SELECTING THE POPULATION WITH THE SMALLEST DISPERSION IN A NONPARAMETRIC SETTING

Milton Sobel
University of California, Santa Barbara
Santa Barbara, California 93106

ABSTRACT

A nonparametric formulation is set up for selecting the best one of k populations where best is defined as the one with the smallest inter(α,β)-range; here inter(α,β)-range is a measure of dispersion defined by the difference of the β^{th} quantile and the α^{th} quantile. The formulation is strictly nonparametric in the sense that the df's are only assumed to be continuous and are not assumed to be stochastically ordered. The formulation and solution are similar to the solution of the corresponding "central tendency" problem treated by Sobel in [5], except that tables have not been prepared.

Appendix A gives a second-order correction term for the probability of a correct selection. Appendix B deals with a related problem of selecting a subset containing the best population and is similar to the solution of the corresponding "central tendency" problem treated by Rizvi and Sobel in [4].

1. INTRODUCTION

There are available k populations with completely unknown (cumulative) distribution functions $F_i(x)$ (or simply F_i) $(i = 1,\ldots,k)$; a pair of numbers α,β are also given with $0 < \alpha < \frac{1}{2} < \beta < 1$ with neither α nor β close to $\frac{1}{2}$. If $x_\alpha(F_i)$ denotes the α^{th} percentile of F_i then $Q_i = Q_{\alpha,\beta}(F_i) = x_\beta(F_i) - x_\alpha(F_i)$ for the given pair (α,β) denotes the inter(α,β)-range of F_i; our principal interest is in $\alpha = 1 - \beta = \frac{1}{4}$ in which case this is the familiar interquartile range. Based on n independent observations from each of the k populations, our goal is to select the "best" population where "best" means that it has the smallest inter(α,β)-range, i.e., it has the smallest Q-value. The df's $F_i(x)$ are each assumed to be continuous in x but are otherwise completely unknown and they need not belong to a common subfamily nor need they have a common support. If either $x_\alpha(F_i)$ or $x_\beta(F_i)$ is not unique then we define, say $x_\alpha(F_i)$, as the midpoint of the set $\{x: F_i(x) = \alpha\}$ and the theory and results of this paper still hold. However we assume below that all quantiles used are uniquely defined in order to avoid cumbersome notation that does not add to the basic ideas.

If we let $F_{[i]}(x) = F_{[i]}$ denote the df with the i^{th} smallest Q-value and use the notation $F_{[i]} \succ F_{[j]}$ to mean that $Q_i \leq Q_j$, then the correct ordering of the k distribution is

$$F_{[1]} \succ F_{[2]} \succ \cdots \succ F_{[k]}. \tag{1.1}$$

No additional information is assumed to be available at the outset concerning the correct pairing of the F_i with the $F_{[j]}$.

As a result of the more general setting that we use, and of the fact that each of two specified quantities ϵ^* and d^* defined below usually doesn't have a comparable parameter in other models, it is difficult to compare our results with the results of other weaker nonparametric formulations. The present format of posing and solving this nonparametric problem is a natural extension of the problem treated in [5] where the goal was to select that one of k populations which has the largest α-quantile, for fixed prespecified α. Thus it is shown by the present paper that (completely) nonparametric ranking procedures can be derived not only for "central tendency" problems but also for "dispersion" problems.

In Appendix B we consider the related problem of selecting a subset of the k populations that

contains the best one, i.e., the one with the smallest Q-value. This bears the same relation to the main problem of this paper that the paper [4] bears to the paper [5], namely the problems are different and separate tables are needed but certain technical details of the analysis are similar and can be handled simultaneously.

The nonparametric solution in [5] was also used in [2] to select fixed-size subsets that contain the best population.

A more detailed exposition of the problem treated in [5] is given in a separate expository paper [Appendix C] written as a companion to the present paper.

2. FORMULATION OF THE PROBLEM: REQUIREMENT AND P(CS)

For any fixed number n of observations let $S_{i\alpha}$ and $S_{i\beta}$ denote the α^{th} and β^{th} sample quantiles from F_i and let $S_i = S_{i\beta} - S_{i\alpha}$ denote the sample inter(α,β)-range. Our procedure will be to simply select the population that gives rise to the smallest value of S_i.

We now formulate a requirement for this problem so that by imposing a procedure that satisfies this requirement we can control the confidence we have in the solution obtained. Let F_0 denote the best population, i.e., the one with the smallest Q-value. If in some well-defined sense the $k-1$ worst populations (which we call the W-set) are sufficiently more disperse than F_0, then we want the probability of a correct selection P(CS) to be at least P^* (preassigned); we now make this more precise. Let $\epsilon^* > 0$ be specified so that the closed intervals $[\alpha-\epsilon^*, \alpha+\epsilon^*]$ and $[\beta-\epsilon^*, \beta+\epsilon^*]$ are disjoint and do not include 0, $\frac{1}{2}$ or 1, i.e., so that $0 < \epsilon^* < \min(\alpha, \frac{1}{2}-\alpha, 1-\beta, \beta-\frac{1}{2})$. We use F_0 and ϵ^* to define the two disjoint closed intervals

$$(2.1) \qquad I_1 = \left[x_{\alpha-\epsilon^*}(F_0), x_{\alpha+\epsilon^*}(F_0) \right], \qquad I_2 = \left[x_{\beta-\epsilon^*}(F_0), x_{\beta+\epsilon^*}(F_0) \right].$$

To separate the W-set from F_0 we suppose that each of the $k-1$ populations in the W-set is larger than F_0 throughout I_1 and smaller than F_0 throughout I_2. Let d denote the minimum vertical distance throughout both intervals between F_0 and any member of the W-set. Then our requirement states that if $d \geq d^*$ (the value of $d^* > 0$ is preassigned), then we require that $P(CS) \geq P^*$, where $\frac{1}{k} < P^* < 1$. Thus ϵ^*, d^* and P^* are all preassigned and the pair (α,β) is given by the problem; our principal interest is in $\alpha = 1-\beta = \frac{1}{4}$ and we usually take $\epsilon^* = d^*$ in the computations. The problem remaining is to find the smallest common sample size n from each of the k populations that will satisfy the above requirement. For the inequalities given on ϵ^*, d^*, P^*, α, β such an n will always exist.

The above formulation makes use of a nested configuration in the sense that if $d > 0$ then the entire interval $J(F_0) = [x_\alpha(F_0), x_\beta(F_0)]$ is included in the corresponding intervals for the $k-1$ worst populations and hence $Q(F_0) \leq Q(F_i)$ for all i. If this nested configuration does not hold or if $d < d^*$ then our requirement states nothing and hence the P(CS) would have to be evaluated for each such configuration for which the P(CS) is desired; the methods used below would of course be useful.

In carrying out the procedure based on S_i we assume that α is rational and restrict our attention to n-values for which $(n+1)\alpha$ and $(n+1)\beta$ are integers. Thus for $\alpha = 1-\beta = \frac{1}{4}$ we take $n+1$ to be a multiple of 4, i.e., $n = 3,7,11,\ldots$.

To derive the P(CS) for the least favorable configuration (LFC), we let X_0, Y_0 denote respectively the $(n+1)\alpha^{th}$ and $(n+1)\beta^{th}$ order statistics from F_0 and we use X_j and Y_j for the same order statistics from F_j $(j = 1,2,\ldots,k-1)$. In the LFC we have for each F_j in the W-set

$$(2.2) \qquad \begin{aligned} F_j(x) &= F_0(x) + d^* \quad \text{for each } x \text{ in } I_1 \\ F_j(y) &= F_0(y) - d^* \quad \text{for each } x \text{ in } I_2 \end{aligned}$$

Then the $P(CS \mid LFC)$ satisfies the inequality

$$(2.3) \qquad P(CS|LFC) \geq P\left\{X_j < X_0 < Y_0 < Y_j \quad (j = 1,2,\ldots,k-1)\right\}$$

$$= P\left\{F_j(X_j) < F_j(X_0), \quad F_j(Y_j) > F_j(Y_0) \quad (j = 1,2,\ldots,k-1)\right\}$$

$$= P\left\{F_j(X_j) < F_0(X_0) + d^*, \quad F_j(Y_j) > F_0(Y_0) - d^* \quad (j = 1,2,\ldots,k-1)\right\} .$$

Letting $U_j = F_j(X_j)$ and $V_j = F_j(Y_j)$ $(j = 0,1,\ldots,k-1)$ and dropping unnecessary subscripts, we make use of well-known results about order statistics to obtain

$$(2.4) \qquad P(CS|LFC) \geq \int_{\beta-\epsilon^*}^{\beta+\epsilon^*}\int_{\alpha-\epsilon^*}^{\alpha+\epsilon^*}\left[C\int_{v_0-d^*}^{1}\int_{0}^{u_0+d^*} u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}dudv\right]^{k-1}$$

$$\cdot\, C u_0^{r-1}(v_0-u_0)^{s-r-1}(1-v_0)^{n-s}du_0dv_0 ,$$

where $r = (n+1)\alpha$, $s = (n+1)\beta$ and C is given by

$$(2.5) \qquad C = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(s-r)\Gamma(n-s+1)} = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} .$$

3. ASYMPTOTIC EVALUATION OF THE $P(CS \mid LFC)$

As defined above (for each j), the chance variables U and V are respectively the r^{th} and s^{th} order statistics from a uniform distribution $U(0,1)$ and they have the following moments for $r \leq s$

$$(3.1) \qquad E(U) = \frac{r}{n+1}, \quad E(V) = \frac{s}{n+1}, \quad E(UV) = \frac{r(s+1)}{(n+1)(n+2)}$$

$$Cov(U,V) = \frac{r(n-s+1)}{(n+1)^2(n+2)}, \quad Corr(U,V) = \sqrt{\frac{r(n-s+1)}{s(n-r+1)}} .$$

Note that the ordering $r < s$ is essential and that the results also hold for $r = s$. Let $\alpha = \frac{r}{n+1}$, $\beta = \frac{s}{n+1}$. Then it is well known (see e.g. [1]) that

$$(3.2) \qquad X = \frac{(U-\alpha)\sqrt{n+2}}{\sqrt{\alpha(1-\alpha)}}, \quad Y = \frac{(V-\beta)\sqrt{n+2}}{\sqrt{\beta(1-\beta)}}$$

asymptotically $(n \to \infty)$ have a joint bivariate normal distribution with zero means, unit variances and correlation $\rho = \sqrt{\alpha(1-\beta)}/\sqrt{\beta(1-\alpha)}$. For the special case $\alpha = 1 - \beta = \frac{1}{4}$ we have $r = (n+1)/4$, $s = 3(n+1)/4$, $\alpha(1-\alpha) = \beta(1-\beta) = \frac{1}{4}(\frac{3}{4})$ and $\rho = \frac{1}{3}$. Note that ρ is exactly the same for small sample theory as for the asymptotic theory and does not tend to zero.

Let U_j, V_j be the U, V random variables and let X_j, Y_j be the X, Y random variables for F_j $(j = 0,1,\ldots,k-1)$. Then for each j the pair (X_j,Y_j) are asymptotically joint bivariate normal with independence between pairs. If we let $c = d^*\sqrt{n+2}/\sqrt{\alpha(1-\alpha)}$ and $c' = d^*\sqrt{n+2}/\sqrt{\beta(1-\beta)}$ then the limits of integration for the inside integrals in (2.4) are

$$(3.3) \qquad X_j < X_0 + c, \quad Y_j > Y_0 - c' \quad (j = 1,2,\ldots,k-1).$$

Hence we obtain from (2.4) by letting S denote the asymptotic $(n \to \infty)$ limit of the right hand side of (2.4)

$$S = \int\int \left[\int_{y-c}^{\infty}, \int_{-\infty}^{x+c} \varphi(u,v \mid \rho)\,dudv\right]^{k-1} \varphi(x,y \mid \rho)\,dxdy , \tag{3.4}$$

where the subscripts on x_j, y_j $(j = 0,1,\ldots,k-1)$ have been dropped, unmarked limits of integration are from $-\infty$ to ∞, and

$$\varphi(x,y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}\right\} \quad (-\infty < x,y < \infty). \tag{3.5}$$

Using the Taylor expansion of $\varphi(x,y \mid \rho)$ about $\rho = 0$ and using $\varphi^{(\alpha)}(x)$ to denote the α^{th} derivative of the standard normal density $\varphi(x)$ (so that $\varphi^{(0)}(x) = \varphi(x)$), we have

$$\varphi(x,y|\rho) = \sum_{\alpha=0}^{\infty} \varphi^{(\alpha)}(x)\varphi^{(\alpha)}(y)\frac{\rho^\alpha}{\alpha!} = \varphi(x)\varphi(y) \sum_{\alpha=0}^{\infty} H_\alpha(x)H_\alpha(y)\frac{\rho^\alpha}{\alpha!}, \tag{3.6}$$

where $H_\alpha(x)$ is the Hermite polynomial with respect to the (standard) normal density kernel, i.e., $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, etc. Denote the bracketed double integral in (3.4) by $B(x,y \mid \rho)$ and let $\Phi(x)$ denote the standard normal df. After replacing y by $-y$ we obtain from (3.4) and (3.6)

$$B(x,-y \mid \rho) = \Phi(x + c)\Phi(y + c') + \sum_{\alpha=1}^{\infty} \varphi^{(\alpha-1)}(x + c)\varphi^{(\alpha-1)}(y + c')\frac{(-\rho)^\alpha}{\alpha!} . \tag{3.7}$$

We use S_i to denote the terms up to and including ρ^i; here we only consider the "linear" result S_1 but if later numerical evaluations show a need we shall also bring in the terms involving ρ^2 and consider S_2. From (3.4), (3.6) and (3.7) we have for S and S_1, respectively,

$$S = \int\int [\Phi(x+c)\Phi(y+c') - \rho\varphi(x+c)\varphi(y+c')]^{k-1}[\varphi(x)\varphi(y) - \rho\varphi^{(1)}(x)\varphi^{(1)}(y)]dxdy, \tag{3.8}$$

$$\begin{aligned} S_1 = {} & \int\int \Phi^{k-1}(x + c)\Phi^{k-1}(y + c')[\varphi(x)\varphi(y) - \rho\varphi^{(1)}(x)\varphi^{(1)}(y)]dxdy \\ & - (k - 1)\rho\int\int \Phi^{k-2}(x + c)\Phi^{k-2}(y + c')\varphi(x + c)\varphi(y + c')\varphi(x)\varphi(y)dxdy . \end{aligned} \tag{3.9}$$

We wish to express these 3 double integrals in terms of A-functions defined with 3 (varying) arguments k', ρ', h' by

$$A_{k'}(\rho',h') = \int \Phi^{k'-1}\left(\frac{x\sqrt{\rho'} + h'}{\sqrt{1 - \rho'}}\right)\varphi(x)dx , \tag{3.10}$$

where $\rho' > 0$ is usually the reciprocal of an integer, h' is usually nonnegative and $k' \geq 1$ is an integer. A straightforward completion of squares and integration-by-parts gives from (3.9)

$$S_1 = A_k(\tfrac{1}{2},h_1)A_k(\tfrac{1}{2},h_1') - \binom{k}{2}\rho\varphi(h_1)\varphi(h_1')A_{k-1}(\tfrac{1}{3},h_2)A_{k-1}(\tfrac{1}{3},h_2'), \tag{3.11}$$

where $h_1 = \frac{c}{\sqrt{2}}$, $h_1' = \frac{c'}{\sqrt{2}}$, $h_2 = \frac{c}{\sqrt{6}}$, $h_2' = \frac{c'}{\sqrt{6}}$. For the special case $\alpha = 1 - \beta = \frac{1}{4}$ we have $c = c'$, $h_i = h_i'$ $(i = 1,2)$, $\rho = \frac{1}{3}$ and (3.11) reduces to

$$S_1 = A_k^2(\tfrac{1}{2},h_1) - \frac{k(k-1)}{6}\varphi^2(h_1)A_{k-1}^2(\tfrac{1}{3},h_2). \tag{3.12}$$

Note that for $k = 2$ this gives a very simple result, namely

$$S_1 = A_2^2(\tfrac{1}{2},h_1) - \tfrac{1}{3}\varphi^2(h_1)A_1^2(\tfrac{1}{3},h_2) = \Phi^2(h_1) - \tfrac{1}{3}\varphi^2(h_1), \tag{3.13}$$

since $A_1(\rho,h) \equiv 1$ for all ρ and h.

Tables of the A-functions needed for the above are available in the R. Milton Table [3] for many values of k, ρ and h; note that for $k = 2$ we only need a standard normal (df and density) table.

To illustrate the calculations we first consider $k = 2$ and use (3.13). For $\alpha = 1 - \beta = \frac{1}{4}$, $k = 2$, $P^* = .90$ and $\epsilon^* = d^* = .1$ the use of (3.13) gives a trial h_1-value, i.e.,

$$\Phi(h_1) = \sqrt{P^*} = .949 \quad \text{or} \quad h_1 = 1.63. \tag{3.14}$$

We now add $\frac{1}{3}\Phi^2(1.63)$ to $.9$ and solve $\Phi^2(h_1) = .9037$ for a new value of h_1. One or two iterations gives the result $h_1 = 1.650$. Since $h_1 = d^*\sqrt{n+2}/\sqrt{2(\frac{1}{4})(\frac{3}{4})}$, we obtain $n = \frac{3}{8}(\frac{1.650}{.1})^2 - 2 = 100.1$, so that 101 observations are needed from each of the $k = 2$ populations to satisfy the requirement. (The omission of the 1^{st} order correction term in (3.13) gives $n = 98$, but the second order correction term derived in Appendix A indicates that our "linear" result $n = 101$ is generally both conservative and close to the correct answer; a second-order correction in Appendix A reduces this to $n = 100$.)

For $k > 2$ it is again useful to solve (3.12) by first assuming that the second term is not present; the correction term needed is usually very small. For example, with $k = 5$ and α, β, P^*, ϵ^*, d^* all as in the above illustration, our first trial value is $h_1 = c/\sqrt{2} = 2.15$ and the first term in (3.12) is

$$A_5^2(\tfrac{1}{2}, 2.15) = (.94879)^2 = .90020, \tag{3.15}$$

and, if this were not adjusted, we would need

$$n = 2\alpha(1-\alpha)\left(\frac{h_1}{d^*}\right)^2 - 2 = \frac{3}{8}(\frac{2.15}{.1})^2 - 2 = 173.3 - 2 = 171.3, \tag{3.16}$$

where the first equality in (3.16) is a general result for the case $\alpha = 1 - \beta$. Thus without correcting h_1 we need 172 observations from each of the $k = 5$ populations. The total value of S_1 in (3.12) for this h_1 is .8973, so that the correction term is only .003 and the terms involving ρ^2 are not needed. If we now use $h_1 = 2.20$ the value of S_1 in (3.12) is

$$S_1(\text{for } h_1 = 2.20) = (.95443)^2 - \tfrac{10}{3}(.03547)^2(.7573)^2 = .9095 > .9. \tag{3.17}$$

By linear interpolation between these 2 results we obtain $h_1 = 2.162$ and hence by (3.16) we find that $n = 173.28$ so that 174 observations are needed from each of the $k = 5$ populations to satisfy the requirement, i.e., a total of $5(174) = 870$ observations.

If we rank the $k = 5$ populations by considering a comparison of 2 populations, repeated four times, then the total number of observations might at first appear to be smaller than 870 since each pairing requires 200 observations and the four repetitions would require a total of 800 observations. However if we use $P^* = .9$ for each pairing then the overall P(CS) lower bound would be only $(.9)^4 = .656$. Moreover, any attempts to avoid taking new observations on the winner of each pairing would tie up the analysis with a low P(CS) that is difficult to estimate. Thus, our above result, 870, does not appear to be excessively large for the general nonparametric setting we are using. As further evidence of this, suppose we used $P^* = .974$ for each pairwise experiment, so that the overall P(CS) lower bound is $(.974)^4 = .900$. Then for each $k = 2$ experiment we obtain $h_1 = 2.229$ and by (3.16) we need $n = 185$ observations per population. Hence this method requires a total of $8(185) = 1480$ observations, which is substantially larger than our total of 870.

4. FURTHER RESEARCH NEEDED THAT IS RELEVANT TO THIS PROBLEM.

It would be desirable to investigate the effect of taking $\epsilon \neq d^*$ to see if the number of observations required is more sensitive to ϵ^* or to d^*. It would also be desirable to have an explicit table in which the user can look up the n-value corresponding to specified values of P^*, ϵ^* and d^*; these are not yet available. It would also be desirable to compare numerically the solution described here with

that obtained in other seminonparametric models, e.g. with the model in which we assume that for some unknown df $F(x)$ all the populations are given by $F(\frac{x-\theta}{\sigma_j})$, θ being a common unknown location parameter and the scale parameters σ_j being the parameters by which the populations are to be ordered. Here the difficult question is how to find quantities comparable with ϵ^* and d^* in the other models so that meaningful comparisons can be made. Finally it would be desirable to generalize the present solution for selecting the $t = 1$ best so that we could also consider (nonparametrically) the problem of selecting the t best populations for $t > 1$, where "best" is again defined by having the smallest inter(α,β)-range.

APPENDIX A

Although the correction terms of order ρ^2 were not needed above, they may not always be negligible and we may want to see how small they are and whether they indicate that our solution above is usually conservative. Since a slightly more general expression arises for a related problem in Appendix B below, it is useful to generalize (3.4) and (3.8) slightly by putting $bx + c$ instead of $x + c$, $b'x + c'$ instead of $x + c'$, ρ_1 for ρ inside $B(x,y\,|\,\rho)$ and ρ_0 for ρ outside $B(x,y\,|\,\rho)$; hence we obtain our result by setting $b = b' = 1$ $\rho_0 = \rho_1$ in the final result. For convenience, we also set $B = b^2$, $B' = (b')^2$ and define for $j = 1,2$

$$\text{(A1)}\qquad F_j = \frac{A_{k-j}\left(\frac{B}{1+(j+1)B}, \frac{c}{\sqrt{(1+jB)[1+(j+1)B]}}\right)}{\sqrt{1+jB}} \prod_{\alpha=1}^{j} \varphi\left(\frac{c}{\sqrt{[1+(\alpha-1)B](1+\alpha B)}}\right),$$

where the A-function is defined by (3.10) and F_j' is defined similarly with b,B,c replaced by b',B',c', respectively. Thus to get a numerical value for F_j and F_j' we only need [3] and a standard normal (density and df) table. There are four terms, T_i $(i = 1,2,3,4)$ in (3.8) involving ρ_0 and ρ_1 to the second order, namely

$$\text{(A2)}\qquad \begin{aligned}
T_1 &= \frac{\rho_0^2}{2}\int \Phi^{k-1}(bx+c)\varphi^{(2)}(x)dx \int \Phi^{k-1}(b'y+c')\varphi^{(2)}(y)dy,\\
T_2 &= \rho_0\rho_1(k-1)\int \varphi(x)\Phi^{k-2}(bx+c)\varphi(bx+c)dx \int \varphi(y)\Phi^{k-2}(b'y+c')\varphi(b'y+c')dy,\\
T_3 &= \frac{\rho_1^2}{2}(k-1)\int \varphi(x)\Phi^{k-2}(bx+c)\varphi^{(1)}(bx+c)dx \int \varphi(y)\Phi^{k-2}(b'y+c')\varphi^{(1)}(b'y+c')dy,\\
T_4 &= \rho_1^2\left(\frac{k-1}{2}\right)\int \varphi(x)\Phi^{k-3}(bx+c)\varphi^2(bx+c)dx \int \varphi(y)\Phi^{k-3}(b'y+c')\varphi^2(b'y+c')dy.
\end{aligned}$$

Using completion of the square and integration-by-parts the resulting second-order correction term $\mathfrak{I}_2$ can be expressed as

$$\text{(A3)}\qquad \mathfrak{I}_2 = c_{22}F_2F_2' + c_{21}F_2F_1' + c_{12}F_1F_2' + c_{11}F_1F_1',$$

where F_j and F_j' are defined in (A1),

$$\text{(A4)}\qquad \begin{aligned}
c_{22} &= \binom{k-1}{2}\left\{\rho_1^2 + \frac{(k-2)bb'}{(1+B)(1+B')}\left[(k-1)bb'\rho_0^2 + 2\rho_0\rho_1 + \rho_1^2 bb'\right]\right\},\\
c_{21} &= \binom{k-1}{2}\frac{c'}{1+B}\left\{B\rho_1^2 - \frac{bb'}{1+B'}\left[(k-1)bb'\rho_0^2 + 2\rho_0\rho_1 + \rho_1^2 bb'\right]\right\},\\
c_{11} &= \left(\frac{k-1}{2}\right)cc'\left\{\rho_1^2\left(1 - \frac{B}{1+B} - \frac{B'}{1+B'}\right) + \frac{bb'}{(1+B)(1+B')}\left[(k-1)bb'\rho_0^2 + 2\rho_0\rho_1 + \rho_1^2 bb'\right]\right\},
\end{aligned}$$

and c_{12} is the same as c_{21} except that (B,c) are interchanged with (B',c'), respectively. For the special case $b = b' = 1$ and $\rho_0 = \rho_1 = \rho$ (say), we have $B = B' = 1$ and

(A5) $$c_{22} = \frac{k^2(k-1)(k-2)}{8}\rho^2, \quad c_{21} = -\frac{k(k-1)(k-2)}{8}cc'\rho^2, \quad c_{11} = \frac{(k-1)(k+2)}{8}cc'\rho^2,$$

and c_{12} is the same as c_{21} with c' replaced by c. In particular, for $k = 2$ we obtain $c_{22} = c_{21} = c_{12} = 0$ and $c_{11} = \frac{cc'}{2}\rho^2 = h_1h_1'\rho^2$, so that (3.13) with $\alpha = 1 - \beta$ and $\rho = \frac{1}{3}$, becomes

(A6) $$S_2 = \Phi^2(h_1) - \tfrac{1}{3}\varphi(h_1) + (\tfrac{1}{3})^2\tfrac{1}{2}h_1^2\varphi^2(h_1);$$

note that J_2 in (A6) can also be written as $\frac{1}{2}\rho^2[\varphi^{(1)}(h_1)]^2$.

In the numerical example above for $k = 2$ the correction term J_2 adds .0016 to the P(CS) lower bound. If we use this term to iterate for a corrected h_1, we obtain $h_1 = 1.642$, which yields the result $n = \frac{3}{8}(\frac{1.642}{.1})^2 - 2 = 99.1$. Hence we need 100 observations from each of the $k = 2$ populations to satisfy the requirement. Thus our previous result of $n = 101$ was both conservative and close to the correct answer.

APPENDIX B

A similar analysis can be used for the problem of selecting a subset of the k populations that contains the "best" population, where "best" is again defined in terms of having the smallest Q-value, where (α,β) are given by the problem.

For the given (α,β), let $S_i^{(\alpha,\beta)}$ denote the sample inter(α,β)-range from the population π_i which has cdf F_i $(i = 1,2,\ldots,k)$; let $S_{min}^{(\alpha,\beta)}$ denote the smallest of these k scalar quantities. For α',β' such that $\alpha < \alpha' < \frac{1}{2}$ and $\frac{1}{2} < \beta' < \beta$ (In the main application of interest we set both $\alpha = 1 - \beta$ and $\alpha' = 1 - \beta'$.), we put π_i in the selected subset if and only if

(B1) $$S_i^{(\alpha',\beta')} \le S_{min}^{(\alpha,\beta)}.$$

Clearly the population giving rise to $S_{min}^{(\alpha,\beta)}$ gets into the selected subset, so that the subset is never empty. We wish to find a maximal interval (α',β') with $\alpha < \alpha'$ and $\beta' < \beta$ such that we can assert that $P(CS) \ge P^*$; of course, if $\beta' = 1 - \alpha'$ then we are looking for the smallest α' (with $\alpha < \alpha'$) having the same property. Letting $j = 0$ denote the best population, the P(CS) for the LFC is given by

(B2) $$P(CS \mid LFC) = P\left\{S_0^{(\alpha',\beta')} \le S_{min}^{(\alpha,\beta)}, \text{ where min is over } j \ne 0\right\}$$
$$\ge P\left\{X_\alpha^{(j)} \le X_{\alpha'}^{(0)}, \; Y_{\beta'}^{(0)} \le Y_\beta^{(j)} \quad (j = 1,2,\ldots,k-1)\right\}.$$

Let $\varepsilon > 0$ be such that $\alpha < \alpha' < \alpha + \varepsilon$ and $\beta - \varepsilon < \beta' < \beta$ and such that the two closed intervals

(B3) $$I_1 = [x_\alpha(F_0), x_{\alpha+\varepsilon}(F_0)], \qquad I_2 = [x_{\beta-\varepsilon}(F_0), x_\beta(F_0)]$$

are disjoint. To separate F_j $(j = 1,2,\ldots,k-1)$ from F_0 we assume that

(B4) $$F_0(x) \le F_j(x) \quad \text{for all } x \text{ in } I_1,$$
$$F_0(x) \ge F_j(x) \quad \text{for all } x \text{ in } I_2.$$

Then in the LFC we have equality in (B4) in the corresponding intervals. Using this, we can continue the chain of inequalities in (B2) to obtain

(B5) $$P(CS \mid LFC) \ge P\left\{F_j(X_\alpha^{(j)}) \le F_0(X_{\alpha'}^{(0)}), F_0(Y_{\beta'}^{(0)}) \le F_j(Y_\beta^{(j)}) \; (j = 1,2,\ldots,k-1)\right\}$$
$$= \int_0^1\int_0^v \left[\int_v^1\int_0^u cw^{r'-1}(z-w)^{s'-r'-1}(1-z)^{n-s'}dydz\right]^{k-1} cu^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}dudv,$$

where $r' = (n+1)\alpha'$, $s' = (n+1)\beta'$, C is given by (2.5) and C' is the same with (r,s) replaced by (r',s'), respectively.

For the asymptotic theory the limits $w < u$ and $v < z$ in (B5) translate into the inequalities

$$x_j < bx_0 + c, \quad y_j > b'y_0 - c' \quad (j = 1,2,\ldots,k-1), \tag{B6}$$

where b, b', c and c' are all positive and given by

$$b = \sqrt{\frac{\alpha'(1-\alpha')}{\alpha(1-\alpha)}}, \quad b' = \sqrt{\frac{\beta'(1-\beta')}{\beta(1-\beta)}}, \quad c = \frac{(\alpha'-\alpha)\sqrt{n+2}}{\sqrt{\alpha(1-\alpha)}}, \quad c' = \frac{(\beta-\beta')\sqrt{n+2}}{\sqrt{\beta(1-\beta)}}. \tag{B7}$$

Then the asymptotic $(n \to \infty)$ result S for the right side of (B5) is

$$S = \iint \Big[\int_{b'y-c'}^{\infty} \int_{-\infty}^{bx+c} \varphi(w,z \mid \rho_1)\,dw\,dz \Big]^{k-1} \varphi(x,y \mid \rho_0)\,dx\,dy, \tag{B8}$$

where ρ_0 is given by (3.1) in terms of r and s and ρ_1 simply uses (r',s') in place of (r,s), respectively.

Again using (3.6) (twice) in (B8), we obtain for S

$$\begin{aligned} S &= \iint B^{k-1}(x,y \mid \rho_1)\varphi(x,y \mid \rho_0)\,dx\,dy = \iint B^{k-1}(x,-y \mid \rho_1)\varphi(x,-y \mid \rho_0)\,dx\,dy \\ &= \iint B^{k-1}(x,-y \mid \rho_1)\varphi(x,y \mid -\rho_0)\,dx\,dy \\ &= \iint \Big[\Phi(bx+c) + \Phi(b'y+c') + \sum_{\alpha=1}^{\infty} \varphi^{(\alpha-1)}(bx+c)\varphi^{(\alpha-1)}(b'y+c') \frac{(-\rho_1)^{\alpha}}{\alpha!} \Big]^{k-1} \Big[\sum_{\beta=0}^{\infty} \varphi^{(\beta)}(x)\varphi^{(\beta)}(y) \frac{(-\rho_0)^{\beta}}{\beta!} \Big] dx\,dy. \end{aligned} \tag{B9}$$

Letting $B = b^2$, $B' = (b')^2$ and using S_1 to denote the constant and linear terms in ρ_1 and ρ_0, we now obtain from (B9) after algebraic manipulation

$$\begin{aligned} S_1 &= \int \Phi^{k-1}(bx+c)\varphi(x)\,dx \int \Phi^{k-1}(b'y+c')\varphi(y)\,dy \\ &\quad - (k-1)[\rho_1 + \rho_0 bb'(k-1)] \int \varphi(x)\Phi^{k-2}(bx+c)\varphi(bx+c)\,dx \int \varphi(y)\Phi^{k-2}(b'y+c')\varphi(b'y+c')\,dy \\ &= A_k\Big(\frac{B}{1+B}, \frac{c}{\sqrt{1+B}}\Big) A_k\Big(\frac{B'}{1+B'}, \frac{c'}{\sqrt{1+B'}}\Big) - \frac{(k-1)[\rho_1 + \rho_0 bb'(k-1)]}{\sqrt{1+B}\sqrt{1+B'}} \varphi\Big(\frac{c}{\sqrt{1+B}}\Big)\varphi\Big(\frac{c'}{\sqrt{1+B'}}\Big) \cdot \\ &\quad \cdot A_{k-1}\Big(\frac{B}{1+2B}, \frac{c}{\sqrt{(1+B)(1+2B)}}\Big) A_{k-1}\Big(\frac{B'}{1+2B'}, \frac{c'}{\sqrt{(1+B')(1+2B')}}\Big), \end{aligned} \tag{B10}$$

and the second order correction terms are directly obtainable from (A3) and (A4) in Appendix A. In terms of α, β, α' and β' the values of ρ_0 and ρ_1 are

$$\rho_0 = \sqrt{\frac{\alpha(1-\beta)}{\beta(1-\alpha)}}, \quad \rho_1 = \sqrt{\frac{\alpha'(1-\beta')}{\beta'(1-\alpha')}}. \tag{B11}$$

Here again the first order terms will typically be small and hence the second order correction terms will generally not be needed.

As a numerical example suppose $k = 5$ and $\alpha = 1 - \beta = \frac{1}{4}$ and we assume that $\alpha' = 1 - \beta'$, so that $\rho_0 = \frac{1}{3}$ and we want to find the smallest value of α' (with $\alpha < \alpha'$) for which the $P(CS) \geq P^*$; say, $P^* = .90$. Assume a common sample size of $n = 100$ observations from each of the $k = 5$ populations. Suppose we try $\alpha' = .35$, $\beta' = .65$. Then $b = b' = 1.10$, $c = c' = 4(.1)\sqrt{34} = 2.33$, $B = B' = 1.21$,

$B/(1+B) = B'/(1+B') = .55$, $c/\sqrt{1+B} = c'/\sqrt{1+B'} = 1.56$. Then the leading term in (B10) is (to 2 decimal places)

(B12) $$A_5^2(.55,1.56) = (.838)^2 = .70,$$

which is much too small. On the other hand with $\alpha' = .40$, $\beta' = .60$ we obtain $b = b' = 1.13$, $c = c' = 3.50$, $B = B' = 1.28$, $B/(1+B) = B'/(1+B') = .56$, $c/\sqrt{1+B} = 2.32$, so that

(B13) $$A_5^2(.56,232) = (.967)^2 = .935$$

which is slightly too large since the correction term in (B10) is of the order of magnitude of .002. The value $\alpha' = .39$, $\beta' = .61$ leads to the value $.908 - .003 = .905$, so that $\alpha' = .39$ is a conservative solution for the requirement that we set. Hence our procedure based on $n = 100$ observations and $k = 5$ is to put into the selected subset exactly those populations whose sample inter(.39,.61)-range is less than the smallest of the five sample interquartile ranges.

APPENDIX C

(EXPOSITION OF) SELECTING THE POPULATION WITH THE LARGEST α-QUANTILE IN A NONPARAMETRIC SETTING.

by Milton Sobel

In the paper [5] a solution is given for the nonparametric problem of selecting those t out of k populations which have the largest α-quantiles. The case $\alpha = \frac{1}{2}$ corresponds to the problem of finding exactly which t of the k given populations have the largest population medians. For $t = 1$ tables are provided in [3] that give the smallest odd number n of observations required per population to satisfy a lower bound probability requirement on the probability of a correct selection; this requirement has to hold when the best one is separated from the $k - 1$ worst populations in a certain well-defined sense. The justification for taking an odd number from each population is (i) simply to have the sample median well defined and (ii) the error is at most one extra observation per population.

To explain the above formulation it suffices to consider the case $t = 1$. Let F_0 denote the best distribution function (df), i.e., the one with the largest population α-quantile. Let $\epsilon^* > 0$ be specified so that $0 < \alpha - \epsilon^*$ and $\alpha + \epsilon^* < 1$. (Actually in [5] we consider two values ϵ_1^*, ϵ_2^* but we now take them equal and use ϵ^* to denote the common value.) In terms of ϵ^* and F_0 we define the closed interval

(C1) $$I = [x_{\alpha-\epsilon^*}(F_0), x_{\alpha+\epsilon^*}(F_0)],$$

where $x_\alpha(F)$ denotes the α^{th} quantile of the df F.

In order to separate the df's F_j $(j = 1,2,\ldots,k-1)$ from F_0 we first assume for each j $(j = 1,2,\ldots,k-1)$ that $F_0(x) < F_j(x)$ for all x in I. Let d denote the minimum difference of $F_j(x) - F_0(x)$ over all j $(j = 1,2,\ldots,k-1)$ and all x in I. Our requirement states that when $d \geq d^*$ we want the procedure to have a probability of a correct selection $P(CS)$ of at least P^*. Since the procedure is simply to take a common number n of observations from each of the k populations and select the one with the largest sample α-quantile, we have only to determine the smallest value of n that will satisfy the above requirement. Here $d^* > 0$ and P^* (with $\frac{1}{k} < P^* < 1$) and ϵ^* are all specified and, of course, α is given by the problem. In the computations we take $\epsilon^* = d^*$ and solve for the smallest n such that $(n+1)\alpha$ is an integer.

As a numerical illustration suppose we have $k = 5$ populations and want to find the one with the largest median $(\alpha = \frac{1}{2})$. Suppose that $d^* = \epsilon^* = .10$ and $P^* = .90$. Then by Table 3 of [5], using the first of the four entries, in the cell we find that $n = 169$ observations are needed from each of the 5 populations to satisfy the above requirement.

The second entry in each Table of [5] gives the corresponding n-value if we make the stronger assumption that $F_0(x) \leq F_j(x)$ for all real x and each j $(j = 1,2,\ldots,k-1)$. This adds very little to the assumption that $d \geq d^*$, since the first two entries are essentially the same throughout the tables. In particular, in the above illustration the answer is again $n = 169$.

In another formulation (Formulation 2A) in [5] we simply consider the difference

$$(C2) \qquad d' = \min_{1 \le j \le k-1} \left\{ x_{\alpha-\epsilon^*(F_0)} - x_{\alpha+\epsilon^*(F_j)} \right\}$$

and the requirement states that when $d' \ge 0$ we want to have a $P(CS) \ge P^*$. The required value of n is the third entry in each cell of Table 3. Finally the fourth entry again makes the stronger assumption that for each j $(j = 1,2,\ldots,k-1)$ $F_j(x) \ge F_0(x)$ for all real x. the third and fourth entries are generally quite close throughout Table 3. Thus, in the above illustration with $P^* = .90$ and $d^* = \epsilon^* = .10$ the third and fourth entries are $n = 103$ and $n = 101$, respectively.

APPENDIX D

COMMENTS ON THE PAPER BY DR. G. McDONALD

It is unfortunate that Dr. McDonald was not aware of three papers on nonparametric ranking. One in particular by Rizvi and Sobel deals specifically with the related problem of selecting a subset containing the one with the largest α-quantile. Another paper deals with selecting the t best populations and the third deals with nonparametric procedures for selecting a subset of specified size s which includes the t populations that have the largest α-quantiles $(t \le s \le k-1)$. Since the Rizvi-Sobel paper is more pertinent to the present discussion I restrict my discussion to that one. The exact reference is:
[4] Rizvi, M. H. and Sobel, M. (1967) Nonparametric procedures for selecting a subset containing the population with the largest α-quantile Ann. of Math Statist. 38 1788-1803.

I was asked to apply the procedure in [4] to Dr. McDonald's data on Motor Vehicle Fatality rates (by state and year) and report on the results. The procedure can be described in terms of Y_{si}, which deontes the s^{th} order statistic from the df F_i; we assume a constant number of observations n from each F_i $(i = 1,2,\ldots,k)$ and, of course, all observations are (mutually) independent. Take $r = (n+1)\alpha$, so that for $\alpha = \frac{1}{2}$ we have $r = (n+1)/2$ and $Y_{r,i}$ is the sample median from F_i $(i = 1,2,\ldots,k)$. We define $Y_{0,i}$ to be $-\infty$ and $Y_{n+1,i}$ to be $+\infty$ $(i = 1,2,\ldots,k)$. The procedure in [4] is to put F_i in the selected subset if and only if

$$(D1) \qquad Y_{r,i} \ge \max_{1 \le j \le k} Y_{r-c,j}$$

where c (with $1 \le c \le r-1$) is an integer to be determined; c is the smallest integer for which the $P(CS) \ge P^*$. If we had written $1 \le c \le r$ then a value of c would always exist (even for $P^* = 1$), but for the indicated range $[1, r-1]$ a value of c exists provided P^* is chosen not greater than some function $P_1 = P_1(n,\alpha,k)$ where $\frac{1}{k} < P_1 < 1$. Since $P_1 \to 1$ rapidly as n increases, this restriction is not serious and hence by assuming that $c < r$ we can avoid with probability one the trivial strategy of putting all the populations in the selected subset regardless of what observations are obtained. (It should be noted that even if $P_1 < P^* < 1$, the above procedure still "works" by randomizing between the nondegenerate rule (D1) and the degenerate rule, $c = r$.) From the first expression in (3.6) of [4] which is equivalent to the sum of the first two terms of the rapidly converging alternating sum on the right hand side of (3.5) in [4]), we find that for $k = 49$ and $n = 11$, 13, 15 and 17 the approximate values of P_1 are .70 .87, .95, and .98, respectively; here again we assume n is odd so that the median is simply defined. Hence for $P^* = .90$ (and also for $P^* = .95$) with $k = 49$ we can avoid the trivial strategy by simply checking to see that $n \ge 17$. Since $n = 17$, $k = 49$ and $P^* = .90$, we are in "good shape" to apply the procedure in [4] to Dr. McDonald's data.

The exact asymptotic formulas for the $P(CS)$ are given in [4] along with tables of $(r-c)$-values to carry out the procedure for moderate values of k and n.

For $k = 49$ and $n = 17$ we may start by using the asymptotic formula (5.2) of [4], namely

$$(D2) \qquad \liminf_{n=\infty} P(CS) = \int_{-\infty}^{\infty} \Phi^{k-1}\left(y + \frac{\gamma}{\sqrt{\alpha(1-\alpha)}} \right) \varphi(y)dy,$$

where $\Phi(x)$, $\varphi(x)$ are the standard normal df and density, respectively and $\gamma = c\sqrt{n}/(n+1)$ is to be determined by setting the right side of (D2) equal to the specified P^*. From known tables we find (by interpolation) for $\alpha = \frac{1}{2}$ and $k = 50$ that $\gamma = 1.83$. Setting $n = 17$, we then find that $c = 7.99$

and $r - c = 9 - 7.99 = 1.01$. The nearest integer is 1 and since this result changes slowly with k at $k = 50$ we also take $r - c$ to be 1 for $k = 49$. Note that this result is consistent with the results for the small k tabulated under $P^* = .900$ in Table 3 of [4]. The exact calculation needed to confirm this result is obtained from (3.3) of [4] and we would like to see whether

$$(D3) \qquad \frac{(17)!}{8!8!}\int_0^1 [1 - 17u(1-u)^{16} - (1-u)^{17}]^{48} u^8 (1-u)^8 du$$

is less than .90 and with the middle term in the brackets removed the value is greater than .90. The values obtained are .914 and .982 respectively. (See discussion second paragraph of the next page)

For $r - c = 1$ the procedure is to compute the $\max_{(\text{state})} \min_{(\text{year})}$ constant of the data and put F_i in the selected subset if its median is equal to or greater than that constant. For the data of Dr. McDonald we obtain $\max \min = 5.3$, which happens to come from the state of New Mexico. There are 25 states (including New Mexico) with sample median equal to or greater than 5.3 and this is the size of the desired subset. It includes all the states taken in by R_1 and is included among the states taken in by R_2. Note that large α-quantiles correspond to "bad" (or "worst") populations in this example, although they may have been called "best" populations in [4].

For the dual problem of selecting a subset containing the one with the smallest α-quantile, it follows from Section 6 of [4] that for $\alpha = \frac{1}{2}$ we can use the same value of c together with the rule

$$(D4) \qquad \text{"Put } F_i \text{ in the selected subset if and only if } Y_{r,i} \leq \min_{1 \leq j \leq k} Y_{r+c,j}\text{"}.$$

Note that $r + c \leq 2r - 1 = n$, so that with probability one the solution is again not degenerate if we can use the same c-value. From the data for $r - c = 1$ we search for the $\min_{(\text{state})} \max_{(\text{year})}$ constant which is 3.2, and happens to come from the state of Conneticut. There are 3 states (Conneticut, New Jersey and Rhode Island) with sample medians not exceeding 3.2 and these constitute the desired subset.

Several interesting features can now be pointed out for the procedure in [4], although it should be noted that some of these (namely 1, 2 and 3 below) are, strictly speaking, only conjectures.

1. For the same P^* the subsets from the procedure in [4] will generally be smaller in size (as in this example).
2. The two subsets from [4] will generally be disjoint (as in this example).
3. The subsets obtained by [4] will generally be proper subsets of much smaller size than those of procedures R_2 and R_2' and will be in general agreement with the ordering of the states by Rank Sums (as it turned out in this example).
4. The procedure in [4] does not assume that the unknown df's are stochastically ordered.
5. The solution in [4] is <u>strictly</u> nonparametric in the usual wide sense, i.e., in the <u>whole</u> space of all <u>continuous</u> distribution functions and <u>not just</u> in a "small" subspace characterized by the slippage of a parameter θ.
6. There are <u>no parameters</u> in the <u>nonparametric</u> formulation in [4].
7. Even with $c < r$ it is possible under the procedure of [4] that we will put all the populations in the related subset. In practice this will happen very rarely. Note also that if every state has exactly the same data (each year) then we <u>prefer</u> to include every state in both subsets.
8. Under the procedure in [4] the selected subset can never be empty (in either of the two problems). However under procedure R_2 (and also R_2') it seems clear that this can happen. True, it may be improbable, but the probability of this happening should be controlled at least for some special configurations of interest. Has this been investigated?

It should also be pointed out that the two problems above have been considered separately with distinct probability requirements and using the very same data. It would be much more desirable to consider them as a single problem and to control the overall joint probability that the first subset contains the worst population <u>and</u> that the second subset contains the best population. (This logical dilemma gets even more profound and confusing in Dr. McDonald's paper where four procedures are simultaneously used on the very same data.) Even if n were large, $\alpha > \frac{1}{2}$ for the subset to contain the worst population, $\alpha' < \frac{1}{2}$ for the subset to contain the best population, and the two subsets obtained

were disjoint, it would still be difficult to argue that we could treat these two problems as being independent of each other. I presume that the proposed joint problem has not been considered and I recommend it as a challenging problem for some bright young student interested in ranking and selection.

In Appendix B above we apply nonparametric ideas similar to those in [4] to the problem of selecting a subset of the k populations which contains the one with the smallest interquartile range. As a secondary problem it may be useful and interesting to apply this new procedure to Dr. McDonald's data; this has not yet been done, partly because it would add to the logical dilemma of using a P^*-formulation for different problems and then applying these formulations to the very same data.

Since the value of n is only 17 in Dr. McDonald's data, it is necessary to check up on the asymptotic theory by calculating the integral in (D6) above taken from (3.3) of [4]. Surprisingly, it turns out to have the value $.914 > .900$ and hence the solution given above can be improved for $P^* = .900$ by taking $r - c = 2$ (rather than 1); this must give subsets that are of the same size or smaller than those obtained above. Note that the results above do hold for $P^* = .950$ since in that case we have to use $r - c = 1$. To confirm that $r - c = 2$ for $P^* = .900$ we also compute the $P(CS \mid LFC)$ from (3.3) in [4] for $r - c = 3$, to see if it is $< .900$, namely

$$(D5)\quad P(CS/LFC) = \frac{17!}{8!8!}\int_0^1 \left[1 - 136u^2(1-u)^{15} - 17u(1-u)^{16} - (1-u)^{17}\right]^{48} u^8(1-u)^8 du,$$

and obtain the value $.785 < .900$; for these calculations ad hoc quadrature methods were used together with the incomplete beta function table (of Karl Pearson).

Applying this result ($r - c = 2$) for $P^* = .900$ to the data we find that the subset size (for containing the worst state) is reduced from 25 to 17 and that it contains Vermont, Kentucky and the last 15 states in Table 4 of Dr. McDonald's paper. Thus the result is similar to the subset selected by procedure R_1. For the dual problem with $P^* = .900$ and $r - c = 2$ the subset size (for containing the best state) is reduced from three to two, namely either Conneticut or Rhode Island is the best state with confidence $P^* = .900$. Thus our subset sizes are much smaller than those obtained by the method of Rank Sums. However the results we obtain are in general agreement with the ordering of the states that arises by the usage of Rank Sums, since e.g. we include the last 15 states of Table 4 in one subset of size 17 and we include the top 2 states of Table 4 in the other subset.

It may be of some technical interest to point out that for k near 49, the usage of $k = 49 \pm 1$ or 49 ± 2 will not affect the second decimal of our calculations for P_1 in (3.5) of [4] or for $P(CS \mid LFC)$ in (3.3) of [4]. Hence any calcualtions above that were made with $k = 51$ instead of $k = 49$ will still be correct to two decimal places. Further, if someone wants to add or omit certain states from the analysis, most of the analytical and sample calcualtions are still useful under the procedure of [4], whereas the rank sums needs a whole new calculation if we add or subtract a single state.

I would like to thank Dr. McDonald for bringing this interesting set of data to our attention.

REFERENCES

[1] H. A. David (1970). Order Statistics. John Wiley and Sons Inc., New York. p. 201.

[2] M. M. Desu and Milton Sobel (1970). Nonparametric procedures for selecting fixed-size subsets. A contribution to Statistical Decision Theory and Related Topics Edited by S. S. Gupta and J. Yackel. Academic Press, New York. 255-273.

[3] R. Milton (1963). Tables of the Equally Correlated Multivariate Normal Probability Integral, T.R. #27 Department of Statistics, University of Minnesota, Minneapolis, Minnesota.

[4] M. Haseeb Rizvi and Milton Sobel (1967). Nonparametric procedures for selecting a subset containing the population with the largest α-quantile, Ann. of Math. Statist. 38 1788-1803.

[5] M. Sobel (1967). Nonparametric procedures for selecting the t populations with the largest α-quantities, Ann. of Math. Statist. 38 1804-1816.