THE EXTENDED TWO-SAMPLE PROBLEM: NONPARAMETRIC CASE

by

Pranab Kumar Sen

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1140

August 1977

THE EXTENDED TWO-SAMPLE PROBLEM: NONPARAMETRIC CASE*

Pranab Kumar Sen
University of North Carolina, Chapel Hill

The classical two-sample problem is extended here to the case where the distribution functions of the observable random variables are specified functions of unknown distribution functions and the null hypothesis to be tested or the parameter to be estimated relates to these unknown distributions. Various properties of the proposed rank tests and derived estimates are studied.

AMS 1970 Classification Nos: 62G02, 62G10

Key Words & Phrases: Asymptotic efficiency, asymptotic normality, distribution-freeness, equality of basic distributions, rank based estimates, rank order tests.

---

*Work partially supported by the Air Force Office of Scientific Research, U.S.A.F., Grant No. AFOSR 74-2736.

1. INTRODUCTION

In the classical two-sample problem, given two independent samples from two (continuous) distributions F and G, we intend either to test for the identity of F and G or to estimate some parameter which relates G to F in a meaningful way [e.g., $G(x) \equiv F(x-\theta)$ where θ is the difference of locations of the two distributions]. In an extended two-sample problem, we conceive of two independent samples from two continuous distributions F^* and G^* respectively, where

$$F^*(x) \equiv Q_1(F(x)) \ , \quad G^*(x) \equiv Q_2(G(x)) \ , \tag{1.1}$$

$Q_1 = \{Q_1(u),\ 0 \le u \le 1\}$ and $Q_2 = \{Q_2(u),\ 0 \le u \le 1\}$ are *known*, non-decreasing functions (with $Q_1(0) = Q_2(0) = 0$, $Q_1(1) = Q_2(1) = 1$) and the basic (continuous) distributions F and G are not specified. We intend to test for

$$H_0: \ F \equiv G \ \text{ (without necessarily assuming that } \ Q_1 \equiv Q_2) \ . \tag{1.2}$$

[If $Q_1 \equiv Q_2$ then $F \equiv G \Rightarrow F^* \equiv G^*$, so that the classical theory holds.] Also, relating G to F in a meaningful way, we desire to estimate the allied parameters based on the observations from the distributions F^* and G^*. As illustration, we consider the following:

Example 1. Consider a *system* with $k(\ge 1)$ electric cells connected in series (i.e., the low potential pole of the i-th cell is connected to the high potential cell of the $(i+1)$-th one, for $i = 1,\ldots,k-1$) and assume that each cell has a life distribution F. Then, the life distribution of the system is $F^*(x) \equiv 1-[1-F(x)]^k$. Suppose that there is a second system with

$\ell(\geq 1)$ cells in series and the life distribution of each cell is G, so that the system has a life distribution $G^*(x) \equiv 1-[1-G(x)]^{\ell}$. Thus, for samples relating to the two systems, (1.1) holds with $Q_1(u) \equiv 1-(1-u)^k$ $Q_2(u) \equiv 1-(1-u)^{\ell}$.

Example 2. Suppose that in Example 1, the cells are connected in parallel (i.e., all the high potential poles are connected together by a conducting wire and all the low potential poles together in the sameway). Then, $F^*(x) \equiv [F(x)]^k$ and $G^*(x) \equiv [G(x)]^{\ell}$. Also, if C be the individual capacity of the cells and the system becomes inoperative when its capacity is less that C^* (where, for some r: $1 \leq r \leq k$, $(k-r)C < C^* \leq (k-r+1)C$), then $F^*(x) \equiv \sum_{j=r}^{k} \binom{k}{j} [F(x)]^j [1-F(x)]^{k-j}$. A similar case holds for the second system. In both these examples, k and ℓ need not be equal. Similar examples arise in problems of reliability theory and competing risks.

When $Q_1 \equiv Q_2$, $F \equiv G \Rightarrow F^* \equiv G^*$, and hence, the classical nonparametric theory holds. We intend to show that so long as Q_1 and Q_2 are specified, the classical theory can readily be extended and suitable rank tests can be constructed. Towards this, note that under H_0 in (1.2),

$$G^*(x) \equiv Q_0(F^*(x)) \quad \text{where} \quad Q_0(u) \equiv Q_2(Q_1^{-1}(u)) \ , \quad 0 \leq u \leq 1 \ , \qquad (1.3)$$

and Q_0 is a specified, non-decreasing function with $Q_0(0) = 0$ and $Q_0(1) = 1$. This representation (characterizing the distribution-free nature of rank based tests) is exploited in Section 2 in the proposal of suitable rank tests. The asymptotic distribution theory is dealt with in Section 3 and this is incorporated in Section 4 in the study of optimal rank

statistics for some local alternatives. Section 5 deals with the allied estimation problem and some general remarks are made in the concluding section.

2. THE PROPOSED RANK TESTS

Let $X_1,\ldots,X_m$ be independent and identically distributed random variables (i.i.d.r.v.) with a continuous distribution function (df) $F^*(x)$, and let $Y_1,\ldots,Y_n$ be i.i.d.r.v. with a continuous df $G^*(x)$, where F^* and G^* satisfy (1.1). Let $N = m+n$ and $R_{n1},\ldots,R_{Nm},\ldots,R_{NN}$ be respectively the ranks of $X_1,\ldots,X_m$, $Y_1,\ldots,Y_n$ (in the combined sample); by virtue of the assumed continuity of F^* and G^*, ties among the observations may be neglected, in probability, so that $\underset{\sim}{R}_N = (R_{N1},\ldots,R_{NN})$ is a (random) permutation of $(1,\ldots,N)$. Let $a_N(1),\ldots,a_N(N)$ be a set of (real valued) scores and we consider the usual two-sample rank order statistic

$$T_N = m^{-1}\sum_{i=1}^{m} a_N(R_{Ni}) \tag{2.1}$$

and with suitable $\{a_N(i)\}$, we intend to use T_N as a test statistic.

Let $\underset{\sim}{r}_N = (r_1,\ldots,r_N)$ be any permutation of $(1,\ldots,N)$ and let $\mathcal{R}_N$ be the set of all possible $N!$ permutations. Then, we have

$$\begin{aligned} P\{\underset{\sim}{R}_N = \underset{\sim}{r}_N \mid H_0\} &= \left[\int\cdots\int_{z_1<\ldots<z_n} \prod_{i=1}^{m} dF^*(z_{r_i}) \prod_{j=1}^{n} dG^*(z_{r_{m+j}})\right] \\ &= \int\cdots\int_{0<u_1<\ldots<u_N<1} \prod_{i=1}^{m} du_{r_i} \prod_{j=1}^{n} dQ_0(u_{r_{m+j}}) \\ &= \int\cdots\int_{0<u_1<\ldots<u_N<1} \prod_{r=1}^{N} d\left\{u_r^{1-\chi_r}[Q_0(u_r)]^{\chi_r}\right\}, \quad \forall\ \underset{\sim}{r}_N \in \mathcal{R}_N, \end{aligned} \tag{2.2}$$

where

$$\chi_r = \begin{cases} 1\ , & r = r_{m+i}\ , \quad \text{for} \quad 1 \le i \le n \\ 0\ , & r = r_i\ , \quad \text{for} \quad 1 \le i \le m\ , \quad \text{for} \quad 1 \le r \le N\ . \end{cases} \tag{2.3}$$

[Thus, there is an one-to-one correspondence between $\underset{\sim}{r}_N$ and the χ_r, $1 \le r \le N$.] Since $Q_0(u)$, $0 \le u \le 1$ is specified under H_0 in (1.2), the distribution of $\underset{\sim}{R}_N$, under H_0, does not depend on F, so that a test based on $\underset{\sim}{R}_N$ (and hence, on T_N in (2.1)) is distribution-free. Note that for $Q_0(u) \ne u$, the probability function in (2.2) need not relate to a (discrete) uniform distribution over the $n!$ realizations $\{\underset{\sim}{r}_n \in \mathcal{R}_N\}$. Nevertheless, the distribution can be enumerated by using (2.2) for a given Q_0 for finite m and n. For instance, in Example 1, $Q_0(u) = 1-(1-u)^a$ where $a = \ell/k (> 0)$ (as $1-G^*(x) = [1-G(x)]^\ell = ([1-F(x)]^k)^a = (1-F^*(x))^a$), and hence, (2.2) reduces to

$$a^n \left\{ \prod_{r=1}^{n} \left[\sum_{j=1}^{r} \left(1 - \chi_{N-j+1} + a\chi_{N-j+1} \right) \right] \right\}^{-1} , \quad \forall\ \underset{\sim}{r}_N \in \mathcal{R}_N\ . \tag{2.4}$$

This familiar expression arises in the so called Lehmann [5] alternative case for the classical two-sample problem (though, there it relates to the alternative hypothesis, whereas, here, it relates to the null case). Similar expressions hold for the second example.

We may note that for every r: $1 \le r \le N$,

$$P\{R_{N1} = r|H_0\} = \Big[\sum_{s=1}^{r}\binom{m-1}{s-1}\binom{n}{r-s}\int_{-\infty}^{\infty}[F^*(x)]^{s-1}[1-F^*(x)]^{n-s}\cdot$$

$$[G^*(x))]^{r-s}[1-G^*(x)]^{n-r+s}dF^*(x)\Big]_{F\equiv G} \qquad (2.5)$$

$$= \sum_{s=1}^{r}\binom{m-1}{s-1}\binom{n}{r-s}\int_0^1 u^{s-1}(1-u)^{m-s}[Q_0(u)]^{r-s}(1-Q_0(u)]^{n-r+s}du\ ,$$

so that

$$E[T_N|H_0] = E[a_N(R_{N1})|H_0]$$

$$= \sum_{r=1}^{N} a_N(r)P\{R_{N1} = r|H_0\}\ . \qquad (2.6)$$

Similarly,

$$E[T_N^2|H_0] = m^{-2}\{mE[a_N^2(R_{N1})|H_0] + m(m-1)E[a_N(R_{N1})a_N(R_{N2})|H_0\}$$

$$= m^{-1}\sum_{r=1}^{N} a_N^2(r)P\{R_{N1} = r|H_0\} \qquad (2.7)$$

$$+ (1-m^{-1})\sum_{r\neq s=1}^{N} a_N(r)a_N(s)P\{R_{N1} = r,\ R_{N2} = s|H_0\}\ ,$$

where for every $1 \le r \le s \le N$,

$$P\{R_{N1} = r,\ R_{N2} = s|H_0\} = P\{R_{N2} = r,\ R_{N1} = s|H_0\}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s-r}\frac{(m-1)!}{(i-1)!(j-1)!(m-i-j)!}\frac{n!}{(r-i)!(s-r-j)!(n-s+i+j)!}\cdot$$

$$\cdot \iint_{0<u<v<1} u^{i-1}(v-u)^{j-i-1}(1-v)^{m-i-j}[Q_0(u)]^{r-i}[Q_0(v)-Q_0(u)]^{s-r-j}\cdot$$

$$\cdot [1-Q_0(v)]^{n-s+i+j}dudv\ . \qquad (2.8)$$

Thus, for a specified Q_o, $E(T_n|H_0)$ and $V(T_N|H_0)$ can be computed. As m or n increases, the computation of the exact null distribution

of T_N (or its mean and variance) becomes prohibitively laborious. For this reason, in the next section, we take recourse to the asymptotic case and provide simple approximations under appropriate regularity conditions.

3. ASYMPTOTIC DISTRIBUTION OF T_N

Let $u(1) = 1$ or 0 according as t is $\geq$ or < 0 and define

$$F_m^*(x) = m^{-1}\sum_{i=1}^{m} u(x-X_i) \ , \quad G_n^*(x) = n^{-1}\sum_{j=1}^{n} u(x-Y_j) \quad (-\infty < x < \infty) \tag{3.1}$$

$$\lambda_N = m/N \quad \text{and} \quad H_N^*(x) = \lambda_N F_m^*(x) + (1-\lambda_N)G_n^*(x) \ , \ -\infty < x < \infty . \tag{3.2}$$

Thus, F_m^*, G_n^* and H_N^* are respectively the first, second and combined sample empirical df's. Let then

$$H^*(x) = \lambda_N F^*(x) + (1-\lambda_N)G^*(x) \ , \ -\infty < x < \infty . \tag{3.3}$$

[Note that H^* may depend on N through λ_N; for notational simplicity, this dependence is understood.] We conceive of a *score function* $\phi = \{\phi(u), 0 < u < 1\}$ such that $\phi(u) = \phi_1(u) - \phi_2(u)$, $0 < u < 1$ where both ϕ_1 and ϕ_2 are absolutely continuous and non-decreasing with

$$\int_0^1 |\phi_j(t)|\{t(1-t)\}^{-\frac{1}{2}}dt < \infty \ , \quad \text{for} \quad j = 1,2 \ . \tag{3.4}$$

[This is slightly more restrictive than the square integrability condition of the ϕ_j, but, is less restrictive than $\int_0^1 |\phi_j(u)|^r du < \infty$ for some $r > 2$, $j = 1,2$.] Then, we assume that the scores $\{a_N(i)\}$ are defined by

$$a_N(i) = \phi(i/(N+1)) \quad \text{or} \quad E\phi(U_{Ni}) \ , \quad \text{for} \quad i = 1,\ldots,N(\geq 1) \ , \tag{3.5}$$

where $U_{N1} < \ldots < U_{NN}$ are the ordered random variables of a sample of size N from the rectangular $(0,1)$ df. In particular, when $\phi(u) \equiv u$ (or the inverse of the standard normal df), $a_N(i) = i/(N+1)$ (or the normal scores). Let us then define

$$\mu = \int_{-\infty}^{\infty} \phi(H^*(x))dF^*(x) \ , \tag{3.6}$$

$$\sigma_1^2 = 2 \iint\limits_{-\infty<x<y<\infty} F^*(x)[1-F^*(y)]\phi'(H^*(x))\phi'(H^*(y))dG^*(x)dG^*(y) \tag{3.7}$$

$$\sigma_2^2 = 2 \iint\limits_{-\infty<x<y<\infty} G^*(x)[1-G^*(y)]\phi'(H^*(x))\phi'(H^*(y))dF^*(x)dF^*(y) . \tag{3.8}$$

Note that under H_0 in (1.2), we have

$$\mu = \mu_0 = \int_0^1 \phi(\lambda_N u + (1-\lambda_N)Q_0(u))du \tag{3.9}$$

$$\sigma_1^2 = \sigma_{10}^2 = 2 \iint\limits_{0<u<v<1} u(1-v)\phi'(\lambda_N u+(1-\lambda_N)Q_0(u))\phi'(\lambda_N v+(1-\lambda_N)Q_0(v))dQ_0(u)dQ_0(v) \tag{3.10}$$

$$\sigma_2^2 = \sigma_{20}^2 = 2 \iint\limits_{0<u<v<1} Q_0(u)[1-Q_0(v)]\phi'(\lambda_N u+(1-\lambda_N)Q_0(u))\phi'(\lambda_N v+(1-\lambda_N)Q_0(v))dudv. \tag{3.11}$$

[The dependence of μ, μ_0, σ_1^2, σ_{10}^2, σ_2^2 and σ_{20}^2 on N through λ_N is understood.] Finally, we assume that there exist a λ_0 $(0 < \lambda_0 \leq \frac{1}{2})$ and a_n $N_0(\geq 2)$, such that

$$(0 <)\lambda_0 \leq \lambda_N \leq 1-\lambda_0 (< 1) \ , \quad \forall \ N \geq N_0 \ . \tag{3.12}$$

Then we have the following.

Theorem 1. *Under the assumptions made above, $\max(\sigma_1^2,\sigma_2^2) > 0$ insures that as $N \to \infty$, for every real $x(-\infty < x < \infty)$,*

$$P\{N\ (T_N-\mu)/\sigma_N \le x\} \to (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,dt\ , \tag{3.13}$$

where

$$\sigma_N^2 = (1-\lambda_N)^2\{\sigma_1^2/\lambda_N + \sigma_2^2/(1-\lambda_N)\} \tag{3.14}$$

and further

$$N^{\frac{1}{2}}|E(T_N)-\mu| \to 0 \quad \textit{and} \quad NV(T_N)/\sigma_N^2 \to 1\ . \tag{3.15}$$

The proof of (3.13) follows directly from Theorem 2.3 of Hájek [2] after noting that our T_N is a special case of his statistic (where $c_1 = \ldots = c_m = 1$, $c_{m+1} = \ldots = c_N = 0$), so that his expressions in (2.9) and (2.10) simplify considerably and also our (3.4) insures his square integrability condition. The second ascertain in (3.15) also follows from Hájek's Theorem 2.3. In fact, we have strengthened his square integrability condition to (3.4) with the objective of using Theorem 1 of Hoeffding [4] which insures the first ascertion in (3.15). In view of these, the details are omitted.

Note that Theorem 1 covers both the null and non-null cases. In the null case, Q_0 is specified, so that by (3.9)-(3.11), all the quantities μ_0, σ_{10}^2 and σ_{20}^2 are also specified, and hence if $\max(\sigma_{10}^2,\sigma_{20}^2) > 0$, then

$$Z_N^* = N^{\frac{1}{2}}(T_N-\mu_0)/\sigma_{N0} \sim N(0,1) \quad (\text{under } H_0) \tag{3.16}$$

where

$$\sigma_{N0}^2 = (1-\lambda_N)^2\{\sigma_{10}^2/\lambda_N + \sigma_{20}^2/(1-\lambda_N)\}\ . \tag{3.17}$$

Thus, a large sample test can be based on Z_N^* using the appropriate percentile point of the standard normal df. As illustration, we consider the case in Example 1 and Wilcoxon statistic (i.e., $\phi(u) \equiv u$). Then, we have

$$\mu_0 = \int_0^1 \{\lambda_N u + (1-\lambda_N)[1-(1-u)^a]\}du = 1 - \tfrac{1}{2}\lambda_N - (1-\lambda_N)/(a+1) , \tag{3.18}$$

$$\sigma_{10}^2 = 2a^2 \iint_{0<u<v<1} u(1-v)(1-u)^{a-1}(1-v)^{a-1}dudv = 2a^2[(a+1)(2a+1)(2a+2)]^{-1} \tag{3.19}$$

$$\sigma_{20}^2 = 2a^2 \iint_{0<u<v<1} [1-(1-u)^a](1-v)^a dudv = 2a[(a+1)(a+2)(2a+2)]^{-1} \tag{3.20}$$

where $a = \ell/k$. Hence, (3.17) holds whenever $0 < a < \infty$.

4. LOCALLY OPTIMAL RANK TESTS

Here, we shall consider some local alternative hypotheses (relating G to F), and in this context, study the optimal choice of score functions. First, we consider a sequence $\{K_N\}$ of Pitman-type translation alternatives, where

$$K_N: G(x) = G_{(N)}(x) \equiv F(x+N^{-\frac{1}{2}}\theta), \quad \theta \text{ real (and fixed)} . \tag{4.1}$$

Also, we assume that

$$\lim_{N\to\infty}\lambda_N = \lambda \text{ exists and } 0 < \lambda < 1 . \tag{4.2}$$

Further, we assume that $\phi(u)$, $Q_1(u)$, $Q_2(u)$ have continuous first order derivatives $\phi'(u)$, $q_1(u)$ and $q_2(u)$ respectively for almost all $u(0<u<1)$ and $F(x)$ possesses an absolutely continuous probability

density function $f(x)$ for almost all x. Let $\mu_{(n)}(\theta)$ be the value of μ in (3.7) when K_N holds. Then, by some standard steps, it follows that

$$\lim_{N\to\infty} N^{\frac{1}{2}}(\mu_{(N)}(\theta)-\mu_0) = \theta(1-\lambda)\int_{-\infty}^{\infty} f^2(x)q_1(F(x))q_2(F(x))\phi'(\lambda Q_1(F(x))+(1-\lambda)Q_2(F(x)))dx$$

$$= \theta(1-\lambda)B(F,Q_1,Q_2,\phi,\lambda) \ , \quad \text{say} \ ; \tag{4.3}$$

$$\lim_{N\to\infty} \sigma^2_{1(N)} = \bar{\sigma}^2_{10} \quad \text{and} \quad \lim_{N\to\infty} \sigma^2_{2(N)} = \bar{\sigma}^2_{20} \tag{4.4}$$

where $\sigma^2_{1(N)}$ and $\sigma^2_{2(N)}$ are defined by (3.7) and (3.8) with $G^*(x) \equiv Q_2(F(x-N^{-\frac{1}{2}}\theta))$ and $\bar{\sigma}^2_{10}$ and $\bar{\sigma}^2_{20}$ are defined by (3.10)-(3.11) with λ_N being replaced by λ. Hence, under $\{K_N\}$, $N^{\frac{1}{2}}(T_N-\mu_0)/\sigma_{N0}$ has asymptotically a normal distribution with unit variance and mean

$$\theta B(F,Q_1,Q_2,\phi,\lambda)/\{\bar{\sigma}^2_{10}/\lambda + \bar{\sigma}^2_{20}/(1-\lambda)\}^{\frac{1}{2}} \tag{4.5}$$

where both $\bar{\sigma}^2_{10}$ and $\bar{\sigma}^2_{20}$ also depend on F, Q_0, ϕ and λ. Thus, an optimal choice of ϕ should maximize (4.5) (for a fixed θ) and, in general, this depends on F, Q_1, Q_2 and λ in a rather involved way. In particular when $Q_1(u) \equiv Q_2(u) \equiv Q(u)$ ($\Rightarrow q_1(u) = q_2(u) = q(u)$, but $q(u)$ not necessarily equal to 1 for all $0<u<1$), (4.5) reduces to

$$\lambda(1-\lambda)\theta\left\{\int_{-\infty}^{\infty} f^2(x)q^2(F(x))\phi'(Q(F(x)))dx\right\}\Big/\left\{\int_0^1 \phi^2(u)du-\left\{\int_0^1 \phi(u)du\right\}^2\right\}^{\frac{1}{2}} . \tag{4.6}$$

If we define

$$\psi(u) = -\,[f(x)q'(F(x))/q(F(x)) + f'(x)/f(x)]_{Q(F(x))=u}$$

$$= -\,[(d/dx)\log\, q(F(x))f(x)]_{Q(F(x))=u} \ , \quad 0<u<1 \ , \tag{4.7}$$

and note that $\int_0^1 \psi(u)du = 0$, we obtain then by partial integration on the numerator of (4.6) that (4.6) is equal to

$$\lambda(1-\lambda)\theta\left[\int_0^1 (\phi(u)-\bar{\phi})\psi(u)du\right] \Big/ \left\{\int_0^1 (\phi(u)-\bar{\phi})^2 du\right\}^{\frac{1}{2}} \tag{4.8}$$

where $\bar{\phi} = \int_0^1 \phi(u)du$. Thus, here an optimal choice of ϕ is

$$\phi(u) \equiv \psi(u) \ , \quad 0 < u < 1 \ . \tag{4.9}$$

This is a direct extension of the parallel result for the calssical two-sample problem where $q(u) \equiv 1$ so that $\psi(u) = -f'(F^{-1}(u))/f(F^{-1}(u))$, $0 < u < 1$.

Let us next consider a sequence $\{K_N^*\}$ of scale alternatives, where

$$K_N^*: \ G(x) = G_{(N)}(x) \equiv F(x(1+N^{-\frac{1}{2}}\theta)) \ , \ \theta \text{ real (and fixed)} \tag{4.10}$$

(with $\theta > -K$, $K \geq 1$ for every $N \geq K^2$). In this case, in (4.3), the integral has to be replaced by

$$\int_{-\infty}^{\infty} xf^2(x)q_1(F(x))q_2(F(x)) \ \phi'(\lambda Q_1(F(x)) + (1-\lambda)Q_2(F(x)))dx \tag{4.11}$$

and a similar change is needed in (4.6). Similarly, in (4.7) and (4.9), $\psi(u)$ will have to change to

$$\begin{aligned}\psi^*(u) &= -1 - [x\{(d/dx)\log q(F(x))f(x)\}]_{Q(F(x))=u} \qquad (4.12)\\ &= -1 - \{F^{*-1}(u)\}\psi(u) \ , \ 0 < u < 1 \text{ where } F^*(x)=Q(F(x)) \ .\end{aligned}$$

In either case, it may be observed that for $Q_1 \equiv Q_2$, though the null distribution of T_N agrees with that in the classical two-sample problem,

the asymptotic power function and the optimal score function are generally different and depend on $q'(u)$ as well as $f'(x)$.

5. ESTIMATION BASED ON T_N

In the classical two-sample problem, the problem of estimation based on linear rank statistics has been treated by Hodges and Lehmann [3] and Sen [7]. In view of the results in Section 3, we shall extend it to the extended two-sample problem as follows. Suppose now that (1.1) holds with

$$G(x) \equiv F(x+\theta) \quad \text{where} \quad \theta \quad \text{(real) is unknown} , \tag{5.1}$$

and our problem is to estimate θ (based on $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$). Let us denote then

$$\begin{aligned} H_\theta^*(x) &= \lambda_N F^*(x) + (1-\lambda_N)G^*(x) \\ &= \lambda_N Q_1(F(x)) + (1-\lambda_N)Q_2(F(x+\theta)) \end{aligned} \tag{5.2}$$

and since Q_1, Q_2 are non-decreasing, it follows that $H_\theta^*(x)$ is also non-decreasing in θ. Also, $P\{Y_i + \theta \le x\} = G^*(x-\theta) = Q_2(F(x))$, $\forall\ i \ge 1$. Thus, $T_N(X_1,\ldots,X_m,Y_1+\theta,\ldots,Y_n+\theta)$ has the same distribution as of $T_n = T(X_1,\ldots,X_m,Y_1,\ldots,Y_n)$ under $\theta = 0$. On the otherhand, if $a_N(1) \le \ldots \le a_N(N)$, then the rank of X_i among $X_1,\ldots,X_m$, $Y_1+a,\ldots,Y_n+a$ (denoted by $R_{Ni}(a)$) is $\searrow$ in $a(-\infty < a < \infty)$ for every $1 \le i \le m$. Hence, $T_N(a) = T_N(X_1,\ldots,X_m, Y_1+a,\ldots,Y_n+a)$ is also $\searrow$ in a. Thus, if we denote the right hand side of (2.6) by μ_{N0}, then by alignment, we

may consider the following estimator. Let

$$\hat{\theta}_{N,1} = \sup\{a;\ T_N(a) > \mu_{N0}\} \tag{5.3}$$

$$\hat{\theta}_{N,2} = \inf\{a;\ T_N(a) > \mu_{N0}\}\ . \tag{5.4}$$

The proposed estimator is

$$\hat{\theta}_N = \frac{1}{2}\,(\hat{\theta}_{N,1} + \hat{\theta}_{N,2})\ , \tag{5.5}$$

As in the case of the classical two-sample problem, $\hat{\theta}_N$ is a translation-invariante, robust and consistent estimator of θ. Further by virtue of Theorem 1 and the asymptotic simplifications made in Section 4, the proof of Theorem 4 of Hodges and Lehmann [3] can be directly adapted and this yields that under the assumptions made in Sections 3 and 4, as $N \to \infty$

$$N^{\frac{1}{2}}(\hat{\theta}_N - \theta) \sim N(0, \{\bar{\sigma}_{12}^2/\lambda + \bar{\sigma}_{20}^2/(1-\lambda)\}/B^2(F,Q_1,Q_2,\phi,\lambda)) \tag{5.6}$$

where $B(F,Q_1,Q_2,\phi,\lambda)$ is defined by (4.3). For instance, in Example 1, if we use the Wilcoxon scores (i.e., $\phi(u) \equiv u$), we obtain from (3.19), (3.20), (4.3) and (5.6) that

$$N^{\frac{1}{2}}(\hat{\theta}_N - \theta) \sim N(0, \gamma_{k\ell}^2)\ , \tag{5.6}$$

where

$$\gamma_{k\ell}^2 = \left[\frac{a}{(a+1)^2}\left\{\frac{a}{\lambda(2a+1)} + \frac{1}{(1-\lambda)(a+2)}\right\}\Big/\left(k\ell\int_{-\infty}^{\infty} f^2(x)\,(1-F(x))^{k+\ell-2}\,dx\right)^2\right]$$

$$= \left[\frac{1}{k\ell\,(k+\ell)^2}\left\{\frac{k}{\lambda(2k+\ell)} + \frac{\ell}{(1-\lambda)(k+2\ell)}\right\}\Big/\left\{\int_{-\infty}^{\infty} f^2(x)\,(1-F(x))^{k+\ell-2}\,dx\right\}^2\right].$$

Similar expressions can be derived for the second example. In this context, we may refer to Brown [1] for some related estimation problems based on the maximum likelihood procedure. A natural question arises whether $\gamma_{k\ell}$ is a minimum for $k = \ell = 1$? In general, it need not be.

6. SOME GENERAL REMARKS

By virtue of (1.3), it is intuitively appealing to consider a Kolmogorov-Smirnov type test statistic:

$$D_N = \sup_x N^{\frac{1}{2}} |G_n^*(x) - Q_0(F_m^*(x))| \tag{6.1}$$

(or the one-sided case) where the empirical df's are defined by (3.1). With the χ_r defined in (2.3), we have

$$D_N = N^{\frac{1}{2}} \left\{ \max_{1\le i\le N} \left| \frac{1}{N} \sum_{j=1}^{i} \chi_j - Q_0\left[\frac{1}{m} \sum_{j=1}^{i} (1-\chi_j)\right] \right| \right\} . \tag{6.2}$$

As such by using (2.2) and the one-to-one correspondence between $\underset{\sim}{R}_N$ and $\{\chi_r,\ 1 \le r \le N\}$, the small sample distribution of D_N can be obtained (under H_0). For large m, n this becomes quite complicated. In fact, if we consider a stochastic process $W_N = \{W_N(u),\ 0 \le u \le 1\}$ by letting

$$W_N(u) = N^{\frac{1}{2}}[\{G_n^*(x) - G^*(x)\} - \{Q_0(F_m^*(x)) - Q_0(F^*(x))\}]_{u=F^*(x)} \tag{6.3}$$

then under H_0 in (1.2),

$$D_N = \sup_{0\le u\le 1} |W_N(u)| . \tag{6.4}$$

It can be shown by standard steps that under H_0 in (1.2),

$$W_N \xrightarrow{\mathcal{D}} W = \{W(u),\ 0 \le u \le 1\}\ ,\quad \text{as}\quad N \to \infty\ , \tag{6.5}$$

where W is Gaussian with $EW(u) = 0,\ 0 \le u \le 1$ and

$$EW(u)W(v) = \lambda^{-1} q_0(u) q_0(v) u(1-v) + (1-\lambda)^{-1} Q_0(u)[1-Q_0(v)] \tag{6.6}$$

for $0 \le u \le v \le 1$. In general (for $Q_0(u) \not\equiv u$), (6.6) differs from the covariance structure of a Brownian bridge, and hence, the asymptotic distribution theory of the classical two-sample Kolmogorov-Smirnov statistic does not hold in our case. For W, the boundary crossing probabilities for general Q_0 are not precisely known, and hence, the prospect of using D_N as a large sample test statistic does not appear to be very bright. A similar criticism applies to the Cramér-von Mises type test based on $\|G_n^*(x) - Q_0(F_m^*(x))\|$.

It may also be intuitively appealing to use $\int Q_0(F_m^*(x))dG_n^*(x)$ (or more generally, $\int \phi(Q_0(F_m^*(x))dG_n^*(x))$ as a test statistic. The development of the asymptotic distribution theory of such a statistic poses no serious problem and can be made by the usual expansion of F_m^* and G_n^* around F^* and G^* respectively, and then using the Gaussian nature of the functions $\sqrt{N}\,[F_m^*-F^*]$ and $\sqrt{n}\,[G_n^*-G]$; the techniques are very similar to the ones employed in Section 3.6 of Puri and Sen [6] and hence, the details are omitted.

REFERENCES

[1] Brown, G.H., "Estimation from smallest values," *Communications in Statistics* 7 (1978), to appear.

[2] Hájek, J., "Asymptotic normality of simple linear rank statistics," *Annals of Mathematical Statistics* 39 (1968), 324-46.

[3] Hodges, J.L. Jr., and Lehmann, E.L., "Estimates of location based on rank tests," *Annals of Mathematical Statistics* 34 (1963), 598-611.

[4] Hoeffding, W., "On the centering of a simple linear rank statistic," *Annals of Statistics* 1 (1973), 54-66.

[5] Lehmann, E.L., The power of rank tests, *Annals of Mathematical Statistics* 24 (1953), 23-43.

[6] Puri, M.L., and Sen, P.K., *Nonparametric Methods in Multivariate Analysis*. John Wiley: New York, 1971.

[7] Sen, P.K., "On the estimation of relative potency in dilution (direct) essays by distribution-free methods," *Biometrics* 19 (1953), 532-52.