreasing everywhere. Because of these undesirable properties, $\hat{F}_1$ and $\hat{F}_2$ can not be regarded as distribution functions, and hence, we need to consider some alternative estimators.

Definition. An estimator of $F_i(x)$ having the fundamental properties that (i) it is non-decreasing in x everywhere, (ii) it lies in the closed interval [0,1] and (iii) it tends to 0 or + 1 according as $x \longrightarrow -\infty$ or $+\infty$, is termed

a characteristic preserving estimate.

We shall consider here characteristic preserving estimate of F_1 and F_2. First, for simplicity of presentation, we consider the case of real valued random variables for which both G_1 and G_2 are continuous. As before, the order statistics of the ith sample are denoted by $X_{i,o}(=-\infty)<X_{i,1}<\ldots<X_{i,n_i}<X_{i,n_i+1}(=\infty)$, i=1,2, where by virtue of the continuity of G_1,G_2, ties can be neglected, in probability. Since, by (4·10)-(4·11), $\hat{dF}_i$ is positive only at the n_i order statistics of the ith sample, for i=1,2, we propose our estimates of F_1,F_2, using these n_1+n_2 points. For this, let

$$\hat{F}_{1,j} = [q_2(j/n_1)-q_1G_{n_2}(X_{1,j})]/(p_1-p_2), j=1,\ldots,n_1, \hat{F}_{1,0}=0; \qquad (4\cdot12)$$

$$\hat{F}_{2,j} = [p_1(j/n_2)-p_2G_{n_1}(X_{2,j})]/(p_1-p_2), j=1,\ldots,n_2, \hat{F}_{2,0}=0; \qquad (4\cdot13)$$

where G_{n_1} and G_{n_2} are defined by (3·13). As mentioned earlier, $\hat{F}_{i,j}$ is not necessarily $\nearrow$ in j $(0\leq j\leq n_i)$, i=1,2. Our proposed estimators $\tilde{F}_i(x;n_1,n_2)$, i=1,2, are then defined by

$$\tilde{F}_i(x;n_1,n_2) = \tilde{F}_i(X_{i,j}) \text{ for } X_{i,j}\leq x< X_{i,j+1}, 0\leq j\leq n_i, i=1,2, \qquad (4\cdot14)$$

where

$$\tilde{F}_1(x_{1,0}) = 0 = \tilde{F}_2(x_{2,0}), \ \tilde{F}_1(X_{1,n_1}) = \tilde{F}_2(X_{2,n_2}) = 1; \qquad (4\cdot15)$$

$$\tilde{F}_i(X_{i,j}) = \begin{cases} \tilde{F}_i(X_{i,j-1}), & \text{if } \hat{F}_{i,j}\leq\hat{F}_{i,j-1}, \\ \hat{F}_{i,j}, & \text{if } \tilde{F}_{i,j-1}<\hat{F}_{i,j}<1, \\ 1, & \text{if } \hat{F}_{i,j}\geq 1, \end{cases} \qquad (4\cdot16)$$

for $1\leq j\leq n_i$, i=1,2.

Now, by (4·14)-(4·16), $\tilde{F}_i(x)$ is non-decreasing in x, lies in the closed interval [0,1], and attains the lower and upper bounds for $x<X_{i,1}$ and $x\geq X_{i,n_i}$, respectively, for i=1,2. Thus, $\tilde{F}_1$ and $\tilde{F}_2$ are characteristic preserving estimates of F_1 and F_2. Note that (4·16) can be written as

$$\tilde{F}_i(X_{i,j}) = \min\{\max_{0\le k\le j} \hat{F}_{i,k}, 1\}, 0\le j\le n_i, i=1,2. \qquad (4\cdot17)$$

Further, note that for $X_{1,j}\le x<X_{1,j+1}$, $G_{n_1}(x)$ remains stationary, whereas $G_{n_2}(x)$ may increase. Hence, by (4·1), $\hat{F}_1(x;n_1,n_2)$ is non-increasing for $X_{1,j}\le x<X_{1,j+1}$, $j=0,\ldots,n_1$. Similarly, $\hat{F}_2(x;n_1,n_2)$ is non-increasing for $X_{2,j}\le x<X_{2,j+1}$, $j=0,\ldots,n_2$. Consequently,

$$\max_{0\le k\le j} \hat{F}_{i,k} = \sup\{\hat{F}_i(x;n_1,n_2) : x\le X_{i,j}\}, \qquad (4\cdot18)$$

for $0\le j\le n_i$, $i=1,2$. Thus, by (4·17) and (4·18),

$$\tilde{F}_i(X_{i,j}) = \min[\sup\{\hat{F}_i(x;n_1,n_2) : x\le X_{i,j}\}, 1], \qquad (4\cdot19)$$

and by (4·14), (4·19) and a few standard steps, we get that

$$\tilde{F}_i(x) = \min[\sup\{\hat{F}_i(y;n_1,n_2) : y\le x\}, 1], -\infty<x<\infty. \qquad (4\cdot20)$$

[Note that $\hat{F}_i(y;n_1,n_2)=0$, $\forall\, y<\min(X_{1,1},X_{2,1})$, so that $\tilde{F}_i(x)\ge 0$.] The last definition is quite flexible and it readily suits the cases where G_1 and G_2 (or F_1 and F_2) are multi-dimensional distributions or are not necessarily continuous everywhere. Thus, for general $k(\ge 1)$ variate distributions $F_1(x)$, $F_2(x), x\in R^k$ on denoting by $a\le b$, the coordinatewise inequalities, our proposed estimates are

$$\tilde{F}_i(x;n_1,n_2) = \min[\sup\{\hat{F}_i(y;n_1,n_2) : y\le x\}, 1], i=1,2. \qquad (4\cdot21)$$

where the df's $\hat{F}_1$ and $\hat{F}_2$ are defined by (3·16) and (4·2). Actually, these are obtained by Smoothing $\hat{F}_1$ and $\hat{F}_2$. If we let $n=n_1+n_2$, $\lambda_n=n_1/n$, and assume that there exists a λ_o : $0<\lambda_o\le\frac{1}{2}$, such that for all n,

$$\lambda_o\le\lambda_n\le 1-\lambda_o, \qquad (4\cdot22)$$

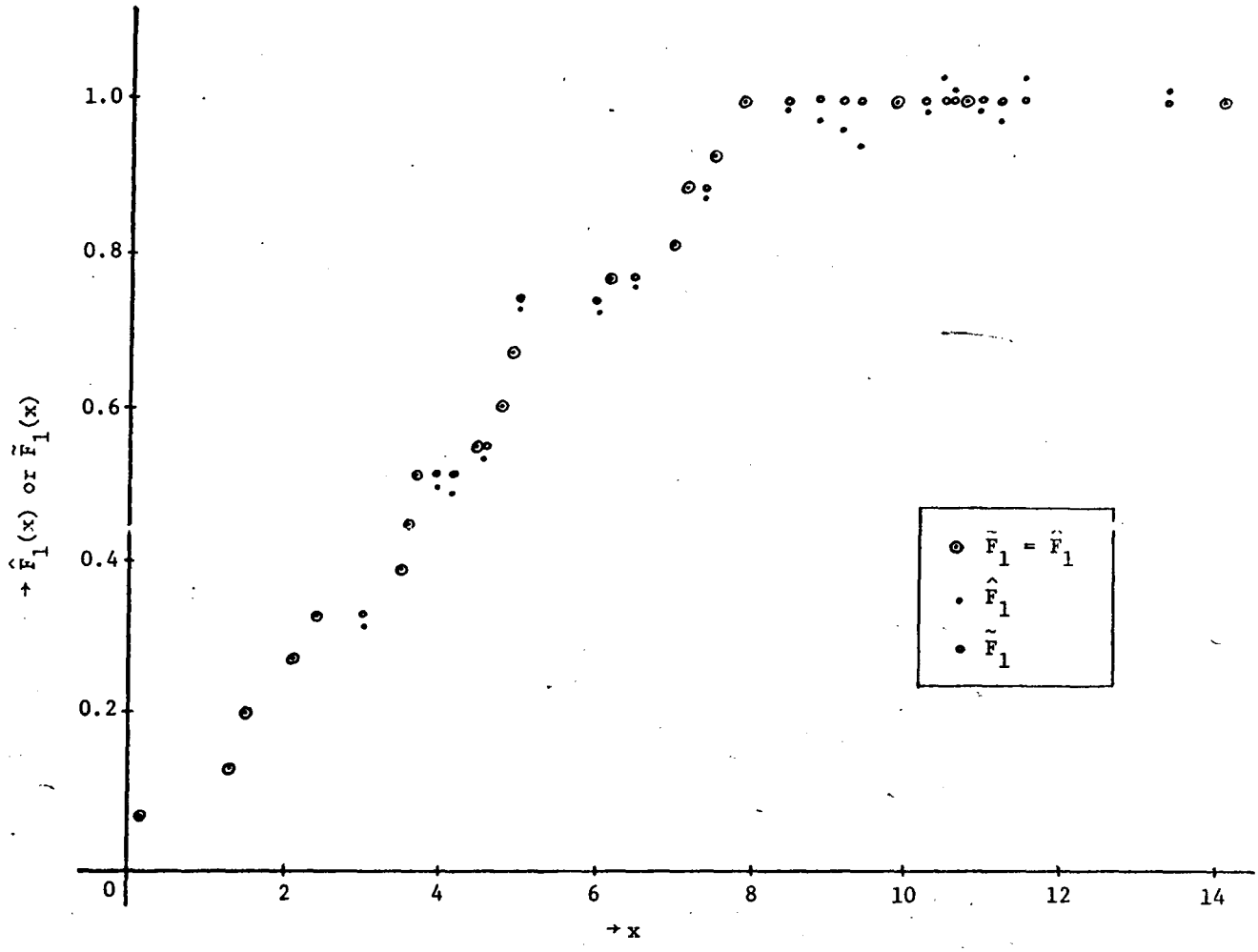
then we have the following.

THEOREM 4·1. *If F_1 and F_2 are continuous everywhere and* (4·22) *holds, then*

$$\sup_x\left\{n^{\frac{1}{2}}\left|\tilde{F}_i(x;n_1,n_2)-\hat{F}_i(x;n_1,n_2)\right|\right\}\longrightarrow 0 \text{ almost surely} \qquad (4\cdot23)$$

as $n\longrightarrow\infty$, *for* $i=1,2$.

FIGURE 4.1. The empirical df's $\hat{F}_1$ and $\tilde{F}_1$



The proof of the theorem is sketched in the Mathematical appendix. To illustrate the relative behavior of $\hat{F}$ and $\tilde{F}$, we obtain by the use of random numbers and tables for standard normal random deviates the following two samples, each of size 20, when F_1 and F_2 are normal df's with means 5 and 8 and standard deviations 2 and 3, respectively, and $p_1=q_2=0{\cdot}8$. The ordered variables for the two samples are, respectively, 0·18, 1·32, 1·46, 2·06, 2·40, 3·50, 3·62, 3·66, 4·52, 4·80, 4·92, 4·94, 6·18, 7·00, 7·18, 7·52, 7·85, 9·89, 10·46, 11·45 and 2·98, 3·98, 4·11, 4·55, 5·00, 5·92, 6·58, 7·37, 8·45, 8·93, 9·20, 9·41, 10·28, 10·46, 10·58, 10·71, 10·91, 11·15, 13·43, 14·12. On using (4·1) and computing $\hat{F}_1$ and $\hat{F}_2$, we immediately observe that $\hat{F}_2(x)<0$ for $x<3{\cdot}98$ and $\hat{F}_1(x)>1$ for $x\geq 10{\cdot}46$. On the same graph paper, we plot $\hat{F}_1(x)$ and $\tilde{F}_1(x)$, see figure 4·1. Only the points of discontinuity of $\hat{F}_1(x)$ or $\tilde{F}_1(x)$ are spotted on the graph. Whereas $\hat{F}_1$ can have both positive and negative jumps, $\tilde{F}_1$ is non-decreasing. The maximum displacement between $\hat{F}_1$ and $\tilde{F}_1$, in this case is 0·067, and it occurs at $x=9{\cdot}41$. Similar conclusions hold for $\hat{F}_2$ and $\tilde{F}_2$.

(Figure 4·1 goes here)

For discrete F_1 and F_2, we have to impose the following condition that the df's F_1 and F_2 both have the common jump points; otherwise, the identity of the question A or B may be revealed by a look at the response. If x be a jump point of F_1 and F_2 (and hence, of G_1 and G_2), on noting that

$$(q_2\{G_1(x)-G_1(x-0)\} - q_1\{G_2(x)-G_2(x-0)\})/(p_1-p_2)=\{F_1(x)-F_1(x-0)\}>0, \quad (4{\cdot}24)$$

it can be shown that as $n_1,n_2 \longrightarrow \infty$,

$$\hat{F}_1(x;n_1,n_2)-\hat{F}_1(x-0;n_1,n_2) > 0 \text{ almost surely,} \quad (4{\cdot}25)$$

so that Theorem 4·1 readily extends to this situation. The same result holds when the X_{ij} are recorded on suitable interval scale, where the df's can only

be estimated for the cell boundaries.

In the remaining of this section, we consider the case of continuous and univariate F_1 and F_2, and provide suitable confidence bands to them. For this, let us define

$$\beta = (p_1-p_2)/q_1,\ \eta = (p_1-p_2)/q_2 \text{ and } \delta = \eta/\beta = q_1/q_2. \qquad (4\cdot26)$$

We shall provide two alternative confidence bands. The first one, analogous to the confidence bands for $P\{X<Y\}$, based on two independent samples, considered by Birnbaum and McCarty (1958), is based on the two Kolmogorov statistics for G_{n_1} and G_{n_2}. The second one is based on the technique of Sen, Bhattacharyya and Suh (1973, Section 4).

Note that from (2·2) and (4·1), we have

$$\hat{F}_1(x)-F_1(x)=\eta^{-1}[G_{n_1}(x)-G_1(x)]-\beta^{-1}[G_{n_2}(x)-G_2(x)], \qquad (4\cdot27)$$

so that

$$\sup_x \left| \hat{F}_1(x)-F_1(x) \right| \leq \eta^{-1}\left\{\sup_x \left| G_{n_1}(x)-G_1(x)\right|\right\}+\beta^{-1}\left\{\sup_x \left| G_{n_2}(x)-G_2(x)\right|\right\}$$

$$=\eta^{-1}D_{n_1}+\beta^{-1}D_{n_2} = D_{n_1n_2} \text{ (Say)}, \qquad (4\cdot28)$$

where D_{n_1} and D_{n_2} are the one-sample Kolmogorov statistics whose distributions do not depend on G_1 and G_2, when these are continuous. Let us denote the df of D_{n_1} by $\Phi(x;n_1)$, $n_1\geq 1$. Note that for small n_1, extensive tables for $\Phi(x;n_1)$ are available [viz., Birnbaum (1952), Owen (1962)], while for large n_1,

$$\Phi(n_1^{-\frac{1}{2}}x;n_1) \longrightarrow L(x)=1-2\Sigma_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}, \qquad (4\cdot29)$$

for every fixed $x(\geq 0)$. From (4·28), we have

$$P\{D_{n_1n_2}\leq x\} = P\{\eta^{-1}D_{n_1}+\beta^{-1}D_{n_2}\leq x\}$$

$$= P\{D_{n_1}+\delta D_{n_2}\leq \eta x\}$$

$$= \int_0^{\beta x}\Phi(\eta x-\delta y;n_1)\,d\Phi(y;n_2), \qquad (4\cdot30)$$

so that for every $a>0$,

$$P\left\{\sup_x \left|\hat{F}_1(x)-F_1(x)\right| \psi\, a\right\} \geq \int_0^{\beta a} \Phi(\eta a-\delta y;n_1)\,d\Phi(y;n_2), \tag{4·31}$$

which provides the desired distribution-free confidence bounds to F_1, by equating the right hand side of (4·31) to the desired confidence coefficient $1-\alpha, 0<\alpha<1$, by a proper choice of a. For large n_1,n_2, we let [as in (4·22)] $n_1/n=\lambda_n$ and assume that

$$\lim_{n\to\infty} \lambda_n=\lambda \text{ exists and } \lambda_0 \leq \lambda \leq 1-\lambda_0. \tag{4·32}$$

Let then

$$D_n^*=\sqrt{n}D_{n_1n_2} \text{ and } D_{n_i}^* =\sqrt{n_i}D_{n_i}, \ i=1,2.$$

Then, by (4·29) and (4·30), under (4·32),

$$P\{D_n^*<a\}=P\{D_{n_1}^* +\delta\sqrt{\lambda/(1-\lambda)}D_{n_2}^* \leq a\eta\sqrt{\lambda}\}$$

$$\longrightarrow \int_0^{\beta a\sqrt{1-\lambda}} L(a\eta\sqrt{\lambda}-u\delta\sqrt{\lambda/(1-\lambda)})\,dL(u) \tag{4·33}$$

$$= 1-\{[1-L(\beta a\sqrt{1-\lambda})]+ \int_0^{\beta a\sqrt{1-\lambda}} \{1-L(a\eta\sqrt{\lambda}-u\delta\sqrt{\lambda/(1-\lambda)})\}dL(u)\}$$

$$= 1-2\Sigma_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\beta^2(1-\lambda)^2a^2}-$$

$$2\Sigma_{j=1}^{\infty}\Sigma_{\ell=1}^{\infty}(-1)^{j+\ell} \int_0^{\beta a\sqrt{1-\lambda}} e^{-2j^2\lambda[\eta a-\delta u/\sqrt{1-\lambda}]^2} 4\ell^2 u e^{-2\ell^2u^2}\,du,$$

for every $a>0$. The series approximation in (4·33) is usually quite rapidly convergent, and for specific β, η and a, only a few terms on the right hand side of (4·33) gives an adequate approximation.

For the second procedure, we define

$$\nu = (\lambda\eta^2)^{-1}+((1-\lambda)\beta^2)^{-1} \text{ and } \xi = (\lambda\eta^2)^{-1}\nu^{-1}. \tag{4·34}$$

Then, we have the following.

THEOREM 4·2. Under (4·32), for every $a>0$,

$$\lim_{n\to\infty} P\left\{\sup_x \sqrt{n}\left|\hat{F}_1(x;n_1,n_2)-F_1(x)\right| \leq a\right\}$$

$$\geq L(a/\sqrt{\nu}) = 1-2\Sigma_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2a^2\nu^{-1}}, \tag{4·35}$$

where the equality sign holds when $G_1 \equiv G_2$.

The proof of the theorem is sketched in the matematical appendix. Since n_1, n_2, p_1 and p_2 are all specified, ν is specified, so that by equating the right hand side of (4·35) to our desired confidence coefficient $1-\alpha$ $(0<\alpha<1)$, one gets a confidence band for F_1. The case of F_2 follow similarly. Computationally, (4·35) is a lot simpler than (4·33).

5. THE CASE OF MORE THAN TWO QUESTIONS.

Suppose that $A_1,\ldots,A_t, t\geq 2$, are the t questions, and in the ith sample, a respondent selects (at random) the question A_j with the probability p_{ij}, for $j=1,\ldots,t, i=1,\ldots,t$. The actual distributions of the responses for the questions $A_1,\ldots,A_t$ are denoted by $F_1(x),\ldots,F_1(x)$, respectively, while the distribution of the response for the ith sample observations is denoted by $G_i(x), i=1,\ldots,t$. Then

$$G_i(x) = \Sigma_{j=1}^{t} p_{ij} F_j(x), \quad 1\leq i\leq t. \tag{5·1}$$

Let us denote by

$$\underset{\sim}{P} = \begin{pmatrix} p_{11}\cdots p_{1t} \\ \vdots \\ p_{t1}\cdots p_{tt} \end{pmatrix} \tag{5·2}$$

and assume that $\underset{\sim}{P}$ is positive definite, so that $\underset{\sim}{P}^{-1}$ exists. [For t=2, $p_1>p_2 \Rightarrow \underset{\sim}{P}$ is positive definite. But for t>2, more stringent condition on the p_{ij} may be needed.] Also, let

$$\underset{\sim}{G}(x) = \begin{bmatrix} G_1(x) \\ \vdots \\ G_t(x) \end{bmatrix} \text{ and } \underset{\sim}{F}(x) = \begin{bmatrix} F_1(x) \\ \vdots \\ F_t(x) \end{bmatrix} \tag{5·3}$$

Then by (5·1) - (5·3),

$$\underset{\sim}{G}(x) = \underset{\sim\sim}{P}\underset{\sim}{F}(x) \quad \text{i.e., } \underset{\sim}{F}(x) = \underset{\sim}{P}^{-1}\underset{\sim}{G}(x) \tag{5·4}$$

The last equation provides the necessary clue for the estimation of $\underset{\sim}{F}$ and its functionals, and the results of Sections 3 and 4 can be readily extended for $t\geq 2$. For intended brevity, these are not reproduced.

MATHEMATICAL APPENDIX

1. The proof of Theorem 4·1. We prove (4·23) only for i=1, as the same proof holds for i=2. Also, for simplicity of proof, we consider the univariate case where F_1, F_2 (and hence, G_1 and G_2) are defined on the real line $(-\infty,\infty)$; the multivariate extension is straightforward, and hence, is not considered. We are to show that as $n \longrightarrow \infty$,

$$\sup_{0\leq t\leq 1}\left\{\sqrt{n}\left|\tilde{F}_1(F_1^{-1}(t);n_1,n_2)-\hat{F}_1(F_1^{-1}(t);n_1,n_2)\right|\right\}\longrightarrow 0 \text{ a.s. (almost surely).} \quad (A\cdot 1)$$

We make use of the elegant Bahadur representation of sample quantiles, as extended to the case of non-identically distributed random variables [viz., Sen (1968)] along with the basic inequality in Theorem 4·2 of Sen and Ghosh (1971), and following a few standard steps, obtain that under (4·22), as $n \longrightarrow \infty$,

$$\sup\{\left|\hat{F}_1(F_1^{-1}(u);n_1,n_2)-\hat{F}_1(F_1^{-1}(t);n_1,n_2)-(u-t)\right|: u,t\in[0,1] \text{ and } |u-t|<n^{-\frac{1}{2}}\log n\} = O(n^{-3/4}\log n), \quad (A\cdot 2)$$

almost surely. As such, as $n \longrightarrow \infty$, for every $\varepsilon>0$,

$$\sup_{0<t\leq 1}\left|[\hat{F}_1(F_1^{-1}(t+n^{-\frac{1}{2}}\varepsilon);n_1,n_2)-\hat{F}_1(F_1^{-1}(t);n_1,n_2)-\varepsilon n^{-\frac{1}{2}}]\right| = O(n^{-3/4}\log n), \quad (A\cdot 3)$$

almost surely, which implies that

$$\hat{F}_1(F_1^{-1}(t+n^{-\frac{1}{2}}\varepsilon);n_1,n_2)>\hat{F}_1(F_1^{-1}(t);n_1,n_2) \text{ a.s. for all } 0\leq t<1. \quad (A\cdot 4)$$

Therefore, by (4·21) and (A·4), as $n\longrightarrow\infty$, for every $\varepsilon>0$,

$$\hat{F}_1(F_1^{-1}(t-n^{-\frac{1}{2}}\varepsilon);n_1,n_2\leq\tilde{F}_1(F_1^{-1}(t);n_1,n_2)\leq\hat{F}_1(F_1^{-1}(t+n^{-\frac{1}{2}}\varepsilon);n_1,n_2) \quad (A\cdot 5)$$

for all $0<t<1$, almost surely. Consequently, as $n\longrightarrow\infty$,

$$\begin{aligned}&\sup_{0<t<1}\sqrt{n}\left|\tilde{F}_1(F_1^{-1}(t);n_1,n_2)-\hat{F}_1(F_1^{-1}(t);n_1,n_2)\right|\\ &\leq \sup_{0<t<1}\sqrt{n}\left|\hat{F}_1(F_1^{-1}(t+n^{-\frac{1}{2}}\varepsilon);n_1,n_2)-\hat{F}_1(F_1^{-1}(t-n^{-\frac{1}{2}}\varepsilon);n_1,n_2)\right| \quad (A\cdot 6)\\ &= 2\varepsilon + O(n^{-1/4}\log n) \text{ a.s., by (A·3).}\end{aligned}$$

Since $\varepsilon(>0)$ is arbitrary and $n^{-1/4}\log n \longrightarrow 0$ as $n \longrightarrow \infty$, the proof follows. Q.E.D.

2. The proof of Theorem 4·2. By (4·27), (4·32) and standard results on the weak convergence of empirical processes to Gaussian functions, it follows that as $n\longrightarrow\infty$, $\{n^{\frac{1}{2}}[\hat{F}_1(F_1^{-1}(t);n_1,n_2)-t],\ 0\leq t\leq 1\}$ converges in law to a Gaussian function $W = \{W(t),0\leq t\leq 1\}$ where $EW(t)=0$, $0\leq t\leq 1$, and for $0\leq s\leq t\leq 1$,

$$EW(s)W(t) = G_1(F_1^{-1}(s))[1-G_1(F_1^{-1}(t))]/\lambda\eta^2+G_2(F_1^{-1}(s))[1-G_2(F_1^{-1}(t))]/(1-\lambda)\beta^2. \quad (A\cdot 7)$$

Using (4·34), (A·7) simplifies to

$$\nu\{H(F_1^{-1}(s))[1-H(F_1^{-1}(t))]-\xi(1-\xi)[G_1(F_1^{-1}(s))-G_2(F_1^{-1}(s))][G_2(F_1^{-1}(t))-G_2(F_1^{-1}(t))]\}, \quad (A\cdot 8)$$

where

$$H(x) = \xi G_1(x)+(1-\xi)G_2(x) \text{ for } -\infty<x<\infty. \quad (A\cdot 9)$$

Let now $\{W^o(t),0\leq t\leq 1\} = W^o$ be a Gaussian function with $EW^o(t) = 0$ for $0\leq t\leq 1$, and for $0\leq s\leq t\leq 1$,

$$EW^o(s)W^o(t) = \nu H(F_1^{-1}(s))[1-H(F_1^{-1}(t))]. \quad (A\cdot 10)$$

Note that $\{\nu^{-\frac{1}{2}}W^o(F_1(H^{-1}(t))),0\leq t\leq 1\}$ is a standard Brownian bridge, so that for every $d\geq 0$,

$$P\left\{\sup_{0\leq t\leq 1}\left|W^o(t)\right|\leq d\right\} = P\left\{\sup_{0\leq t\leq 1} v^{-\frac{1}{2}}\left|W^o(F_1(H^{-1}(t)))\right|\leq \nu^{-\frac{1}{2}}d\right\}$$
$$= L(d/\nu^{\frac{1}{2}}) = 1-2\Sigma_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2d^2\nu^{-1}}, \quad (A\cdot 11)$$

where the last step follows from the well-known result on the Brownian bridge on [0,1], viz., Billingsley (1968,p. 85). Also, note that for every $m(\geq 1)$ and arbitrary $\underset{\sim}{t}_m = (t_1,\ldots,t_m)$ (with $0\leq t_1<\cdots<t_m\leq 1$), if $\underset{\sim}{D}_m(\underset{\sim}{t}_m)$ and $\underset{\sim}{D}^o_m(\underset{\sim}{t}_m)$ be the dispersion matrices of $[W(t_1),\cdots,W(t_m)]$ and $[W^o(t_1),\cdots,W^o(t_m)]$, respectively, then by (A·8) and (A·10),

$$\underset{\sim}{D}_m(\underset{\sim}{t}_m) = \underset{\sim}{D}^o_m(\underset{\sim}{t}_m) - \underset{\sim}{D}^*_m, \quad (A\cdot 12)$$

where $\underset{\sim}{D}^*_m = \nu\xi(1-\xi)(([G_1(F_1^{-1}(t_i))-G_2(F_1^{-1}(t_i))][G_1(F_1^{-1}(t_j))-G_2(F_1^{-1}(t_j))]))$ is positive semi-definite. Consequently, by Lemma 4·4 of Sen, Bhattacharyya and Suh (1973), we have for every $d\geq 0$,

$$P\left\{\max_{1\le j\le m}\left|W(t_j)\right|\le d\right\}\ge P\left\{\max_{1\le j\le m}\left|W^o(t_j)\right|\le d\right\}, \qquad (A\cdot 13)$$

where the equality sign holds when $\underset{\sim}{D}^*_m = \underset{\sim}{0}$. Since (A·13) holds for every $m\ge 1$ and arbitrary $0\le t_1<\cdots<t_m\le 1$, passing on to the limit ($m\longrightarrow\infty$), we conclude that for every $d\ge 0$,

$$P\left\{\sup_{0\le t\le 1}\left|W(t)\right|\le d\right\}\ge P\left\{\sup_{0\le t\le 1}\left|W^o(t)\right|\le d\right\}, \qquad (A\cdot 14)$$

where the equality sign holds when $G_1\equiv G_2$. Therefore, by the weak convergence of $\{\sqrt{n}[\hat{F}_1(F_1^{-1}(t);n_1,n_2)-t], 0\le t\le 1\}$ to W, and by (A·11) and (A·14), we obtain that for every $a\ge 0$,

$$\begin{aligned}\lim_{n\longrightarrow\infty} P\left\{\sup_x \sqrt{n}\left|\hat{F}_1(x;n_1,n_2)-F_1(x)\right|\le a\right\} &= P\left\{\sup_{0\le t\le 1}\left|W(t)\right|\le a\right\}\\ &\ge P\left\{\sup_{0\le t\le 1}\left|W^o(t)\right|\le a\right\}\\ &= L(a/\sqrt{\nu}) = 1-2\Sigma_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2a^2\nu^{-1}}, \qquad (A\cdot 15)\end{aligned}$$

which completes the proof.

REFERENCES

Abernathy, J.R., Greenberg, B.G., Horvitz, D.G. "Estimates of induced abortion in Urban North Carolina," Demography 7, (1970) 19-29.

Billingsley, P. (1968). Convergence of Probability Measures. New York, John Wiley.

Birnbaum, Z.W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample sizes. Jour. Amer. Statist. Assoc. 47, 425- .

Birnbaum, Z.W., McCarty, R.C. (1958). A distribution-free upper bound for $P_r\{X<Y\}$ based on independent samples of X and Y. Ann. Math. Statist. 29, 558-583.

Fraser, D.A.S. (1953). Completeness of the order statistics. Canadian Jour. Math. 6, 42- .

Greenberg, B.G., Kuebler, R.R., Jr., Abernathy, J.R., Horvitz, D.G. "Application of the randomized response technique in obtaining quantitative data". Jour. Amer. Statist. Assoc. 66, (1971) 243-250.

Halmos, P.R. (1946). The theory of unbiased estimation. Ann. Math. Statist. 17, 34-43.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19, 293-325.

Mises, R. von (1967). On the asymptotic distribution of differentiable statistical functions. Ann. Math. Statist. 18, 309-348.

Owen, D.B. (1962). Handbook of Statistical Tables. Reading, Mass; Addison-Wesley.

Puri, M.L., Sen, P.K. (1971). Nonparametric Methods in Multivariate Analysis. New York, John Wiley.

Sen, P.K. (1960). On some convergence properties of U-statistics. Calcutta Statist. Assoc. Bull. 10, 1-18.

Sen, P.K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. Ann. Math. Stat. 39, 1724-1730.

Sen, P.K. Bhattacharyya, B.B., Suh, M.W. (1973). Limiting behavior of the extremum of certain sample functions. Ann. Statist. 1, .

Sen, P.K., Ghosh, M. (1971). On bounded length sequential confidence intervals based on one-sample rank order process. Ann. Math. Statist. 42, 189-203.