AD 677313

ON SELECTION PROCEDURES BASED ON RANKS: COUNTEREXAMPLES CONCERNING LEAST FAVORABLE CONFIGURATIONS

BY
M. HASEEB RIZVI AND GEORGE G. WOODWORTH

TECHNICAL REPORT NO. 114
October 28, 1968

SUPPORTED BY THE ARMY, NAVY AND AIR FORCE UNDER
CONTRACT Nonr-225(53)(NR-042-002)
WITH THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

On Selection Procedures Based on Ranks:

Counterexamples Concerning Least Favorable Configurations

By

M. Haseeb Rizvi[1] and George G. Woodworth

TECHNICAL REPORT NO. 114

October 28, 1968

Supported by the Army, Navy and Air Force under
Contract Nonr-225(53) (NR-042-002)
with the Office of Naval Research

Gerald J. Lieberman, Project Director

[1]The research of this author was supported by the Federal Highway Administration Contract FH 11-6667 with Stanford University, and NASA Contract No. NGR36-008-040 with the Ohio State University.

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

The next counterexample shows that (2.7) is false; and it seems to us that this invalidates $R(\delta^*, P^*)$ as a reasonable procedure since the infimum of $P[CS]$ is not controlled even asymptotically. The expedient of the authors of the latest version of [7] of considering only that part of the parameter space where $\theta_{[k]} - \theta_{[1]} = O(n^{-\frac{1}{2}})$ is difficult to translate into practice. Does it mean that one should use $R(\delta^*, P^*)$ only when one is convinced that $\theta_{[k]} - \theta_{[1]} = O(n^{-\frac{1}{2}})$?

Counterexample 2.

Consider the logistic cdf $F(x) = (1 + e^{-x})^{-1}$ and let $\underset{\sim}{\theta}(\delta^*) \in D(\delta^*)$ be a sequence of $\underset{\sim}{\theta}$-values depending on δ^* as follows:

$$\theta_1 = \cdots = \theta_{k-t-1} = -\theta_0,\ \theta_{k-t} = 0,\ \theta_{k-t+1} = \delta^*, \tag{2.8}$$

$$\theta_{k-t+2} = \cdots = \theta_k = \theta_0,$$

where θ_0 is a fixed positive constant and $\delta^* < \theta_0$.

We now prove the following assertion: For each $k \geq 3$ and each $t < k$, there exists a value of P^*, say P_0^*, $\binom{k}{t}^{-1} < P_0^* < 1$, such that

In problem II the experimenter sets only the P^*-value and requires that, with probability greater P^*, the selected subset contains the index of the largest θ-value. This problem might arise in screening drugs as cancer cures; one would want to reduce the number of drugs which are to be submitted to further tests but at the same time be reasonably sure of not eliminating any drug which is a potential cure.

In this paper we examine certain procedures which have been claimed elsewhere to be solutions to these problems. We show by means of specific examples that these procedures are in fact not solutions and should be used with caution if they are used at all.

On Selection Procedures Based on Ranks: Counterexamples Concerning Least Favorable Configurations

By

M. Haseeb Rizvi and George G. Woodworth

1. Introduction

Let $\pi_1, \pi_2, \ldots, \pi_k$ denote $k > 2$ univariate populations differing only in location; that is, an observation X_i drawn from π_i has cumulative distribution function (cdf) $F(x - \theta_i)$ where F is a known continuous cdf with square integrable density f but the location parameter vector $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ is unknown. Let the ordered values of the location parameters be denoted by $\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}$.

Selecting the t best populations.

The decision problem here is to select the populations corresponding to the $t < k$ largest θ-values. The goal of the decision maker is to find a procedure, say R, and a sample size n such that the probability of a correct selection using rule R, $P[CS|R,\underset{\sim}{\theta}]$, has the property that

$$\inf_{\underset{\sim}{\theta} \in D(\delta^*)} P[CS|R,\underset{\sim}{\theta}] \geq P^*, \tag{1.1}$$

where

$$D(\delta^*) = \{\underset{\sim}{\theta}: \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta^*\}, \tag{1.2}$$

and $\binom{k}{t}^{-1} < P^* < 1$ and $\delta^* > 0$ are preassigned constants.

Selecting a subset containing the best population.

The decision problem here is to select a subset of the k populations containing the population associated with $\theta_{[k]}$. The goal of the decision maker is to find for fixed n and preassigned $P^* < 1$ a procedure, say R', such that

$$\inf_{\underset{\sim}{\theta}} P[CS|R', \underset{\sim}{\theta}] \geq P^*. \tag{1.3}$$

We consider two procedures (proposed elsewhere) based on rank sums and show by counterexamples in sections 2 and 3 that they do not satisfy (1.1) (or (1.3)).

2. A procedure based on rank sums for selecting the t best populations.

Let $\{X_{ij}: i = 1, \ldots, k, j = 1, \ldots, n\}$ be k samples each of size n (n is to be determined by (1.1)), X_{ij} being the $j^{\underline{th}}$ observation from π_i, and let R_{ij} be the rank of X_{ij} among all the observations.

Define the rank sums

$$T_{in} = \frac{1}{n^2} \sum_{j=1}^{n} R_{ij}, \quad i = 1, \ldots, k \tag{2.1}$$

$$= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{s=1}^{n} \sum_{r=1}^{k} I(X_{ij} > X_{rs}) + \frac{1}{n}, \tag{2.2}$$

where $I(\cdot)$ is the indicator of the event in parentheses.

The proposed selection rule, call it $R(n)$, is as follows:

i) Draw samples of size n from each population and compute T_{in} for $i = 1, \ldots, k$.

ii) Select the t populations having the largest T_{in}-values, resolving ties by the obvious randomization.

The problem now is to find a value $n = n(\delta^*, P^*; k, t, F)$ such that $R(n)$ satisfies (1.1).

In solving this problem a crucial role is played by the slippage configuration $\underset{\sim}{\theta}_0$:

$$\theta_{[1]} = \cdots = \theta_{[k-t]} = \theta_{[k-t+1]} - \delta^* = \cdots = \theta_{[k]} - \delta^*. \tag{2.3}$$

Many selection rules, for example the rule based on the sample means, have the property that the infimum in (1.1) is attained when $\underset{\sim}{\theta}$ is in the slippage configuration; in other words for many rules the slippage configuration is the least favorable configuration. For such rules it is a relatively easy task to find the appropriate value of n (see, for instance, Example 1 of [1]). The following counterexample, kindly communicated to the authors by E. L. Lehmann, shows that for the rank-sum rule $R(n)$ the slippage configuration is not least favorable.

Counterexample 1 (E. L. Lehmann).

Let $k = 3$, $t = 1$ and let F be a continuous cdf which places probability q and $p = 1 - q$ respectively on the intervals $(0, \varepsilon)$ and $(1, 1 + \varepsilon)$; $\varepsilon < 1/3$ is a constant. Let $\delta^* = \varepsilon$ and consider two parameter values:

$$\underset{\sim}{\theta}_0 = (0, 0, \delta^*), \quad \underset{\sim}{\theta}_1 = (0, \delta^*, 2\delta^*).$$

For $n = 2$, we show that

$$P[CS|R(2), \underset{\sim}{\theta}_0] > P[CS|R(2), \underset{\sim}{\theta}_1]. \tag{2.4}$$

Since $\underset{\sim}{\theta}_0$ is in the slippage configuration and $\underset{\sim}{\theta}_0$, $\underset{\sim}{\theta}_1 \in D(\delta^*)$, defined by (1.2), this provides the required counterexample.

<u>Proof</u>: The supports of the distributions of the populations under the two parameter configurations can be depicted as shown in Figure 1.

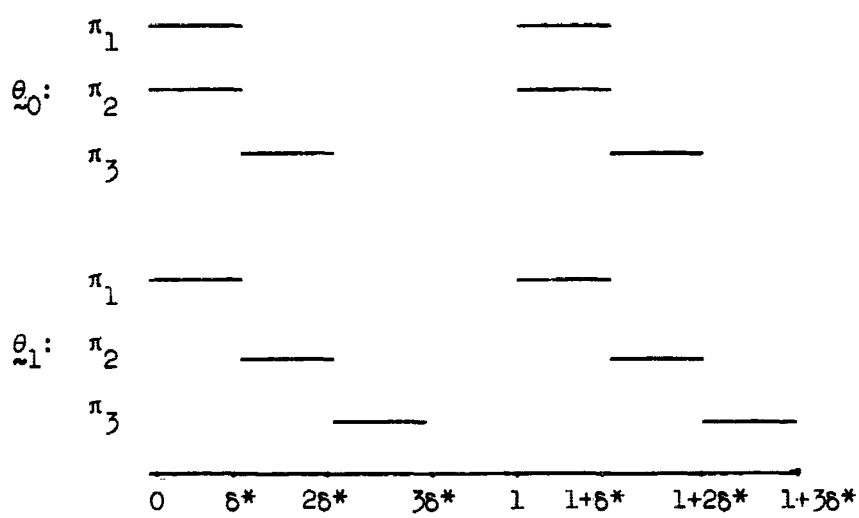

Figure 1: Supports of Distributions.

Let B_i be 0, 1 or 2 according as 0, 1 or 2 observations from π_i are in the upper interval of the support of its distribution, $\underset{\sim}{B} = (B_1, B_2, B_3)$ and $\underset{\sim}{b} = (b_1, b_2, b_3)$ is a realization of $\underset{\sim}{B}$. Clearly $P[\underset{\sim}{B} = \underset{\sim}{b} | \underset{\sim}{\theta}] = \prod_{i=1}^{3} \binom{2}{b_i} p^{b_i} q^{2-b_i}$ for $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ or $\underset{\sim}{\theta}_1$.

$\underset{\sim}{R} = \{R_{ij} : i = 1, 2, 3, j = 1, 2\}$ is the vector of ranks and $\underset{\sim}{r} = \{r_{ij}\}$ is a realization of $\underset{\sim}{R}$. Given $\underset{\sim}{R} = \underset{\sim}{r}$ a correct selection (selection of π_3) occurs with probability 1 if $r_{31} + r_{32} > \max(r_{21} + r_{22}, r_{11} + r_{12})$, with probability $\frac{1}{2}$ if $r_{31} + r_{32} = r_{21} + r_{22} > r_{11} + r_{12}$ or $r_{31} + r_{32} = r_{11} + r_{12} > r_{21} + r_{22}$, and with probability 1/3 if $r_{31} + r_{32} = r_{21} + r_{22} = r_{11} + r_{12}$. The conditional probability that $\underset{\sim}{R} = \underset{\sim}{r}$ given $\underset{\sim}{B} = \underset{\sim}{b}$ is easy to compute, for example

$$P[\underset{\sim}{R} = (1, 2; 3, 4; 5, 6) \mid \underset{\sim}{B} = (0, 0, 0), \underset{\sim}{\theta}_i] = \begin{cases} 1/48 & i = 0 \\ 1/8 & i = 1. \end{cases}$$

Thus, for each of the 27 values of $\underset{\sim}{b}$ one can determine the conditional probability of a correct selection given $\underset{\sim}{B} = \underset{\sim}{b}$ under $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_1$. For most of the $\underset{\sim}{b}$ the probability is the same under $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_1$ but in the six cases listed in Table 1 there is a difference.

Table 1

$\underset{\sim}{b}$	$P[\underset{\sim}{B} = \underset{\sim}{b}]$	$P[CS \mid \underset{\sim}{B} = \underset{\sim}{b}, \underset{\sim}{\theta}]$ $\underset{\sim}{\theta}_0$	$\underset{\sim}{\theta}_1$
(0, 1, 0)	$2pq^5$	5/6	1/2
(1, 0, 0)	$2pq^5$	5/6	1
(1, 1, 0)	$4p^2q^4$	1/6	0
(1, 2, 1)	$4p^4q^2$	1/2	0
(2, 1, 1)	$4p^4q^2$	1/2	1
(2, 2, 1)	$2p^5q$	1/9	0

Thus

$$P[CS \mid R(2), \underset{\sim}{\theta}_0] - P[CS \mid R(2), \underset{\sim}{\theta}_1]$$

$$= \frac{1}{3} pq^5 + \frac{2}{3} p^2q^4 + \frac{2}{9} p^5q > 0,$$

which establishes counterexample 1.

The possibility still remains that the slippage configuration is asymptotically $(\delta^* \to 0)$ least favorable; an asymptotic solution based on this assumption has been claimed by various authors ([4], [7] and [8]). This solution is as follows:

Let $A(P^*; k, t)$ be the solution of

$$\int \Phi^{k-t}(x + A)\,d\Phi^{t}(x) = P^* \tag{2.5}$$

where Φ is the standard normal cdf, and define $n(\delta^*, P^*; k, t, F)$ to be the smallest integer larger than

$$A^2(P^*; k, t) \,/\, 12[\delta^* \int f^2(x)dx]^2, \tag{2.6}$$

where f is the derivative of F. The selection rule $R(\delta^*, P^*; k, t, F) = R(\delta^*, P^*)$ is the rule $R(n)$ with n set equal to $n(\delta^*, P^*; k, t, F)$. The natural inclination to call $R(\delta^*, P^*)$ "distribution-free" must be resisted; obviously one needs to know F to carry out this procedure.

If $\underset{\sim}{\theta}$ is in the slippage configuration (2.3), then it can be shown ([7] or [8]) that

$$\lim_{\delta^* \to 0} P[CS | R(\delta^*, P^*), \theta_0] = P^*$$

The authors of [4] and [8] have incorrectly asserted that the slippage configuration is least favorable (this was also asserted in earlier versions of [7]) from which it would follow that $R(\delta^*, P^*)$ satisfies (1.1) asymptotically as $\delta^* \to 0$; i.e. for fixed P^*, it has been claimed that

$$\lim_{\delta^* \to 0} \inf_{\underset{\sim}{\theta} \in D(\delta^*)} P[CS | R(\delta^*, P^*), \underset{\sim}{\theta}] = P^*. \tag{2.7}$$

The next counterexample shows that (2.7) is false; and it seems to us that this invalidates $R(\delta^*, P^*)$ as a reasonable procedure since the infimum of $P[CS]$ is not controlled even asymptotically. The expedient of the authors of the latest version of [7] of considering only that part of the parameter space where $\theta_{[k]} - \theta_{[1]} = O(n^{-\frac{1}{2}})$ is difficult to translate into practice. Does it mean that one should use $R(\delta^*, P^*)$ only when one is convinced that $\theta_{[k]} - \theta_{[1]} = O(n^{-\frac{1}{2}})$?

Counterexample 2.

Consider the logistic cdf $F(x) = (1 + e^{-x})^{-1}$ and let $\underset{\sim}{\theta}(\delta^*) \in D(\delta^*)$ be a sequence of $\underset{\sim}{\theta}$-values depending on δ^* as follows:

$$(2.8) \qquad \theta_1 = \cdots = \theta_{k-t-1} = -\theta_0, \ \theta_{k-t} = 0, \ \theta_{k-t+1} = \delta^*,$$

$$\theta_{k-t+2} = \cdots = \theta_k = \theta_0,$$

where θ_0 is a fixed positive constant and $\delta^* < \theta_0$.

We now prove the following assertion: For each $k \geq 3$ and each $t < k$, there exists a value of P^*, say P_0^*, $\binom{k}{t}^{-1} < P_0^* < 1$, such that

$$(2.9) \qquad \lim_{\delta^* \to 0} P[CS|R(\delta^*, P_0^*), \underset{\sim}{\theta}(\delta^*)] < P_0^*,$$

which clearly contradicts (2.7).

Lemma 1.

$$(2.10) \qquad \lim_{\delta^* \to 0} P[CS|R(\delta^*, P^*), \underset{\sim}{\theta}(\delta^*)]$$

$$\leq \Phi(2^{-\frac{1}{2}} A^* \rho(\theta_0)),$$

where

$$(2.11) \qquad A^* = A(P^*; k, t),$$

$$(2.12) \qquad \rho(\theta_0) = 3^{\frac{1}{2}} \int H_{\theta_0}(2F - 1)dF/[\int H_{\theta_0}^2 dF - (\int H_{\theta_0} dF)^2]^{\frac{1}{2}}$$

and

$$(2.13) \quad H_{\theta_0}(x) = k^{-1}[(k - t - 1)F(x + \theta_0) + 2F(x) + (t - 1)F(x - \theta_0)].$$

Proof: Notice first that if $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$, then

$$(2.14) \qquad P[CS|R(\delta^*, P^*), \underset{\sim}{\theta}]$$

$$\leq P[\max_{1 \leq i \leq k-t} T_{in} \leq \min_{k-t<j\leq k} T_{jn}|\underset{\sim}{\theta}]$$

$$\leq P[T_{k-t+1,n} - T_{k-t,n} \geq 0|\underset{\sim}{\theta}],$$

where n is the smallest integer greater than (2.6). From (2.2) one has, with probability one when $\underset{\sim}{\theta} = \underset{\sim}{\theta}(\delta^*)$,

$$\begin{aligned}
T_{k-t+1,n} - T_{k-t,n} &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{s=1}^{n}\{2I(X_{k-t+1,j} > X_{k-t,s}) - 1 \\
&\qquad + \sum_{\substack{i\neq k-t \\ \text{or } k-t+1}}[I(X_{k-t+1,j} > X_{is}) \\
&\qquad\qquad - I(X_{k-t,j} > X_{is})]\}
\end{aligned}$$

$$\begin{aligned}
(2.15)\quad &= \frac{1}{n}\sum_{j=1}^{n}\sum_{i\neq k-t,k-t+1}\{F(X_{ij} - \delta^*) - F(X_{ij})\} \\
&\quad - \frac{1}{n}\sum_{j=1}^{n}\{2F(X_{k-t,j} - \delta^*) + (k-t-1)F(X_{k-t,j} + \theta_0) \\
&\qquad + (t-1)F(X_{k-t,j} - \theta_0)\} \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\{2F(X_{k-t+1,j}) + (k-t-1)F(X_{k-t+1,j} + \theta_0) \\
&\qquad + (t-1)F(X_{k-t+1,j} - \theta_0)\} \\
&\quad + 1 - 2\int F(x+\delta^*)dF(x) + (k-t-1)\int F(x+\theta_0)d(F(x-\delta^*) - F(x)) \\
&\qquad + (t-1)\int F(x-\theta_0)d(F(x-\delta^*) - F(x)) \\
&\qquad + \varepsilon_n(\theta_0, \delta^*),
\end{aligned}$$

where $E\,\varepsilon_n^2(\theta_0, \delta^*) \leq C/n^2$ and C is an absolute constant. Note that (2.15) is obtained by U-statistic arguments in imitation of, say, the proof of Theorem 5.6, p. 229 of [3].

Let

$$(2.16)\qquad W_n = n^{\frac{1}{2}}(T_{k-t+1,n} - T_{k-t,n}),$$

routine calculation yields

$$EW_n = n^{\frac{1}{2}}\{2\int F(x+\delta^*)dF(x) - 1$$

$$+ (k-t-1)\int(F(x-\theta_0) - F(x-\theta_0-\delta^*))dF(x)$$

$$+ (t-1)\int(F(x+\theta_0) - F(x+\theta_0-\delta^*))dF(x)\}.$$

By (2.6) and (2.11) one has $n^{\frac{1}{2}}\delta^* \to A^*[12\int f^2]^{-\frac{1}{2}}$ as $\delta^* \to 0$; thus, by Olshen's Lemma (p. 1766 of [5])

$$(2.17)\qquad \lim_{\delta^*\to 0} EW_n = \frac{A^*}{\sqrt{12\int f^2}}\{2\int f^2(x)dx + (k-t-1)\int f(x-\theta_0)f(x)dx$$
$$+ (t-1)\int f(x+\theta_0)f(x)dx\}.$$

Also

$$(2.18)\qquad \lim_{\delta^*\to 0} \mathrm{Var}(W_n) = 2k^2\{\int H_{\theta_0}^2 dF - (\int H_{\theta_0} dF)^2\},$$

where H_{θ_0} is defined by (2.13).

If we set $F(x) = (1+e^{-x})^{-1}$, then $f(x) = F(x)(1-F(x))$ and $\int f^2 = 1/6$, so that (2.17) becomes, after integrating by parts,

$$\lim_{\delta^*\to 0} EW_n = 3^{\frac{1}{2}}A^*k\int H_{\theta_0}(2F-1)dF.$$

Since (2.15) is asymptotically normal by Liapunov's theorem, it follows that

$$\lim_{\delta^*\to 0} P[CS|R(\delta^*, P^*), \underset{\sim}{\theta}(\delta^*)]$$

$$\leq \lim_{\delta^*\to 0} P[T_{k-t+1,n} - T_{k-t,n} \geq 0|\underset{\sim}{\theta}(\delta^*)]$$

$$= \lim_{\delta^* \to 0} P[(W_n - EW_n)/(\mathrm{Var}(W_n))^{\frac{1}{2}} \geq -EW_n/(\mathrm{Var}(W_n))^{\frac{1}{2}} | \underset{\sim}{\theta}(\delta^*)]$$

$$= \Phi(2^{-\frac{1}{2}} A^* \rho(\theta_0)),$$

which proves Lemma 1.

Remark. For $\theta_0 > 0$, H_{θ_0} is clearly not a linear function of F and, since H_{θ_0} and F are both monotone increasing, we have

$$(2.19) \qquad 0 \leq \rho(\theta_0) < 1.$$

Lemma 2.

For any k and t

$$(2.20) \qquad \lim_{P^* \to 1} 2^{\frac{1}{2}} \Phi^{-1}(P^*)/A^* = 1,$$

where $A^* = A(P^*; k, t)$ and A is defined by (2.5).

Proof: Let $Z_1, \ldots, Z_k$ be independent normal (0,1) random variables. Then,

$$1 - P^* = 1 - \int \Phi^{k-t}(x + A^*) d\Phi^t(x)$$

$$= P[\max_{1 \leq i \leq k-t} Z_i > \min_{k-t < j \leq k} Z_j + A^*]$$

$$= P[\bigcup_{1 \leq i \leq k-t < j \leq k} \{Z_i > Z_j + A^*\}]$$

$$\leq t(k - t) P[Z_1 > Z_k + A^*]$$

$$= t(k - t)\,[1 - \Phi(2^{-\frac{1}{2}} A^*)].$$

Also clearly

$$1 - P^* \geq [1 - \Phi(2^{-\frac{1}{2}} A^*)].$$

Lemma 2 now is a consequence of the following easily verifiable fact

$$\lim_{u \to 1} \Phi^{-1}(u)/[-2\log(1-u)]^{\frac{1}{2}} = 1$$

and of the well known approximation to Mills' ratio.

Counterexample 2 now follows from (2.10), (2.19) and (2.20) by selecting P_0^* large enough so that

$$2^{-\frac{1}{2}} A(P_0^*;\ k,\ t)/\Phi^{-1}(P_0^*) < 1/\rho(\theta_0).$$

<u>A remark on the scale parameter case.</u>

Suppose π_i has cdf $F(x/\sigma_i)$ where $F(x) = 0$ for $x < 0$, F is known, and $\underset{\sim}{\sigma} = (\sigma_1, \ldots, \sigma_k)$ is unknown (if $F(x) \neq 0$ for $x < 0$ then replace x by $|x|$). $R(n)$, with X_{ij} replaced by $-X_{ij}$, could be used to select the t smallest σ-values; in [6] it is asserted that, for any constant $\theta^* > 1$, $P[CS|R(n), \underset{\sim}{\sigma}]$ attains its minimum, subject to the condition

$$\sigma^2_{[t+1]}/\sigma^2_{[t]} \geq \theta^* > 1,$$

when

$$\theta^*\sigma^2_{[1]} = \ldots = \theta^*\sigma^2_{[t]} = \sigma^2_{[t+1]} = \ldots = \sigma^2_{[k]}.$$

That this is false, even asymptotically $(\theta^* \to 1)$, follows from Counterexample 2 by considering the random variable $Y = -\log(X)$, since if X has cdf $F(x/\sigma)$ then Y has cdf $1-F(\exp(\mu - y))$, where $\mu = -\log\sigma$, and Y_{ij} has the same rank as $-X_{ij}$.

3. A procedure based on rank sums for selecting a subset containing the best population.

The authors of [2] propose the following procedure, call it $R'(n)$: Put π_i in the selected subset iff

$$T_{in} \geq \max_j T_{jn} - c_n,$$

where

$$c_n = (12n)^{-\frac{1}{2}} kA^* + o(n^{-\frac{1}{2}}) \tag{3.1}$$

and $A^* = A(P^*; k, 1)$, defined by (2.5). We shall show that the slippage configuration: $\theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[k]}$ is not least favorable by proving the following:

Counterexample 3.

Let $\underset{\sim}{\theta}_1$ denote the configuration

$$\theta_1 = \cdots = \theta_{k-2} = -1, \ \theta_{k-1} = \theta_k = 0$$

and let $\underset{\sim}{\theta}_0$ denote the slippage configuration for this problem: $\theta_1 = \theta_2 = \cdots = \theta_k$. If $F(x)$ is as in (3.7) and $k \geq 3$, then

$$\lim_{n \to \infty} P[CS|R'(n), \underset{\sim}{\theta}_1] < P^* = \lim_{n \to \infty} P[CS|R'(n), \underset{\sim}{\theta}_0]. \tag{3.2}$$

Proof: The equality is established in [2] and the inequality below. Clearly

$$P[CS|R'(n), \underset{\sim}{\theta}_1] \leq P[T_{kn} - T_{k-1,n} \geq -c_n | \underset{\sim}{\theta}_1]. \tag{3.3}$$

It follows as in the proof of Lemma 1 that $W_n = n^{\frac{1}{2}}(T_{kn} - T_{k-1,n})$ has a limiting normal distribution with zero mean and variance

$$\sigma^2(H) = 2k^2\{\int H^2 dF - (\int H\, dF)^2\},$$

where

$$(3.4) \qquad H(x) = k^{-1}[(k - 2)F(x + 1) + 2F(x)].$$

Thus by (3.1) and (3.3)

$$\lim_{n \to \infty} P[CS|R'(n), \underset{\sim}{\theta}_1] = \Phi(k(12)^{-\frac{1}{2}}A^*/\sigma(H)).$$

It follows from (2.20) that for any $\varepsilon > 0$ there exists $\frac{1}{2} < P^*_\varepsilon < 1$ such that

$$A^* = A(P^*_\varepsilon;\ k,\ 1) \leq (1 + \varepsilon)2^{\frac{1}{2}}\Phi^{-1}(P^*_\varepsilon).$$

Thus the counterexample will be proved if it can be shown that

$$(3.5) \qquad \sigma^2(H) > k^2/6.$$

From (3.4)

$$(3.6) \quad \sigma^2(H)/2 = 4/12 + 4(k-2)\mathrm{Cov}(F(\mathbf{X}),\ F(\mathbf{X}+1)) + (k - 2)^2\,\mathrm{Var}\,(F(\mathbf{X} + 1)),$$

where $\mathbf{X}$ has cdf F.

Now let

$$(3.7) \qquad F(x) = \begin{cases} 1/2 + x/2b & -b < x \leq 0 \\ 1/2 & 0 < x \leq 1 \\ 1/2 + (x - 1)/2a & 1 < x \leq 1 + a, \end{cases}$$

where $0 < a < 1 < b$ are constants to be determined below.

Thus,

$$F(x + 1) = \begin{cases} 1/2 + (x + 1)/2b & -(b + 1) < x \leq -1 \\ 1/2 & -1 < x \leq 0 \\ 1/2 + x/2a & 0 < x \leq a \\ 1 & a < x \end{cases}$$

or, except for a set having zero $F(x)$-measure,

$$(3.8) \quad F(x + 1) = \begin{cases} F(x) + 1/2b & 0 < F(x) \leq 1/2 - 1/2b \\ 1/2 & 1/2 - 1/2b < F(x) \leq 1/2 \\ 1 & 1/2 < F(x) \leq 1 . \end{cases}$$

If X has cdf F then $F(X)$ is a uniform random variable and it follows from (3.6) and (3.8) that

$$(3.9) \quad \sigma^2(H)/2 = k^2/12 + (13k - 10)(k - 2)/192 - \beta(3k^2 - 8k + 4)/8$$
$$+ 3\beta^2(k - 2)^2/8 + \beta^3(k^2 - 4)/6$$
$$- \beta^4(k - 2)^2/4$$

where $\beta = (2b)^{-1}$. It is clear that for sufficiently small β (large b) the right side of (3.9) can be made larger than $k^2/12$ so that (3.5) is satisfied and Counterexample 3 is proved.

4. Concluding remark.

Procedures $R(n)$ and $R'(n)$ are special cases of the scores procedures proposed in [2], [4], [6], [7] and [8]. The second counterexample probably works for any scores procedure when F (instead of being logistic) is the cdf against which the scores are locally most powerful.

REFERENCES

[1] Barr, D. R. and Rizvi, M. H. (1966). An introduction to ranking and selection procedures. Jour. Amer. Stat. Assoc. **61** 640-646.

[2] Bartlett, N. S. and Govindarajulu, Z. (1965). Some distribution-free statistics and their application to the selection problem. Dittoed manuscript. Abstract in Ann. Math. Statist. **36** 1597-1598.

[3] Fraser, D. A. S. (1957). Nonparametric Methods in Statistics. John Wiley and Sons, Inc., New York.

[4] Lehmann, E. L. (1963). A class of selection procedures based on ranks. Math. Annalen **150** 268-275.

[5] Olshen, Richard A. (1967). Sign and Wilcoxon tests for linearity. Ann. Math. Statist. **38** 1759-1769.

[6] Puri, M. L. and Puri, P. S. (1967). Selection procedures based on ranks: scale parameter case. Mimeo, Series No. 105, Dept. of Statistics, Purdue University. Abstract in Ann. Math. Statist. **37** p. 554.

[7] Puri, M. L. and Puri, P. S. (1968). Multiple decision procedures based on ranks for certain problems in analysis of variance. Unpublished manuscript. Abstract in Ann. Math. Statist. **37** p. 1068.

[8] Woodworth, G. G. (1965). An extension of a result of Lehmann on the asymptotic efficiency of selection procedures based on ranks. Technical Report No. 66, Department of Statistics, University of Minnesota.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Department of Operations Research and Department of Statistics Stanford University Stanford, California	2a. REPORT SECURITY CLASSIFICATION 2b. GROUP

3. REPORT TITLE
On Selection Procedures Based on Ranks: Counterexamples Concerning Least Favorable Configurationa

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (Last name, first name, initial)
Rizvi, M. Haseeb and Woodworth, George G.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 28, 1968	16	8

8a. CONTRACT OR GRANT NO. Nonr-225(53) b. PROJECT NO. NR-042-002 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report No. 114 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Logistics and Mathematical Science Branch Office of Naval Research, Washington, D.C.

13. ABSTRACT

This paper is concerned with certain multiple-decision procedures based on ranks which have been proposed for analyzing data in a one-way layout:

$$X_{ij} = \theta_i + \epsilon_{ij} , i = 1, \ldots , k, j = 1, \ldots , n$$

where the errors (ϵ_{ij}) are independent, have the same known cumulative distribution function (cdf) F and where $\underset{\sim}{\theta} = (\theta_1, \ldots , \theta_k)$ is unknown. Two problems are considered: I. Select the indices of the t largest θ-values. II. Select a subset containing the index of the largest θ-value. In problem I the experimenter sets a preassigned separation threshold $\delta^* > 0$ and a preassigned probability threshold $P^* < 1$ and requires that the procedure he uses have the property that the probability of correct selection is greater than or equal to P^* whenever the t largest θ-values are at least δ^* larger than the rest of the θ-values. This problem might arise if there were k different batches of raw materials available for purchase and one wanted to select the t best batches. In problem II the experimenter sets only the P^*-value and requires that, with probability greater P^*, the selected subset contains the index of the largest θ-value. This problem might arise in screening drugs as cancer cures; one would want to reduce the number of drugs which are to be submitted to further tests but at the same time be reasonably sure of not eliminating any drug which is a potential cure. In this paper we examine certain procedures which have been claimed elsewhere to be solutions to these problems. We show by means of specific examples that these procedures are in fact not solutions and should be used with caution if they are used at all.

DD FORM 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
multiple decision						
ranking						
rank sum						
least favorable						
counterexamples						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.