

NONPARAMETRIC PROCEDURES FOR SELECTING THE t POPULATIONS WITH THE LARGEST α -QUANTILES¹

BY MILTON SOBEL

University of Minnesota

1. Introduction. There are given k populations with cumulative distribution functions (cdf) $F_i(x) = F_i$ ($i = 1, 2, \dots, k$); a number α with $0 < \alpha < 1$ and an integer t with $1 \leq t \leq k - 1$ are preassigned. The goal is to select those t of the k populations which have the largest α -quantiles based on n independent observations from each of the k populations. The form of the cdf F_i is unknown but it is assumed that $F_i(x)$ is continuous in x ($i = 1, 2, \dots, k$). Two different formulations of this nonparametric problem will be given; for each formulation there will be a Case A and a Case B according to whether a certain assumption is made. A subset approach to this problem with a similar nonparametric formulation is considered in a companion paper [1]. This paper deals only with exact results; a discussion of the asymptotics for this paper will be published later in a separate paper.

Let $x_\alpha(F)$ denote the α th-quantile of the distribution F ; if this quantile is not unique we can define it as any arbitrary point (or any convex combination of points) in the closure of the set $\{x: F(x) = \alpha\}$. However, we shall assume that each F_i has a unique α -quantile in order to avoid extra notation which does not add to the basic ideas of the problem. Let $F_{[i]}(x) = F_{[i]}$ denote the cdf with the i th smallest α -quantile. The correct ordering of the k distributions is

$$(1.1) \quad F_{[1]} \leq F_{[2]} \leq \dots \leq F_{[k]}$$

where $F_{[i]} \leq F_{[j]}$ means that $x_\alpha(F_{[i]}) \leq x_\alpha(F_{[j]})$. It is assumed that no *a priori* information (in the form of a distribution or otherwise) is available concerning the correct pairing of the F_i and the $F_{[i]}$. We do not assume that the k distributions have the same supports, nor do we require that they differ only in a location parameter, i.e., the problem and the solution are nonparametric.

It is clear that if the $F_i(x)$ are ordered uniformly in x , i.e., if there are no crossovers; then the problem of selecting the t smallest of the $F_i(x)$ is equivalent to the problem of selecting the t populations with the largest α -quantiles.

FORMULATION 1. Let $\epsilon_\gamma^* > 0$ ($\gamma = 1, 2$) be two specified numbers such that $\epsilon_1^* \leq \alpha \leq 1 - \epsilon_2^*$. Let $x_\beta(F)$ and $\bar{x}_\beta(F)$ denote the infimum and supremum of the set $\{x; F(x) = \beta\}$. Unless stated otherwise, the indices i, j will run over the ranges $i = 1, 2, \dots, k - t$ and $j = k - t + 1, k - t + 2, \dots, k$, respectively. Let $\underline{F} = \underline{F}(x)$ denote the min $F_{[i]}(x)$ and $\bar{F} = \bar{F}(x)$ denote the max $F_{[j]}(x)$. Define the closed interval

$$(1.2) \quad I = [\bar{x}_{\alpha-\epsilon_1^*}(\bar{F}), x_{\alpha+\epsilon_2^*}(\bar{F})].$$

Received 16 November 1965; revised 10 July 1967.

¹ Supported by NSF under Grant No. 3813.

Figure for Formulation 1

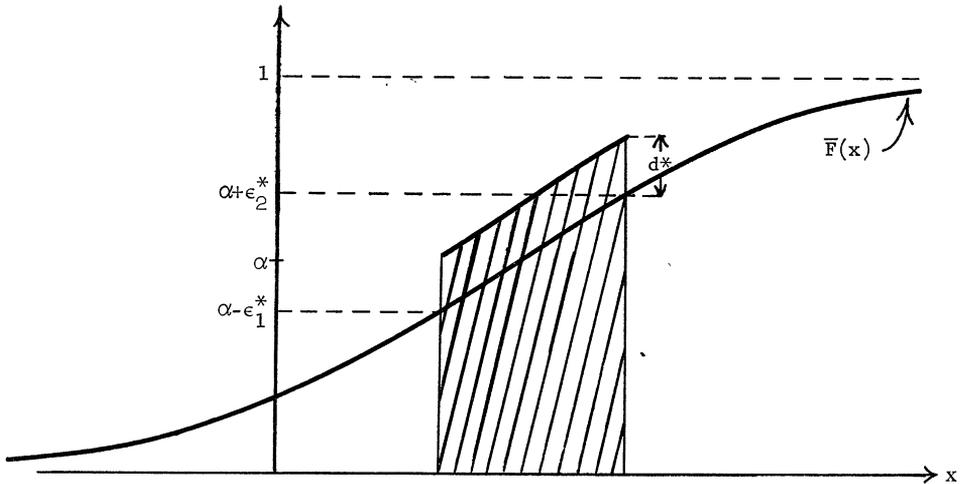


FIGURE for Formulation 1. To be in the preference zone (i.e., in the complement of the indifference zone) each $F_{[i]}(x)$ ($i = 1, 2, \dots, k - t$) must avoid the shaded area, i.e., $\underline{F}(x)$ must avoid the shaded area.

Let $d_{ij} = \inf (F_{[i]}(x) - F_{[j]}(x))$ over all $x \in I$ and let

$$(1.3) \quad d = \min_{(i,j)} d_{ij} = \inf_{x \in I} (\underline{F}(x) - \bar{F}(x)).$$

Finally let d^* and P^* denote specified constants with $d^* > 0$ and $1 > P^* > 1/\binom{k}{t}$. We define $F_{[k-t]} < F_{[k-t+1]}$ to mean that $x_\alpha(F_{[k-t]}) < x_\alpha(F_{[k-t+1]})$ and when this is the case we define a correct selection (CS) to mean the selection of those t populations with the largest α -quantiles; we shall not need the definition of a CS when $F_{[k-t]} = F_{[k-t+1]}$.

Our goal is to find a procedure R which satisfies the probability requirement

$$(1.4) \quad P\{CS | R\} \geq P^* \quad \text{when} \quad d \geq d^*,$$

where $d^* > 0$ and P^* , with $1 > P^* > 1/\binom{k}{t}$, are preassigned constants. Any vector or point (F_1, F_2, \dots, F_k) for which $d < d^*$ for at least one pair (i, j) is said to be in the "indifference zone" and no probability requirement is made for such points. (See figure for Formulation 1.)

In order to define our basic order statistic in (1.6) below we have to assume that the common sample size n from each population is sufficiently large so that $1 \leq (n + 1)\alpha \leq n$. Let the positive integer r be defined by

$$(1.5) \quad r \leq (n + 1)\alpha < r + 1.$$

It follows from the above that $1 \leq r \leq n$. Yet $Y_{j,i}$ denote the j th order statistics from F_i ($j = r, r + 1; i = 1, 2, \dots, k$) and let

$$(1.6) \quad W_i = [r + 1 - (n + 1)\alpha]Y_{r,i} + [(n + 1)\alpha - r]Y_{r+1,i}.$$

The coefficients in (1.6) are based on a linear combination of two adjacent order statistics to approximate the point where the sample cdf (with slanted lines instead of horizontal lines) is equal to α .

We can now define the

PROCEDURE R. Take n independent observations from each population and select the t populations which give rise to the t largest W -values.

If α is rational then $(n + 1)\alpha$ will be an integer for some arithmetic progression N of n -values and then $r = (n + 1)\alpha$ and $W_i = Y_{r,i}$; for simplicity we shall assume that α is rational. For example, if $\alpha = \frac{1}{2}$ then N is the set of odd integers.

[In most practical problems α can be written as (or approximated by) a fraction with small denominator D . Then the maximum loss due to restricting our attention to n -values in N is that we take at most $D - 1$ extra observations per population. Clearly this is not a deficiency of the procedure R since, if necessary, we could (a) compute the $P\{CS | R\}$ for all integers, and/or (b) investigate whether any error arises in interpolating among the computed probabilities for $n \in N$ to decide what common sample size to use. For many D -values neither of these is necessary; for example, if $\alpha = \frac{1}{2}$ then $D - 1 = 1$ and the inefficiency is clearly negligible even for moderate n -values.]

Hence for our formulation and this procedure R the problem now reduces to finding the smallest integer n such that (1.4) is satisfied; we shall be content with finding the smallest n in N that satisfies (1.4). In the sequel we refer to the above as Formulation 1A and if a certain additional assumption (see (2.11) below) is made we call it Formulation 1B.

FORMULATION 2. Let $\epsilon_\gamma^* > 0$ ($\gamma = 1, 2$) be two specified numbers such that

Figure for Formulation 2

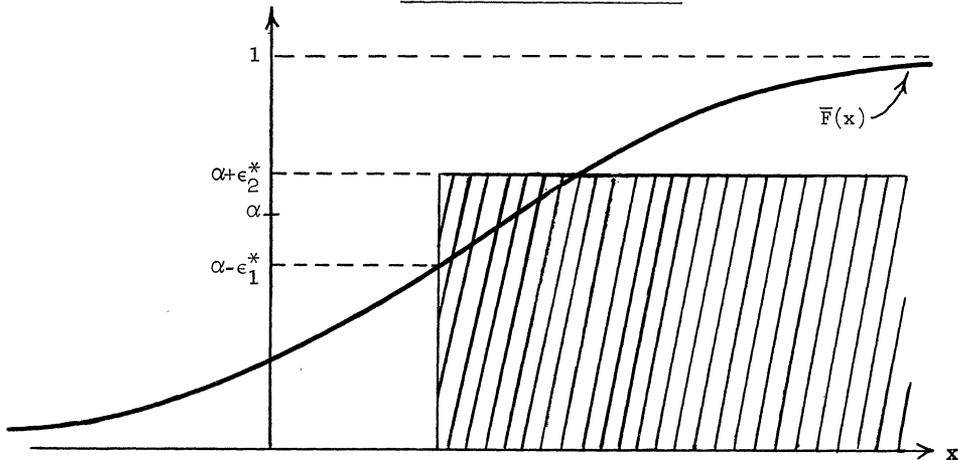


FIGURE for Formulation 2. To be in the preference zone (i.e., in the complement of the indifference zone) each $F_{[i]}(x)$ ($i = 1, 2, \dots, k - t$) must avoid the shaded area, i.e., $F(x)$ must avoid the shaded area.

$\epsilon_1^* \leq \alpha \leq 1 - \epsilon_2^*$. For $i = 1, 2, \dots, k - t$ and $j = k - t + 1, k - t + 2, \dots, k$, let $d'_{ij} = x_{\alpha - \epsilon_1^*}(F_{[i]}) - \bar{x}_{\alpha + \epsilon_2^*}(F_{[j]})$ and let

$$(1.7) \quad d' = \min_{(i,j)} d'_{ij} = x_{\alpha - \epsilon_1^*}(\bar{F}) - \bar{x}_{\alpha + \epsilon_2^*}(\underline{F}).$$

Our goal is to find a procedure R such that

$$(1.8) \quad P\{\text{CS} | R\} \geq P^* \quad \text{when} \quad d' \geq 0,$$

where P^* (with $1 > P^* > 1/\binom{k}{t}$) is a preassigned constant. (See figure for Formulation 2.) The form of the procedure is the same as for Formulation 1 above and the comments made there hold in this case also; in particular, the same assumption that α is rational will be made. Hence the problem again reduces to finding the smallest integer n such that (1.8) is satisfied. The vectors (or points) in the indifference zone are those for which $d'_{ij} < 0$ for at least one pair (i, j) . We refer to the above as Formulation 2A and if a certain additional assumption (see (2.11) below) is made we call it Formulation 2B.

The reason for introducing both ϵ_1^* and ϵ_2^* above is that the experimenter may want to make ϵ_1^* and ϵ_2^* unequal. In particular, he may want to make them proportional to α and $1 - \alpha$, respectively; for $\alpha = \frac{1}{2}$ this means that $\epsilon_1^* = \epsilon_2^* = \epsilon^*$ (say).

In the rest of this paper we find exact expressions for the $P\{\text{CS} | R\}$ for each formulation and these are then used to find the smallest integer n that satisfies the corresponding P^* -condition. Tables of exact n -values are given at the end of the paper for selected values of k, P^* , and $d^* = \epsilon_1^* = \epsilon_2^*$.

2. Probability of a correct selection. For each formulation we calculate a lower bound to the exact $P\{\text{CS} | R\}$ as a function of d for an arbitrary configuration and for the "least favorable" configuration, the latter being defined as the configuration that yields the infimum of the $P\{\text{CS}\}$ in the complement of the indifference zone.

2.1 $P\{\text{CS}\}$ for Formulation 1A. Let $Y_{r,(i)} = Y_{(i)}$ denote the r th order statistic associated with the distribution $F_{[i]}$. The probability element $dH_{r,i}(y) = dH_i(y)$ and the cdf $H_{r,i}(y) = H_i(y)$ of $Y_{(i)}$, respectively, are well known (or easily shown) to be

$$(2.1) \quad dH_i(y) = r \binom{n}{r} F_{[i]}^{r-1}(y) [1 - F_{[i]}(y)]^{n-r} dF_{[i]}(y),$$

$$(2.2) \quad H_i(y) = \sum_{j=r}^n \binom{n}{j} F_{[i]}^j(y) [1 - F_{[i]}(y)]^{n-j} = G[F_{[i]}(y)],$$

where $G(p) = G_{r,n-r+1}(p)$ is used for the standard Incomplete Beta function

$$G(p) = [n! / (r - 1)! (n - r)!] \int_0^p x^{r-1} (1 - x)^{n-r} dx.$$

The probability of a correct selection under the procedure R is given by

$$(2.3) \quad \begin{aligned} P\{\text{CS} | R\} &= P\{\max(Y_{(1)}, \dots, Y_{(k-t)}) < \min(Y_{(k-t+1)}, \dots, Y_{(k)})\} \\ &= \sum_{j=k-t+1}^k P\{\max(Y_{(1)}, \dots, Y_{(k-t)}) < Y_{(j)}\} \\ &= \min(Y_{(k-t+1)}, \dots, Y_{(k)}) \\ &= \sum_{j=k-t+1}^k \int_{-\infty}^{\infty} \prod_{i=1}^{k-t} H_i(y) \prod_{\beta=k-t+1, \beta \neq j}^k [1 - H_{\beta}(y)] dH_j(y) \\ &\geq \sum_{j=k-t+1}^k \int_I + \sum_{j=k-t+1}^k \int_{I^+} = T_1 + T_2 \quad (\text{say}), \end{aligned}$$

where I^+ is the infinite open interval (say, (a, ∞)) to the right of I . We now make use of the well-known result that $G(y)$ is a continuous nondecreasing function of y and the fact that this property holds by assumption for $F_{[i]}(y)$ for each i ; using this it follows from (2.2) that the same property holds for $H_i(y)$ for each i . On I^+ it follows from the definition of d and ϵ_2^* that $F_{[i]}(y) \geq \alpha + \epsilon_2^* + d$ for each i and $F_{[j]}(a) \leq \alpha + \epsilon_2^*$ for each j ; hence $H_i(y) \geq G(\alpha + \epsilon_2^* + d)$ for each i and $H_j(a) \leq G(\alpha + \epsilon_2^*)$ for each j . Bounding one set of factors in (2.3) and integrating the other set we now obtain

$$\begin{aligned} T_2 &\geq G^{k-t}(\alpha + \epsilon_2^* + d) \int_{I^+} \sum_{j=k-t+1}^k \prod_{\beta=k-t+1, \beta \neq j}^k [1 - H_\beta(y)] dH_j(y) \\ &= G^{k-t}(\alpha + \epsilon_2^* + d) [1 - \prod_{j=k-t+1}^k (1 - H_j(y))] \Big|_a^\infty \\ &= G^{k-t}(\alpha + \epsilon_2^* + d) [1 - G(\alpha + \epsilon_2^*)]^t. \end{aligned}$$

Let $\underline{H}(y)$ and $\bar{H}(y)$ be defined as in (2.2) with $F_{[i]}(y)$ replaced by $\underline{F}(y)$ and $\bar{F}(y)$, respectively. Since $H_i(y) \geq \underline{H}(y)$ we can bound T_1 below by writing

$$(2.4) \quad T_1 \geq \sum_{j=k-t+1}^k \int_I \underline{H}^{k-t}(y) \prod_{\beta=k-t+1, \beta \neq j}^k [1 - H_\beta(y)] dH_j(y)$$

but in order to justify replacing $H_\beta(y)$ by $\bar{H}(y)$ in (2.4) we need another expression for $P\{\text{CS} | R\}$. Starting with the same first line as in (2.3) and proceeding as above we can also bound the $P\{\text{CS} | R\}$ by writing it in the form

$$\begin{aligned} P\{\text{CS} | R\} &= \sum_{i=1}^{k-t} P\{\max(Y_{(1)}, \dots, Y_{(k-t)}) \\ &= Y_{(i)} < \min(Y_{(k-t+1)}, \dots, Y_{(k)})\} \\ (2.5) \quad &= \sum_{i=1}^{k-t} \int_{-\infty}^{\infty} \{\prod_{\gamma=1, \gamma \neq i}^{k-t} H_\gamma(y) \prod_{j=k-t+1}^k [1 - H_j(y)]\} dH_i(y) \\ &\geq \sum_{i=1}^{k-t} \int_I \{\prod_{\gamma=1, \gamma \neq i}^{k-t} H_\gamma(y)\} [1 - \bar{H}(y)]^t dH_i(y) \\ &\quad + G^{k-t}(\alpha - \epsilon_1^* + d) [1 - G(\alpha - \epsilon_1^*)]^t, \end{aligned}$$

where the second term now comes from the infinite open interval to the left of I . We note from (2.5) that to minimize the $P\{\text{CS} | R\}$ we have to set $H_j(y)$ equal to $\bar{H}(y)$ for $y \in I$; we proceed to do this now in (2.4). Combining the results for T_1 and T_2 above we obtain

$$(2.6) \quad \begin{aligned} P\{\text{CS} | R\} &\geq t \int_I \underline{H}^{k-t}(y) [1 - \bar{H}(y)]^t d\bar{H}(y) \\ &\quad + G^{k-t}(\alpha + \epsilon_2^* + d) [1 - G(\alpha + \epsilon_2^*)]^t. \end{aligned}$$

We now compute a lower bound for the $P\{\text{CS} | R\}$ at the least favorable configuration (lfc) by setting $d = d^*$ in (2.6) and by setting $\underline{F}(y) = \bar{F}(y) + d^*$. Hence, letting $u = \bar{F}(y)$,

$$(2.7) \quad \underline{H}(y) = G[\underline{F}(y)] = G(u + d^*),$$

$$(2.8) \quad \bar{H}(y) = G[\bar{F}(y)] = G(u).$$

Hence (2.6) takes the form

$$(2.9) \quad \begin{aligned} P\{\text{CS} | R\} &\geq t \int_{\alpha - \epsilon_1^*}^{\alpha + \epsilon_2^*} G^{k-t}(u + d^*) [1 - G(u)]^{t-1} dG(u) \\ &\quad + G^{k-t}(\alpha + \epsilon_2^* + d^*) [1 - G(\alpha + \epsilon_2^*)]^t. \end{aligned}$$

From (2.5) we now obtain the expression

$$(2.10) \quad P\{CS | R\} \geq (k - t) \int_{\alpha+d^*-\epsilon_1^*}^{\alpha+d^*+\epsilon_2^*} G^{k-t-1}(v)[1 - G(v - d^*)]^t dG(v) \\ + G^{k-t}(\alpha - \epsilon_1^* + d^*)[1 - G(\alpha - \epsilon_1^*)]^t,$$

which can also be obtained from (2.9) by an integration-by-parts.

It is interesting to note that in this formulation with fixed n and $d^* \rightarrow 0$ the right side of (2.9) approaches a lower bound which is less than or equal to $1/\binom{k}{t}$, with equality holding only if $\alpha + \epsilon_2^* = 1$ and $\alpha - \epsilon_1^* = 0$. In fact, by differentiation it is easy to see that the right side of (2.9) is increasing in each of ϵ_1^* and ϵ_2^* . Hence the lower bound as $d^* \rightarrow 0$ must lie in the closed interval $\{G^{k-t}(\alpha)[1 - G(\alpha)]^t, 1/\binom{k}{t}\}$, which reduces to $\{2^{-k}, 1/\binom{k}{t}\}$ for $\alpha = \frac{1}{2}$.

For $n \rightarrow \infty$ and $(r/n) \rightarrow \alpha$ we will show below that the right side of (2.9) approaches one for any fixed $d^* > 0$. If simultaneously $d^* \rightarrow 0$ then the limit must be at least $1/\binom{k}{t}$; the latter follows from the fact that the limit ($n \rightarrow \infty$ and $(r/n) \rightarrow \alpha$) obtained after setting d^* equal to the 'inadmissible' value zero in the right side of (2.9) is $1/\binom{k}{t}$.

2.2. $P\{CS\}$ for Formulation 1B. A slight variation of the results in the last section is obtained if we make the assumption that for all x

$$(2.11) \quad F_{[i]}(x) \geq F_{[k-t+1]}(x) = \bar{F}(x) \quad (i = 1, 2, \dots, k - t), \\ F_{[j]}(x) \leq F_{[k-t]}(x) = \underline{F}(x) \quad (j = k - t + 1, k - t + 2, \dots, k).$$

For the special case $t = 1$ we omit the second part of (2.11) and for $t = k - 1$ we omit the first part. [It should be noted that we could have assumed that for all pairs (γ, δ) with $\gamma < \delta$

$$(2.12) \quad F_{[i]}(x) \leq F_{[j]}(x) \quad \text{for all } x,$$

but the assumption (2.11) is a weaker assumption and accomplishes the same purpose, namely to make $F_{[i]} = F_{[j]} = F$ (say) in the worst configuration.] Under the assumption (2.11) the derivation of the first term on the right side of (2.9) i.e., the integral, remains exactly the same as in (2.9) but the contribution from the complement of I is different. The final result for this formulation, using the same methods as above, is

$$(2.13) \quad P\{CS | R\} \geq t \int_{\alpha-\epsilon_1^*}^{\alpha+\epsilon_2^*} G^{k-t}(u + d^*)[1 - G(u)]^{t-1} dG(u) \\ + G^{k-t}(\alpha + \epsilon_2^* + d^*)\{[1 - G(\alpha + \epsilon_2^*)]^t - [1 - G(\alpha + \epsilon_2^* + d^*)]^t\} \\ + \binom{k}{t}^{-1}\{1 - G_{k-t+1,t}[G(\alpha + \epsilon_2^* + d^*)] + G_{k-t+1,t}[G(\alpha - \epsilon_1^*)]\}.$$

The explanation of the second term on the right side of (2.13) is as follows. In the j th term of the sum in (2.3) we let $u = F_{[j]}(y)$ and consider that portion of the integral for which $\alpha + \epsilon_2^* < u < \alpha + \epsilon_2^* + d^*$. As y approaches $y_1 = x_{\alpha+\epsilon_2^*}(F_{[k-t+1]})$ (the right end point of I) from within I , we have $F_{[j]}(y) \rightarrow \alpha + \epsilon_2^*$ and it follows that $F_{[i]}(y) \geq \alpha + \epsilon_2^* + d^*$, and hence by (2.2)

$$H_i(y_1) \geq G(\alpha + \epsilon_2^* + d^*) \quad \text{for } i = 1, 2, \dots, k - t.$$

Hence, letting I_1 denote this part of the integral in (2.3),

$$\begin{aligned}
 (2.14) \quad I_1 &\geq G^{k-t}(\alpha + \epsilon_2^* + d^*) \int_{y_1^-}^{y_1^+} \sum_{j=k-t+1}^k \prod_{\beta=k-t+1, \beta \neq j}^k [1 - H_\beta(y)] dH_j(y) \\
 &= G^{k-t}(\alpha + \epsilon_2^* + d^*) \{1 - \prod_{j=k-t+1}^k [1 - H_j(y)]\}_{y_1^-}^{y_1^+} \\
 &\geq G^{k-t}(\alpha + \epsilon_2^* + d^*) \{[1 - G(\alpha + \epsilon_2^*)]^t - [1 - G(\alpha + \epsilon_2^* + d^*)]^t\}.
 \end{aligned}$$

The last part of (2.13) represents the contributions of the intervals $\alpha + \epsilon_2^* + d^* \leq u \leq 1$ and $0 \leq u \leq \alpha - \epsilon_1^*$, respectively.

For $d^* = 0$ and any pair $\epsilon_1^* \geq 0$ and $\epsilon_2^* \geq 0$, it is easily seen that the right side of (2.13) reduces to $1/\binom{k}{t}$ for any fixed n .

It is natural to “conjecture” that the right side of (2.13) is greater than the right side of (2.9) and hence that Formulation 1B requires an n -value which is not larger than that required by Formulation 1A. This is indeed the case since the difference Δ of these quantities satisfies the inequality

$$(2.15) \quad \Delta \geq t \int_x^1 y^{k-t}(1-y)^{t-1} dy - x^{k-t}(1-x)^t \quad (0 \leq x \leq 1)$$

where $x = G(\alpha + \epsilon_2^* + d^*)$. Differentiating the right side of (2.15) we find that it is decreasing from $1/\binom{k}{t}$ to 0 as x increases from 0 to 1, so that $\Delta \geq 0$. Another proof of this “conjecture” could be based on the fact that the indifference zone under Formulation 1A is included in the indifference zone under Formulation 1B.

It can be shown that the expressions in (2.9) and (2.13) are asymptotically equivalent for large n (or, equivalently, for d^* close to zero) and it will be of some interest to see how much difference it makes in the sample size to make the assumption (2.11). Tables 1, 2, 3 and 4 show that this difference is quite small for $P^* \geq .75$ and that it decreases with increasing P^* and also with increasing k .

2.3. *The $P\{CS\}$ for Formulation 2A.* Using the third expression in (2.3) as a starting point, setting $H_i(y) = G(\alpha + \epsilon_2^*)$ for each i , integrating the resulting expression and letting $x_1 = \min_j \bar{x}_{\alpha-\epsilon_1^*}(F_{[j]})$ gives

$$\begin{aligned}
 (2.16) \quad P\{CS | R\} &\geq G^{k-t}(\alpha + \epsilon_2^*) \{1 - \prod_{j=k-t+1}^k [1 - H_j(y)]\}_{x_1}^\infty \\
 &= G^{k-t}(\alpha + \epsilon_2^*) \prod_{j=k-t+1}^k [1 - H_j(x_1)].
 \end{aligned}$$

Increasing every H_j as much as possible at x_1 we obtain $G(\alpha - \epsilon_1^*)$ for each of them and hence in the least favorable configuration

$$(2.17) \quad P\{CS | R\} \geq G^{k-t}(\alpha + \epsilon_2^*) [1 - G(\alpha - \epsilon_1^*)]^t.$$

It is easily seen that this tends to one as $n \rightarrow \infty$ and the remaining problem is to determine the smallest n for which the right side of (2.17) is at least P^* .

For the special case $\alpha = \frac{1}{2}$ we have $r = (n + 1)/2 = n - r + 1$ and, using the symmetry of the Incomplete Beta function, (2.17) becomes for $\epsilon_1^* = \epsilon_2^* = \epsilon^*$ (say)

$$(2.18) \quad P\{CS | R\} \geq G^k(\frac{1}{2} + \epsilon^*),$$

which does not depend on t .

2.4. *The P{CS} for Formulation 2B.* We now make the same assumption (2.11) as in Formulation 1B and, using the same methods as above, we obtain

$$\begin{aligned}
 (2.19) \quad P\{CS|R\} &\geq t \int_0^{\alpha - \epsilon_1^*} G^{k-t}(u)[1 - G(u)]^{t-1} dG(u) \\
 &\quad + t \int_{\alpha + \epsilon_2^*}^1 G^{k-t}(u)[1 - G(u)]^{t-1} dG(u) \\
 &\quad + G^{k-t}(\alpha + \epsilon_2^*)\{[1 - G(\alpha - \epsilon_1^*)]^t - [1 - G(\alpha + \epsilon_2^*)]^t\} \\
 &= \binom{k}{t}^{-1}[G_{k-t+1,t}(G(\alpha - \epsilon_1^*)) + 1 - G_{k-t+1,t}(G(\alpha + \epsilon_2^*))] \\
 &\quad + G^{k-t}(\alpha + \epsilon_2^*)\{[1 - G(\alpha - \epsilon_1^*)]^t - [1 - G(\alpha + \epsilon_2^*)]^t\}.
 \end{aligned}$$

It is natural to “conjecture” that the right side of (2.19) is greater than the right side of (2.17) and hence Formulation 2B requires an n which is not larger than that required by Formulation 2A. This is indeed the case and the proof is exactly the same as in Formulation 1. We also note that for $\epsilon_1^* = \epsilon_2^* = 0$ the right side of (2.19) reduces to $1/\binom{k}{t}$.

For $k = 2, t = 1$, and $\alpha = \frac{1}{2}$, Tables 1 through 4 give the smallest odd integers that will solve the problem for selected values of $d^* = \epsilon_1^* = \epsilon_2^*$ and P^* , for each of the four formulations, 1A, 1B, 2A and 2B.

3. Calculation of Tables. For $\alpha = \frac{1}{2}$ the expression for the Incomplete Beta function $G(u)$ can be simplified. In fact, if $n = 2m + 1$ then $r = \frac{1}{2}(n + 1) = n - r + 1 = m + 1$ and writing $G_{m+1}(u)$ for $G(u)$ we have, setting $y = x - \frac{1}{2}$

$$\begin{aligned}
 (3.1) \quad G_{m+1}(u) &= [(2m + 1)!/(m!)^2] \int_0^u (x - x^2)^m dx \\
 &= [(2m + 1)!/(m!)^2 2^{2m}] \int_{-\frac{1}{2}}^{u-\frac{1}{2}} (1 - 4y^2)^m dy.
 \end{aligned}$$

Integrating-by-parts it is easy to show that

$$(3.2) \quad G_{m+1}(u) = G_m(u) + \binom{2m}{m}(u - \frac{1}{2})(u - u^2)^m.$$

Since $G_1(u) = u$ we can iterate (3.2) to obtain

$$(3.3) \quad G_{m+1}(u) = \frac{1}{2} + (u - \frac{1}{2}) \sum_{j=0}^m \binom{2j}{j}(u - u^2)^j.$$

This exact expression (3.3) is very helpful in carrying out on a computer the numerical (Gauss-Legendre) quadrature of the integral in (2.9) for any t and k ; the exact solutions in n (rounded upwards to the next odd integer) are given in Tables 1, 2, 3 and 4.

In the remainder of this section we consider bounds for the $P\{CS|R\}$ for Formulations 1A and 1B with $\alpha = \frac{1}{2}, k = 2, t = 1$ and $d^* = \epsilon_1^* = \epsilon_2^* = \epsilon^*$ (say).

FORMULATION 1A. For the special case $k = 2, t = 1, \alpha = \frac{1}{2}$ and $\epsilon^* = d^*$, using the right side of (2.9), the equation determining n for Formulation 1A becomes

$$(3.4) \quad \int_{\frac{1}{2}-d^*}^{\frac{1}{2}+d^*} G(u + d^*) dG(u) + G(\frac{1}{2} + 2d^*)G(\frac{1}{2} - d^*) = P^*,$$

where $G = G_{r,n-r+1}$ is the Incomplete Beta cdf with parameters $r = (n + 1)/2 = n - r + 1$. To evaluate the integral I in (3.4), it is useful to note that for $c > 0$

TABLE 1: ($d^* = \epsilon^* = .20$)[†]

Exact smallest odd number n of observations required per population by procedure R for selecting from k populations the one with the largest median

[The 4 entries in each cell correspond to formulations 1A, 1B, 2A and 2B and are based on (2.9), (2.13), (2.18) and (2.19), respectively]

P^*	k								
	2	3	4	5	6	7	8	9	10
.550	1	3	7	9	11	11	13	15	15
	1	3	7	9	11	11	13	15	15
	3	5	7	9	11	11	13	13	15
	1	5	7	9	9	11	13	13	15
.600	3	5	9	11	13	15	15	17	19
	1	5	9	11	13	15	15	17	19
	3	7	9	11	11	13	15	15	17
	3	5	7	9	11	13	13	15	15
.650	5	7	11	13	15	17	19	21	23
	3	7	11	13	15	17	19	21	23
	5	7	9	11	13	15	15	17	17
	3	7	9	11	13	15	15	17	17
.700	5	9	13	17	19	21	23	25	27
	5	9	13	17	19	21	23	25	27
	5	9	11	13	15	17	17	19	19
	5	9	11	13	15	15	17	19	19
.750	9	13	17	21	23	25	27	29	31
	7	13	17	21	23	25	27	29	31
	7	11	13	15	17	19	19	21	21
	7	11	13	15	17	17	19	21	21
.800	11	17	21	25	29	31	33	35	37
	11	17	21	25	29	31	33	35	37
	9	13	15	17	19	21	21	23	23
	9	13	15	17	19	21	21	23	23
.850	15	23	29	33	35	39	41	43	45
	15	23	29	33	35	39	41	43	45
	13	15	19	21	21	23	25	25	27
	11	15	17	21	21	23	25	25	27
.900	21	31	37	41	45	47	51	53	55
	21	31	37	41	45	47	51	53	55
	15	19	23	25	27	27	29	31	31
	15	19	23	25	25	27	29	29	31
.950	35	45	51	57	61	65	67	69	71
	35	45	51	57	61	65	67	69	71
	23	27	29	31	33	35	37	37	39
	23	27	29	31	33	35	37	37	39
.975	47	59	67	73	77	81	83	87	89
	47	59	67	73	77	81	83	87	89
	29	33	37	39	41	41	43	45	45
	29	33	37	39	41	41	43	45	45
.990	67	79	89	93	99	103	105	107	111
	67	79	89	93	99	103	105	107	111
	39	43	45	47	49	51	53	55	55
	39	43	45	47	49	51	53	55	55

[†] ϵ^* is the common value of $\epsilon_1^* = \epsilon_2^*$; for formulations 2A and 2B only ϵ^* is used.

TABLE 2: ($d^* = \epsilon^* = .15$)†

Exact smallest odd number n of observations required per population by procedure R for selecting from k populations the one with the largest median

[The 4 entries in each cell correspond to formulations 1A, 1B, 2A and 2B and are based on (2.9), (2.13), (2.18) and (2.19), respectively]

P^*	k								
	2	3	4	5	6	7	8	9	10
.550	3	7	11	15	19	21	23	27	29
	1	7	11	15	19	21	23	27	29
	5	9	13	17	19	21	23	25	27
	3	9	11	15	17	21	23	25	27
.600	5	11	15	19	23	27	29	31	33
	3	11	15	19	23	27	29	31	33
	7	11	15	19	21	23	25	27	29
	5	11	15	17	21	23	25	27	29
.650	7	13	19	25	29	31	35	37	41
	7	13	19	25	29	31	35	37	41
	9	13	17	21	23	27	29	31	33
	7	13	17	21	23	25	27	29	31
.700	11	19	25	31	35	39	41	45	47
	9	19	25	31	35	39	41	45	47
	11	17	21	23	27	29	31	33	35
	9	15	19	23	27	29	31	33	35
.750	15	23	31	37	43	47	51	53	55
	13	23	31	37	43	47	51	53	55
	13	19	23	27	31	33	35	37	39
	13	19	23	27	29	33	35	37	39
.800	19	31	39	47	51	57	59	63	67
	19	31	39	47	51	57	59	63	67
	17	23	27	31	35	37	39	41	43
	17	23	27	31	35	37	39	41	43
.850	27	41	51	57	63	69	73	77	79
	27	41	51	57	63	69	73	77	79
	21	29	33	37	41	43	45	47	49
	21	27	33	37	39	43	45	47	49
.900	39	55	67	75	81	85	91	95	97
	39	55	67	75	81	85	91	95	97
	29	35	41	45	47	51	53	55	57
	29	35	41	45	47	51	53	55	57
.950	61	81	93	103	109	115	121	125	129
	61	81	93	103	109	115	121	125	129
	41	49	53	57	61	63	67	69	71
	41	49	53	57	61	63	67	69	71
.975	85	107	121	131	139	145	149	155	159
	85	107	121	131	139	145	149	155	159
	53	61	67	71	75	77	79	81	83
	53	61	67	71	75	77	79	81	83
.990	119	143	159	169	177	183	189	193	197
	119	143	159	169	177	183	189	193	197
	71	79	83	89	91	95	97	99	101
	71	79	83	89	91	95	97	99	101

† ϵ^* is the common value of $\epsilon_1^* = \epsilon_2^*$; for formulations 2A and 2B only ϵ^* is used.

TABLE 3: ($d^* = \epsilon^* = .10$)†

Exact smallest odd number n of observations required per population by procedure R for selecting from k populations the one with the largest median

[The 4 entries in each cell correspond to formulations 1A, 1B, 2A and 2B and are based on (2.9), (2.13), (2.18) and (2.19), respectively]

P^*	k								
	2	3	4	5	6	7	8	9	10
.550	9	17	27	35	43	49	55	61	65
	3	17	27	35	43	49	55	61	65
	11	21	29	37	43	49	53	57	61
	5	17	27	35	41	47	51	55	59
.600	11	23	35	45	53	61	67	73	77
	9	23	35	45	53	61	67	73	77
	15	25	35	41	49	53	59	63	67
	9	23	33	39	47	53	57	61	65
.650	17	31	45	55	65	73	79	87	91
	13	31	45	55	65	73	79	87	91
	19	31	39	47	55	59	65	69	73
	15	29	39	47	53	59	63	69	73
.700	23	41	57	69	79	87	95	103	107
	21	41	57	69	79	87	95	103	107
	25	37	47	55	61	67	73	77	81
	21	35	45	53	61	67	71	75	81
.750	33	53	71	85	97	105	113	121	127
	31	53	71	85	97	105	113	121	127
	31	43	55	63	69	75	81	85	89
	27	43	53	61	69	75	79	85	89
.800	45	71	89	105	117	127	135	143	151
	45	71	89	105	117	127	135	143	151
	39	53	63	73	79	85	91	95	99
	37	51	63	71	79	85	91	95	99
.850	63	93	115	131	145	155	165	173	179
	63	93	115	131	145	155	165	173	179
	49	65	75	85	91	99	103	109	113
	49	63	75	83	91	97	103	109	113
.900	89	125	149	169	183	195	205	213	221
	89	125	149	169	183	195	205	213	221
	65	81	93	103	109	117	121	127	131
	65	81	93	101	109	115	121	127	131
.950	139	183	211	233	249	261	273	281	291
	139	183	211	233	249	261	273	281	291
	95	111	123	133	139	147	153	157	163
	93	111	123	133	139	147	153	157	161
.975	195	243	275	297	313	327	339	349	359
	195	243	275	297	313	327	339	349	359
	123	141	153	163	171	177	183	189	193
	123	141	153	163	171	177	183	189	193
.990	271	325	359	383	401	415	427	439	449
	271	325	359	383	401	415	427	439	449
	163	181	193	203	211	219	225	229	235
	163	181	193	203	211	219	225	229	235

† ϵ^* is the common value of $\epsilon_1^* = \epsilon_2^*$; for formulations 2A and 2B only ϵ^* is used.

TABLE 4: ($d^* = \epsilon^* = .05$)†

Exact smallest odd number n of observations required per population by procedure R for selecting from k populations the one with the largest median

[The 4 entries in each cell correspond to formulations 1A, 1B, 2A and 2B and are based on (2.9), (2.13), (2.18) and (2.19), respectively]

P^*	k								
	2	3	4	5	6	7	8	9	10
.550	31	69	107	141	171	197	221	243	263
	15	67	105	141	171	197	221	243	263
	43	83	117	147	171	193	213	231	247
	17	71	107	137	163	187	207	225	241
.600	47	93	139	179	213	243	269	291	313
	33	93	139	179	213	243	269	291	313
	57	101	137	169	193	217	237	255	271
	35	91	129	161	187	211	231	249	265
.650	67	125	179	223	261	293	323	347	369
	55	125	179	223	261	293	323	347	369
	75	123	161	193	219	241	263	281	297
	57	115	153	185	213	237	257	277	293
.700	95	165	227	277	319	353	385	411	435
	85	165	227	277	319	353	385	411	435
	97	147	187	219	247	271	291	311	327
	81	141	181	215	243	267	287	307	323
.750	131	217	285	341	387	425	457	487	511
	125	217	285	341	387	425	457	487	511
	123	177	219	253	281	305	325	345	361
	111	171	213	247	277	301	323	343	359
.800	181	283	361	421	471	511	547	577	605
	177	283	361	421	471	511	547	577	605
	157	213	257	291	321	345	367	387	403
	147	209	253	289	317	343	365	385	401
.850	253	371	459	527	579	623	661	695	723
	249	371	459	527	579	623	661	695	723
	201	261	307	341	371	397	419	439	457
	193	259	303	339	369	395	417	437	455
.900	361	503	601	675	733	781	823	857	889
	359	503	601	675	733	781	823	857	889
	265	329	377	413	443	469	493	513	531
	261	327	375	411	443	469	491	511	529
.950	563	737	851	933	997	1049	1095	1133	1167
	561	737	851	933	997	1049	1095	1133	1167
	381	449	497	535	567	593	617	639	657
	379	447	497	535	567	593	617	637	657
.975	781	979	1103	1191	1259	1317	1363	1405	1441
	781	979	1103	1191	1259	1317	1363	1405	1441
	499	569	619	659	691	719	743	763	783
	499	569	619	659	691	719	743	763	783
.990	1087	1307	1439	1535	1607	1667	1717	1761	1801
	1087	1307	1439	1535	1607	1667	1717	1761	1801
	661	733	783	825	857	885	909	931	949
	659	733	783	823	857	885	909	931	949

† ϵ^* is the common value of $\epsilon_1^* = \epsilon_2^*$; for formulations 2A and 2B only ϵ^* is used.

and $2d^*/c$ an integer,

$$(3.5) \quad I \cong \frac{1}{2} \sum_{j=1}^{2d^*/c} [G(\frac{1}{2} + jc) + G(\frac{1}{2} + (j - 1)c)][G(\frac{1}{2} - d^* + jc) - G(\frac{1}{2} - d^* + (j - 1)c)];$$

this is obtained by considering non-overlapping squares of side length c whose diagonals form the line segment on $v = u + d^*$ with $\frac{1}{2} - d^* < u < \frac{1}{2} + d^*$ and replacing the integral over the lower triangle of each square by $\frac{1}{2}$ the integral over the square. If we replace the integral over the lower triangle in each square by the integral over the whole square we clearly obtain an upper bound for I ; it follows that an upper bound B_1 to the absolute value of the error in (3.5) is given by

$$(3.6) \quad B_1 = \frac{1}{2} \sum_{j=1}^{2d^*/c} [G(\frac{1}{2} + jc) - G(\frac{1}{2} + (j - 1)c)][G(\frac{1}{2} - d^* + jc) - G(\frac{1}{2} - d^* + (j - 1)c)].$$

We choose $c < d^*$ so that $0 < \Delta = (d^* - c)/2 < \frac{1}{2}$; then the point $(\frac{1}{2}, \frac{1}{2})$ is not included in any of these squares and for each square the distance from the boundary to $(\frac{1}{2}, \frac{1}{2})$ is at least $\Delta 2^{\frac{1}{2}}$. Hence each square has a distance from $\frac{1}{2}$ of at least Δ in one of the two coordinates. Since there are $2d^*/c$ squares it follows that

$$(3.7) \quad B_1 < B_2 = (d^*/c)[G(\frac{1}{2} + \Delta + c) - G(\frac{1}{2} + \Delta)].$$

We note that $B_1 = cd^*$ for $n = 1$ and it can be shown that B_1 decreases with n ; in fact it can be shown that both B_1 and B_2 decrease exponentially fast with n .

FORMULATION 1B. For $k = 2, t = 1, \alpha = \frac{1}{2}$ and $\epsilon^* = d^*$, using the right side of (2.13), the equation determining n for Formulation 1B becomes

$$(3.8) \quad \int_{\frac{1}{2}-d^*}^{\frac{1}{2}+d^*} G(u + d^*) dG(u) + G(\frac{1}{2} - d^*) + \frac{1}{2}[G(\frac{1}{2} + 2d^*) - G(\frac{1}{2} + d^*)]^2 = P^*$$

where G is the same as in (3.4). The only part that presents any difficulty in (3.8) is the integral and it is the same integral as the one already considered for Formulation 1A above; hence all the results above on I can be used here also.

4. Acknowledgment. The author wishes to thank Professor M. Haseeb Rizvi of Ohio State University and Professor George Woodworth of Stanford University for several helpful discussions in connection with this paper; thanks are also due to Professor Aryeh Dvoretzky for suggesting that the constants ϵ_1^* and ϵ_2^* be taken proportional to α and $1 - \alpha$, respectively (see the end of Section 1). The author is also indebted to Miss Marilyn Huyett (now Mrs. Becker) formerly of The Bell Telephone Laboratories and to Mrs. Elaine Frankowski of the Computation Center of the University of Minnesota for their help in the computation of the tables.

REFERENCE

[1] RIZVI, M. HASEEB and SOBEL, MILTON (1967). Nonparametric procedures for selecting a subset containing the population with the largest α -quantile. *Annals Math. Statist.* **38** 1788-1803.