ON SOME DISTRIBUTION-FREE RANKING AND SELECTION PROCEDURES*

by

Gary Carl McDonald

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series #174

January 1969

*This research was partly supported by Contract AF33(615)67-C-1244.

INTRODUCTION

The shortcomings of the classical tests of homogeneity, i.e., testing the hypothesis of equality of parameters, have long been known. Bahadur [2] was one of the earliest authors to recognize this and to contribute to the theory of the k-sample problems. Given k populations and from each population a fixed number of observations whose distribution depends on a parameter θ_i, concluding that all θ_i are not equal may not be sufficient. Many times the experimenter is interested in assessing which population is associated with the largest (or smallest) θ, which populations possess the t largest (or smallest) θ, etc. These questions may be formulated into multiple decision problems, and a more realistic answer may be obtained if a ranking of the parameters is the desired outcome as is often the case.

Suppose the experimenter is interested in identifying which one of the k populations possess the largest θ. This population will be called the ''best'' population. The parameter θ may be, for example, the mean, variance, some quantile, or some function of these parameters. There have been two approaches to ranking and selection problems, the ''indifference zone'' approach and the ''subset selection'' approach. In the first a single population is chosen and is guaranteed to be the best with probability P^*, a given constant. However, some knowledge of the parameter space is assumed known a priori, e.g., the experimenter must be able to guarantee that the largest parameter is separated from all other ranking parameters by a distance not less than d^*. This formulation is due to Bechhofer [8]. Other contributions to this problem are Bechhofer and Sobel [9]; Bechhofer, Dunnett and Sobel [10]; Sobel and Huyett [61]; Chambers and Jarratt [15]; Barr and Rizvi [5]; Eaton [21]; and Mahamunulu [41]. Selection procedures based on ranks (or scores) of observations are studied in Lehmann [40]; Puri and Puri [54].

The second approach assumes no a priori information about the parameter space. A single population is not necessarily chosen; rather a subset of the given k populations is selected which is guaranteed to contain the best population with probability P^*, the basic probability requirement in these procedures. In this sense the size of the selected subset is a random variable. Among decision procedures which satisfy the basic probability requirement, one which yields the smallest expected size of the selected subset is considered in some ways to be the most desirable. Other performance criteria for comparing decision procedures are: expected minimal rank, expected sum of ranks of the populations selected in the subset, and the expected number of the non-best populations in the selected subset. This ''subset selection'' formulation is due to Gupta [24]. Contributions to this aspect of the problem are Seal [59], [60]; Gupta and Sobel [25], [26], [27]; Lehmann [39]; Gupta [28], [30], [31]; Rizvi [55]; Barr and Rizvi [5]; Studden [62]; Gupta and Nagel [33]. Recent application of the subset selection formulation to multivariate normal populations may be found in Alam and Rizvi [1]; Gnanadesikan [22]; Gupta [32]; Gnanadesikan and Gupta [23]; Gupta and Panchapakesan [34]; Gupta and Studden [35]. Barlow and Gupta [3] and Barlow, Gupta and Panchapakesan [4] have considered the problem of selecting a subset containing the largest (smallest) quantile of a given order and a subset containing the largest (smallest) mean. They assume the observations from each population have a distribution which belongs to certain restricted families, e.g., IFR distributions, IFRA distributions, etc. Distribution-free subset selection procedures have been studied by Bartlett and Govindarajulu [7]; Patterson [48]; Rizvi and Sobel [56].

Multiple decision procedures have also been investigated from a Bayesian point of view. Work assuming an a priori distribution on the parameter space has been done by Dunnett [19]; Guttman and Tiao [36]; Deely and Gupta [17]. Deely [16] uses an empirical Bayes approach to multiple decision problems. He assumes only the existence of an a priori distribution, the exact distribution itself remaining unknown, and then uses a decision theoretic framework with a specific loss structure to derive procedures which minimize the Bayes risk.

The sequential and multistage aspects of the ranking and selection problems have been explored by Bechhofer, Dunnett and Sobel [10]; Bechhofer [11]; Bechhofer and Blumenthal [12]; and Paulson [49], [50], [51], [52]. Nearly all of this work in sequential and multistage procedures has been through the indifference zone approach. Bechhofer, Kiefer and Sobel [13] in a recent monograph have considered sequential procedures for selecting the best of k Koopman-Darmois populations, again using an indifference zone approach. Barron [6] considers sequential procedures from the subset selection approach.

The present thesis deals with some nonrandomized distribution-free ranking and selection procedures using the subset selection approach. The main problem is to select a subset of k given populations which contains the ''best'' population with probability at least P^*. The random variables associated with a fixed population are assumed to be independent identically distributed with a continuous distribution function depending on a single parameter. This parameter is assumed to stochastically order the k distribution functions, and the ''best'' population is the stochastically largest (smallest) population. The procedures presented depend on the individual observations of a given population only through the sum of their ranks in the combined sample. Procedures of this nature are labeled ''nonrandomized rank sum procedures'' (NRSP). In other words, one is not required to have at hand the actual observations from each population; it suffices to have the rank of each observation in the combined sample. In some preference type tests or lost data problems, these ranks may be the only information available to an experimenter. In contrast, randomized distribution-free ranking and selection procedures depend on the individual observations of a given population through the sum of random functions of their joint ranks. Typically these random functions are ordered observations from some fixed distribution. These procedures are also termed ''randomized rank sum procedures'' (RRSP). The work of Bell and Doksum [14] has facilitated some distribution problems associated with the RRSP. This is one reason why most of the literature on distribution-free procedures has been devoted to the randomized procedures. However, these procedures are usually more cumbersome to apply than the nonrandomized procedures indicated here. A more serious drawback to the RRSP is that the chosen subset of populations is not uniquely defined by a given set of data, whereas the subset selected by the nonrandomized procedure is uniquely determined.

Chapter I begins by considering two classes of selection rules depending on a continuous, but otherwise arbitrary, distribution G. In Section 1.2 a theorem is given showing that the probability of a correct selection, with any rule in these two classes, is a nondecreasing function in the largest parameter. The distribution G is then fixed to be the uniform distribution over the unit interval, and two rules are singled out for consideration. Examples are given where the probability of a correct selection is a nondecreasing function of the largest two parameters if they are equal, but not necessarily of the second largest parameter if it does not equal the largest. Upper and lower bounds for the probability of a correct selection are also given. Based on these bounds, a theorem is given which provides a conservative method for obtaining the constants needed to implement the rules under consideration.

Chapter II discusses the distribution theory associated with the two distribution-free procedures under consideration. In Section 2.2 exact expressions are derived for the means, variances and covariances of the rank sums associated with each population. In Section 2.3 the distribution of the maximum rank sum minus an individual rank sum is considered. A recursion formula is developed to obtain the exact distribution when all populations are identically distributed, and the exact results are shown to be in close agreement with asymptotic approximations even for small sample sizes (and a small number of populations). Asymptotic expressions are given for the probability of a correct selection and the expected size of the selected subset for a particular selection rule. In Section 2.4 the distribution of the maximum rank sum divided by an individual rank sum is considered. For more than two populations, all the results are of an asymptotic nature. Section 2.5 gives the exact distribution of the maximum rank product divided by an individual rank product when the populations are identically distributed.

Chapter III compares some performance characteristics of the two procedures with those of possible competing procedures. In Section 3.2 the asymptotic relative efficiency (ARE) of the two distribution-free procedures relative to a normal means procedure is computed. In Section 3.3 a similar comparison is made with Gupta's procedure for gamma populations. In Section 3.4 two parametric procedures are developed for ranking gamma populations which differ only in their guaranteed life time. In Section 3.5 these procedures are then compared with distribution-free procedures (discussed earlier) under certain parameter configurations. The final section contains some discussion of previous results along with indications of future work and conjectures.

CHAPTER I

SOME DISTRIBUTION-FREE RANKING AND SELECTION PROCEDURES

1.1 Formulation of Problem and Two Rules

Let $\pi_1,\ldots,\pi_k$ be $k(\geq 2)$ independent populations. The associated random variables X_{ij}, $j=1,\ldots,n_i$, $i=1,\ldots,k$, are assumed independent and to have a continuous distribution $F_{\theta_i}(x)$ where θ_i belong to some interval Θ on the real line. Suppose $F_\theta(x)$ is a stochastically increasing (SI) family of distributions, i.e., if θ_1 is less than θ_2, then $F_{\theta_1}(x)$ and $F_{\theta_2}(x)$ are distinct and $F_{\theta_2}(x) \leq F_{\theta_1}(x)$ for all x. Examples of such families of distributions are: 1) any location parameter family, i.e., $F_\theta(x) = F(x-\theta)$; 2) any scale parameter family, i.e., $F_\theta(x) = F(x/\theta)$, $\theta > 0$; 3) any family of distribution functions whose densities possess the monotone likelihood ratio (or TP_2) property. Implications of such an ordering are discussed in Lehmann [38]. Let R_{ij} denote the rank of the observation x_{ij} in the combined sample, i.e., if there are exactly r observations less than x_{ij} then $R_{ij} = r+1$. These ranks are well-defined with probability one since the random variables are assumed to have a continuous distribution. Let $Z(1) \leq Z(2) \leq \cdots \leq Z(N)$ denote an ordered sample of size $N = \sum_{i=1}^{k} n_i$ from any continuous distribution G. With each of the random variables X_{ij} associate the number $E[Z(R_{ij})|G]$, and define

$$(1.1.1) \qquad H_i = n_i^{-1} \sum_{j=1}^{n_i} E[Z(R_{ij})|G], \quad i = 1,\dots,k.$$

The following two selection procedures, which choose a subset of the k given populations, will be considered:

$$(1.1.2) \qquad R_1(G)\text{: Select } \pi_i \text{ iff } H_i \geq \max_{1\leq j \leq k} H_j - d,\ i=1,\dots,k,\ d \geq 0$$

$$(1.1.3) \qquad R_2(G)\text{: Select } \pi_i \text{ iff } H_i \geq c^{-1} \max_{1\leq j\leq k} H_j,\ i=1,\dots,k,\ c \geq 1.$$

The constants d and c are chosen so as to satisfy a certain probability requirement which will be imposed on the selection procedures. The number of populations included in the selected subset is a random variable which takes values 1 to k inclusive. Let $\theta_{[1]} \leq \theta_{[2]} \leq \dots \leq \theta_{[k]}$ denote the ordered θ_i. Let $\pi_{(i)}$ be the (unknown) population which has the parameter $\theta_{[i]}$, and let $X_{(i)j}$, $j=1,\dots,n_{(i)}$, be the observations associated with $\pi_{(i)}$. Similarly define $H_{(i)}$ (unknown) to be the quantity computed in (1.1.1) based on the random variables $X_{(i)j}$. The ''best'' population is defined to be $\pi_{(k)}$, the population with the largest parameter. In case several populations possess the largest parameter, one is tagged at random and called the ''best.'' A ''correct selection'' (CS) is said to occur if and only if the ''best'' population, say $\pi_{(k)}$, is included in the selected subset. The probability of making a correct selection using any procedure R is denoted by $P(CS|R)$. Let

$$(1.1.4) \qquad \Omega = \{\underline{\theta} = (\theta_1,\ldots,\theta_k): \theta_i \in \Theta, \quad i = 1,\ldots,k\}.$$

The d value defining the rule $R_1(G)$ in (1.1.2) is the smallest non-negative number such that the probability of a correct selection using rule $R_1(G)$ is never less than a specified constant P^*, $k^{-1} < P^* < 1$, for all $\underline{\theta} \in \Omega$, i.e.,

$$(1.1.5) \qquad \inf_{\Omega} P(CS|R_1(G)) \geq P^*.$$

Similarly, the c value defining rule $R_2(G)$ in (1.1.3) is the smallest number not less than 1 such that (1.1.5) holds with $R_1(G)$ replaced by $R_2(G)$.

The corresponding rules for choosing a subset of the k populations which contains the population with the smallest parameter, say $\pi_{(1)}$, are:

$$(1.1.6) \qquad R_1'(G): \text{ Select } \pi_i \text{ iff } H_i \leq \min_{1\leq j\leq k} H_j + d', \quad i=1,\ldots,k, \; d' \geq 0$$

$$(1.1.7) \qquad R_2'(G): \text{ Select } \pi_i \text{ iff } H_i \leq c' \min_{1\leq j\leq k} H_j, \quad i=1,\ldots,k, \; c' \geq 1.$$

The constants d' and c' are obtained as before. No more consideration will be given to these two rules; results and methods developed for $R_1(G)$ and $R_2(G)$ will have an obvious analogue for $R_1'(G)$ and $R_2'(G)$, respectively.

1.2 Probability of a Correct Selection

A. Monotonicity in $\theta_{[k]}$

To find these d and c values it is essential to know the $\underline{\theta}$ configurations which minimize the $P(CS|R_i(G))$, $i=1,2$. These configurations are partially specified in Theorem 1.2.1 which is obtained after a few preliminary lemmas. The first of these follows simply from Lehmann ([38], pg. 112, No. 11).

Lemma 1.2.1. Let $F_\theta(x)$, $\theta \in \Theta$, be a SI family of distribution functions on the real line. If Ψ is any nondecreasing (nonincreasing) function of x, then $E_\theta[\Psi(X)]$ is a nondecreasing (nonincreasing) function of θ.

By induction Lemma 1.2.2 is obtained.

Lemma 1.2.2. Let $X_1,\ldots,X_n$ be independent identically distributed with distribution $F_\theta(x)$, $\theta \in \Theta$, a SI family of distribution functions on the real line. Let Ψ be a function of $x_1,\ldots,x_n$ which is nondecreasing (nonincreasing) in each of its arguments. Then $E_\theta[\Psi(X_1,\ldots,X_n)]$ is a nondecreasing (nonincreasing) function of θ.

The following lemma is a slightly different version of Lemma 4.2 in Mahamunulu [41] and Lemma 2.1 in Alam and Rizvi [1].

Lemma 1.2.3. Let $\underline{X} = (X_{11},\ldots,X_{1n_1},\ldots,X_{k1},\ldots,X_{kn_k})$ be a vector valued random variable of $\sum_{i=1}^{k} n_i(\geq 1)$ independent components with X_{ij} having the distribution $F_{\theta_i}(x)$, $j=1,\ldots,n_i$, $i=1,\ldots,k$. Suppose $F_\theta(x)$ is a SI family of distributions. Let Ψ be a function of $x_{11},\ldots,x_{1n_1},\ldots,$ $x_{k1},\ldots,x_{kn_k}$ which, for any fixed i, is a nondecreasing (nonincreasing) function of $x_{i1},\ldots,x_{in_i}$ when the other components of $\underline{x}$ are held fixed. Then $E_{\underline{\theta}}[\Psi(\underline{X})]$ is a nondecreasing (nonincreasing) function of θ_i.

Proof. Let $\underline{\theta} = (\theta_1,\ldots,\theta_i,\ldots,\theta_k)$, $\underline{\theta}' = (\theta_1,\ldots,\theta_i',\ldots,\theta_k)$ for some fixed i, $1 \leq i \leq k$. Suppose Ψ is nondecreasing in $x_{i1},\ldots,x_{in_i}$, and $\theta_i \leq \theta_i'$. Let E^* denote expectation with respect to the random variables $X_{i1},\ldots,X_{in_i}$ while $X_{\ell j}$, $\ell \neq i$, are held fixed. Then $E_{\underline{\theta}}[\Psi(\underline{X})] = E[E^*_{\theta_i}(\Psi(\underline{X})]$. By Lemma 1.2.2, $E^*_{\theta_i}[\Psi(\underline{X})] \leq E^*_{\theta_i'}[\Psi(\underline{X})]$, so $E_{\underline{\theta}}[\Psi(\underline{X})] \leq E[E^*_{\theta_i'}(\Psi(\underline{X}))] = E_{\underline{\theta}'}[\Psi(\underline{X})]$. Similar arguments complete the proof if $\Psi(\underline{X})$ is nonincreasing.

The parameter space has been designated by Ω. Now define $\Omega' \subset \Omega$ as

$$\Omega' = \{\underline{\theta} \in \Omega\colon \theta_{[k]} = \theta_{[k-1]}\}. \tag{1.2.1}$$

Theorem 1.2.1. For $i=1,2$, $P_{\underline{\theta}}(CS|R_i(G))$ is a nondecreasing function of $\theta_{[k]}$. Hence

$$\inf_{\Omega} P_{\underline{\theta}}(CS|R_i(G)) = \inf_{\Omega'} P_{\underline{\theta}}(CS|R_i(G)). \tag{1.2.2}$$

Proof. Consider the rule $R_1(G)$. A correct selection occurs if and only if $H_{(k)} \geq \max_{1\leq j\leq k-1} H_{(j)}-d$ since $d \geq 0$. Let

$$\Psi(\underline{X}) = \begin{cases} 1 & \text{if } H_{(k)} \geq \max_{1 \leq j \leq k-1} H_{(j)} - d \\ 0 & \text{otherwise.} \end{cases}$$

Let $R_{(i)j}$ be the rank of $X_{(i)j}$, $j=1,\ldots,n_{(i)}$, and consider an observation $x_{(k)\ell}$ for some fixed ℓ, $1 \leq \ell \leq n_{(k)}$. As $x_{(k)\ell}$ increases and the other observations remain fixed, either:

(1) $x_{(k)\ell}$ surpasses first an $x_{(i)j}$, $i \neq k$, so $R_{(k)\ell}$ increases by 1 and $R_{(i)j}$ decreases by 1.

(2) $x_{(k)\ell}$ surpasses first an $x_{(k)j}$, $j \neq \ell$, so $R_{(k)\ell}$ increases by 1 and $R_{(k)j}$ decreases by 1.

(3) $x_{(k)\ell}$ does not surpass any other observation, so all ranks remain the same.

In all three cases, $H_{(k)}$ is nondecreasing and $H_{(j)}$, $j \neq k$, is nonincreasing and hence so is $\max_{1 \leq j \leq k-1} H_{(j)}$. Therefore $\Psi(\underline{x})$ is a nondecreasing function of $x_{(k)j}$, $j=1,\ldots,n_{(k)}$. By Lemma 1.2.3, $E_{\underline{\theta}}[\Psi(\underline{X})] = P_{\underline{\theta}}(CS|R_1(G))$ is a nondecreasing function of $\theta_{[k]}$. A similar argument proves the result for $R_2(G)$.

<u>Remarks</u>: (1) If H_i in (1.1.1) is redefined to be $H_i^* = n_i^{-1} \sum_{j=1}^{n_i} Z(R_{ij})$ and rules $R_1^*(G)$ and $R_2^*(G)$ are defined by (1.1.2) and (1.1.3) with H_i replaced by H_i^*, $i=1,\ldots,k$, then Theorem 1.2.1 holds with $R_i(G)$ replaced by $R_i^*(G)$. Thus, Theorem 1.2.1 is valid for randomized, as well as nonrandomized, rank sum procedures. (2) Let H_i in (1.1.1) be redefined as $H_i^{**} = n_i^{-1} \sum_{j=1}^{n_i} h(X_{ij})$, where h is any nondecreasing function such that $h(X_{ij})$ is independent of all the random variables except X_{ij}.

Let $R_1^{**}(h)$ and $R_2^{**}(h)$ be defined by (1.1.2) and (1.1.3) when H_i is replaced by H_i^{**}, $i=1,\ldots,k$. Then the function $\Psi(\underline{X})$ defined in the proof of Theorem 1.2.1 is nondecreasing in $X_{(k)j}$, $j=1,\ldots,n_{(k)}$ and also nonincreasing in $X_{(i)j}$, $i=1,\ldots,k-1$, $j=1,\ldots,n_{(i)}$. Hence $\inf_{\Omega} P_{\underline{\theta}}(CS|R_i^{**}(h))$, $i=1,2$, occurs in the subspace where all the parameters are equal.

Of special interest in this thesis is the case where the same number of observations are taken from each population and the distribution G is the uniform distribution over the unit interval. Let n be the common sample size and define

(1.2.3) $$T_i = \sum_{j=1}^{n} R_{ij}, \quad i=1,\ldots,k.$$

The quantity T_i is the sum of ranks of the observations on the random variables associated with population π_i. Rules $R_1(G)$ and $R_2(G)$ given by (1.1.2) and (1.1.3) now have the form:

(1.2.4) $$R_1\text{: Select } \pi_i \text{ iff } T_i \geq \max_{1\leq j\leq k} T_j - d, \quad i=1,\ldots,k,\ d\geq 0.$$

(1.2.5) $$R_2\text{: Select } \pi_i \text{ iff } T_i \geq c^{-1} \max_{1\leq j\leq k} T_j, \quad i=1,\ldots,k,\ c\geq 1.$$

From Theorem 1.2.1, if $k=2$ the probability of a correct selection using either rule R_1 or rule R_2 is minimized when the two populations are identically distributed. The same result is true in a slippage configuration, i.e., if $\theta_{[1]} = \cdots = \theta_{[k-1]}$ then the probability with each are associated two independent random variables denoted by X_{ij}, j=1,2, i=1,2,3. Let X_{ij} have a uniform distribution over the interval $(0,\theta_i)$ where $\Theta = [1,\infty)$ and $\theta_{[1]} = 1$. The random variables X_{ij} can be ordered in 720 different ways, but interchanging random variables from the same population has no effect on the rank sums. Thus, in computing the probabilities of the various rank sum configurations only 90 arrangements of the random variables must be considered and the probability of each of these arrangements is easily calculated. By suitable summation of these probabilities, $P(CS|R_1)=P(T_{(3)} \geq \max_{1\leq j\leq 3} T_{(j)}-d)$ can be obtained as a function of $\theta_{[2]}$ and $\theta_{[3]}$ for d=0,1,...,8. Examination of these equations yields the following results:

(1) $P_{\underline{\theta}}(CS|R_1)$ is a nondecreasing function in $\theta_{[3]}$. This is guaranteed by Theorem 1.2.1.

(2) If $\underline{\theta}' = (\theta_1,\theta_2,\theta_3)$ with $\theta_{[2]} = \theta_{[3]} = \theta'$, then $P_{\underline{\theta}'}(CS|R_1)$ is a nondecreasing function in θ'.

(3) Hence $\inf_{\Omega} P_{\underline{\theta}}(CS|R_1)$ occurs at $\underline{\theta}_0 = (1,1,1)$.

Case 2: Suppose π_1,π_2,π_3 are three independent populations and with each are associated two independent random variables. Let $\Theta = [0,\infty)$, $\theta_{[1]} = 0$ and X_{ij} have the distribution $F(x-\theta_i)$ where

$$F(x-\theta) = \begin{cases} 1 - e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta \end{cases}.$$

After carrying out the calculations indicated in case 1, the following results are obtained:

(1) $P_{\underline{\theta}}(CS|R_1)$ is a nondecreasing function in $\theta_{[3]}$.

(2) If $\underline{\theta}' = (\theta_1, \theta_2, \theta_3)$ with $\theta_{[2]} = \theta_{[3]} = \theta'$, then $P_{\underline{\theta}'}(CS|R_1)$ is a nondecreasing function in θ'.

(3) Hence $\inf_{\Omega} P_{\underline{\theta}}(CS|R_1)$ occurs at $\underline{\theta}_0 = (0,0,0)$.

Case 3: Again consider three independent populations and with each associate two independent random variables. Let X_{ij}, $j=1,2$, $i=1,2,3$, have the continuous distribution

$$
(1.2.6) \qquad F(x-\theta_i) = \begin{cases} 0 & , \quad x \le \theta_i - \epsilon \\ q(x-\theta_i+\epsilon)/2\epsilon & , \quad \theta_i - \epsilon \le x \le \theta_i + \epsilon \\ q & , \quad \theta_i + \epsilon \le x \le \theta_i + 1-\epsilon \\ q+p(x-\theta_i-1+\epsilon)/2\epsilon & , \quad \theta_i+1-\epsilon \le x \le \theta_i+1+\epsilon \\ 1 & , \quad x \ge \theta_i + 1 + \epsilon \end{cases}
$$

where $0 < p < 1$, $p+q=1$. It should be pointed out that the distribution (1.2.6) does not have MLR in x. The probability of a correct selection will be computed for various parameter configurations. For the first of these configurations choose the location parameters and ϵ so that $0 = \theta_{[1]} < \theta_{[2]} < \theta_{[3]}$ and

$$
(1.2.7) \qquad \epsilon < \theta_{[2]}-\epsilon < \theta_{[2]} + \epsilon < \theta_{[3]} - \epsilon < \theta_{[3]} + \epsilon < 1-\epsilon.
$$

The following method provides a systematic way to calculate $P[T_{(3)} \ge \max_{1 \le j \le 3} T_{(j)} - d]$ for $0 \le d \le 8$. Denote a sample outcome by $a_{(1)1}\, a_{(1)2},\ a_{(2)1} a_{(2)2},\ a_{(3)1} a_{(3)2}$, where

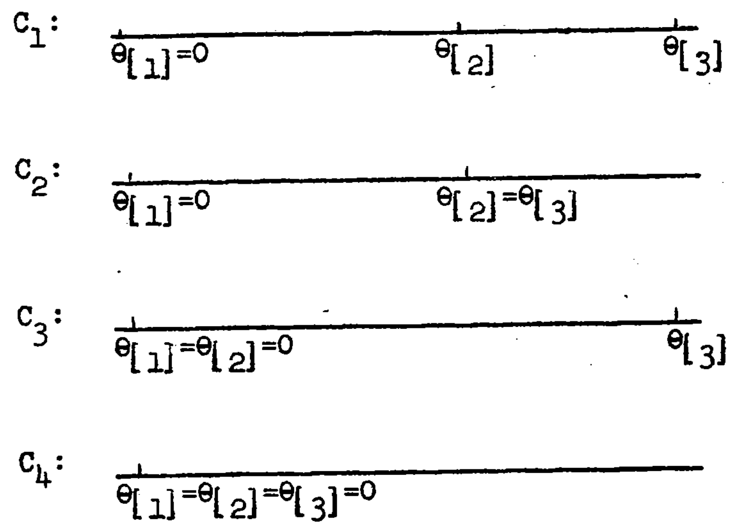
$$(1.2.8) \qquad a_{(i)j} = \begin{cases} 0 \quad , & \theta_{[i]}-\epsilon \le x_{(i)j} \le \theta_{[i]} + \epsilon \\ & \\ 1 \quad , & 1+\theta_{[i]}-\epsilon \le x_{(i)j} \le 1+\theta_{[i]} + \epsilon \end{cases} ,$$

$j=1,2,\ i=1,2,3$. It follows from (1.2.6) that $P[a_{(i)j}=0] = q$ and $P[a_{(i)j}=1] = p$. Thus for each sample outcome its probability and rank sums can be easily computed. Interchanging $a_{(i)1}$ with $a_{(i)2}$ has no effect on the rank sums, so there are 27 sample outcomes to be noted. A few of these are

Sample	Probability	$T_{(1)}, T_{(2)}, T_{(3)}$
00,00,00	q^6	3,7,11
01,01,00	$4q^4p^2$	6,8, 7
11,11,01	$2q\,p^5$	5,9, 7

Now to obtain $P(T_{(3)} \ge \max_{1\le j\le 3} T_{(j)}-d)$ for a fixed d, merely sum the probabilities corresponding to a configuration which satisfies $T_{(3)} \ge T_{(2)}-d$ and $T_{(3)} \ge T_{(1)}-d$.

The above choice of location parameters will be termed a C_1-configuration. A C_2-configuration exists if $\theta_{[3]}$ is reduced to coincide with $\theta_{[2]}$. A C_3-configuration exists if $\theta_{[3]}$ remains fixed and $\theta_{[2]}$ is reduced to coincide with $\theta_{[1]}$. A C_4-configuration exists if both $\theta_{[3]}$ and $\theta_{[2]}$ are reduced to coincide with $\theta_{[1]}(= 0)$. Pictorially the configurations are:



Let $P_{C_i}(CS|R_1)$ represent the probability of a correct selection using rule R_1 when the location parameters are arranged according to configuration C_i. Then the following remarks hold:

(1) $P_{C_1}(CS|R_1) \geq P_{C_2}(CS|R_1) \geq P_{C_4}(CS|R_1)$, $d=0,1,\ldots,8$.

(2) $P_{C_3}(CS|R_1) \geq P_{C_4}(CS|R_1)$, $d=0,1,\ldots,8$

(3) $P_{C_1}(CS|R_1) \leq P_{C_3}(CS|R_1)$, $d=0,1,\ldots,8$, $d \neq 4$.

(4) $P_{C_1}(CS|R_1) > P_{C_3}(CS|R_1)$, $d = 4$.

In this example, $P_{\underline{\theta}}(CS|R_1)$ is nondecreasing in $\theta_{[3]}$, but not in $\theta_{[2]}$ for fixed $\theta_{[3]}$.

C. Upper and Lower Bounds

Upper and lower bounds for the probability of a correct selection can be obtained for rules R_1 and R_2 by making use of some elementary inequalities. First consider rule R_1. Let X_{ij} have a continuous distribution $F_{\theta_i}(x)$, j=1,...,n, i=1,...,k. Then

$$(1.2.9) \qquad P(CS|R_1) = P(T_{(k)} \geq \max_{1 \leq j \leq k-1} T_{(j)} - d),$$

where d is the smallest nonnegative integer so that

$$(1.2.10) \qquad \inf_{\Omega} P(CS|R_1) \geq P^*.$$

Now the rank sums satisfy the inequalities

$$(1.2.11) \qquad (k-1)^{-1} \sum_{j=1}^{k-1} T_{(j)} \leq \max_{1 \leq j \leq k-1} T_{(j)} \leq n(2kn-n+1)/2.$$

The second of these inequalities follows since $T_{(j)}$ is maximized when it is the sum of the last n integers from 1 to (k-1)n. For convenience of notation, let

$$(1.2.12) \qquad v(d,k,n) = \frac{1}{2}\, n(2kn-n+1)-d$$

$$(1.2.13) \qquad u(d,k,n) = \frac{1}{2}\, n(kn+1) - \frac{1}{k}\,(k-1)d\ .$$

Using (1.2.11) in (1.2.9) yields

$$(1.2.14) \qquad P(T_{(k)} \geq v) \leq P(CS|R_1) \leq P(T_{(k)} \geq u),$$

and hence

$$\inf_{\Omega} P(T_{(k)} \geq v) \leq \inf_{\Omega} P(CS|R_1) \leq \inf_{\Omega} P(T_{(k)} \geq u). \tag{1.2.15}$$

The infimum of the left and right hand probabilities in (1.2.15) can be shown to be attained when $\theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[k]}$. This follows directly from Lemma 1.2.3 and by an argument similar to the proof of Theorem 1.2.1. Thus, $P(CS|R_1)$ is bounded below (and above) by a function which in turn attains its minimum value when all parameters are equal, and the bound is independent of the common θ-value and the underlying distribution F. This lower bound is useful in the sense that it gives a conservative method for determining the constant d to implement rule R_1. This is made more precise in the next theorem. Recal that the Mann-Whitney U statistic is calculated using observations from two independent (not necessarily identically distributed) populations with sample sizes p and q, say. Denoting these observations by $x_1,\ldots,x_p$, and $y_1,\ldots,y_q$ then U is the number of times an x_i precedes a y_j. The distribution of the random variable U is well tabulated when the two populations are identically distributed. If T_x denotes the rank sum of the x's in the combined sample, then U and T_x are related by

$$U + T_x = pq + \frac{1}{2} p(p+1). \tag{1.2.16}$$

<u>Theorem 1.2.2</u>. If U is the Mann-Whitney statistic associated with sample sizes n and (k-1)n from identically distributed populations, then

$$\inf_{\Omega} P(CS|R_1) \geq P(U \leq d). \tag{1.2.17}$$

<u>Proof</u>. From (1.2.15) and the remarks following those inequalities, all that remains to be shown is $[T_{(k)} \geq v]$ iff $[U \leq d]$ when all populations are identically distributed. But this is a direct consequence of (1.2.16) with $p=n$, $q=(k-1)n$.

Similar results also hold for rule R_2. Inequalities (1.2.14) hold when R_1 is replaced by R_2 and v and u are replaced by v' and u' where

(1.2.18) $$v'(c,k,n) = n(2kn-n+1)/2c,$$

(1.2.19) $$u'(c,k,n) = \frac{1}{2}\, kn(kn+1)[c(k-1)+1]^{-1}.$$

CHAPTER II

DISTRIBUTION THEORY

2.1. Introduction

In order to actually implement rules R_1 and R_2 defined by (1.2.4) and (1.2.5) it is necessary to obtain the appropriate constants which will guarantee the basic P* condition. The main objective of this chapter is to present exact and asymptotic methods for determining these constants.

In the second section exact expressions are derived for the means, variances and covariances of the T_i, the sum of ranks associated with population π_i. The assumption of continuity is the only condition imposed on the distribution of the random variables associated with each population. In the third section the distribution of the statistic $\max_{1 \leq j \leq k} T_j - T_i$ is considered for k=2, and for higher values of k when the observations are identically distributed and equal in number from each population. Asymptotic results are also given that allow the use of existing tables to approximate the d-constant required by rule R_1 under the conditions discussed in the remarks after (1.2.5). These approximate values are in close agreement with derived exact values. Asymptotic expressions are also given for the probability of a correct selection and for the expected subset size using rule R_1; an example is given where the underlying distribution is exponential. In Section

2.4, similar results are given for the statistic $\max_{1\leq j\leq k} T_j/T_i$ and the rule R_2. However, exact distribution results are not available for $k > 2$. Section 2.5 contains some exact distribution analysis of the statistic $\max_{1\leq j\leq k} Z_j/Z_i$, where Z_i is the product of ranks associated with π_i. A ranking and selection rule could be based on this statistic and easily used when the quantity kn is small.

In the case where an equal number of independent random variables are associated with k independent identically distributed populations, the statistic $\max_{1\leq j\leq k} T_j-T_i$ is stochastically equivalent to $T_i-\min_{1\leq j\leq k} T_j$; and $\max_{1\leq j\leq k} Z_j/Z_i$ is stochastically equivalent to $Z_i/\min_{1\leq j\leq k} Z_j$. If $k=2$, $\max_{1\leq j\leq k} T_j/T_i$ is stochastically equivalent to $T_i/\min_{1\leq j\leq k} T_j$, but for higher values of k this equivalence does not hold.

2.2 Some Moment Results on Rank Sums

Suppose $\pi_1,\ldots,\pi_k$ are k independent populations. With each population n independent random variables are associated and denoted by X_{ij}, $j=1,\ldots,n$, $i=1,\ldots,k$. The continuous distribution function of the X_{ij} is denoted by $F_i(x)$ for $j=1,\ldots,n$. Let T_i be the sum of ranks associated with population π_i.

The main objective in this section is to obtain exact expressions for the means, variances and covariances of the T_i, $i=1,2,\ldots,k$. These results will be useful in certain asymptotic expressions associated with the ranking and selection procedures introduced in Chapter I which are based on the rank sums T_i. The method of derivation is based on an extension of the U-statistic introduced by Mann and Whitney [42] for a two population problem and later extended to three populations by Whitney [63]. Define

$$(2.2.1) \qquad U_{ij}^{(m,\ell)} = \begin{cases} 1 \text{ if } x_{mi} < x_{\ell j};\ i,j=1,\dots,n,\ \ell=m+1,\dots,k,\ m=1,\dots,k-1 \\ 0 \text{ otherwise} \end{cases}$$

and

$$(2.2.2) \qquad U^{(r)} = \sum_{\ell=r+1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} U_{ij}^{(r,\ell)} + \sum_{m=1}^{r-1} \sum_{i=1}^{n} \sum_{j=1}^{n} (1-U_{ij}^{(m,r)}),\ r=1,\dots,k,$$

where $\sum_{\alpha=a}^{b} [\cdot] = 0$ if $b < a$.

The quantity $U^{(r)}$ represents the number of times an observation associated with π_r is less than an observation associated with any other population. The relation between $U^{(i)}$ and T_i is given in the first lemma.

Lemma 2.2.1. For $i=1,2,\dots,k$,

$$(2.2.3) \qquad T_i + U^{(i)} = n(2nk-n+1)/2.$$

Proof. Equation (2.2.3) follows from (1.2.16) by letting the observations from π_i correspond to the x's and all the other observations correspond to the y's.

Define

(2.2.4) $$c_{ij} = n^2 \int_{-\infty}^{\infty} [1-F_i(x)]dF_j(x), \quad i,j=1,\ldots,k.$$

Theorem 2.2.1. The expected value of T_1 is given by

(2.2.5) $$E(T_1) = n(2nk+1)/2 - \sum_{\ell=1}^{k} c_{\ell 1}.$$

Proof. From (2.2.2),

(2.2.6) $$E(U^{(1)}) = \sum_{\ell=2}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} P(X_{1i} < X_{\ell j}) = \sum_{\ell=2}^{k} c_{\ell 1}.$$

Since $c_{11} = n^2/2$,

(2.2.7) $$E(U^{(1)}) = \sum_{\ell=1}^{k} c_{\ell 1} - n^2/2.$$

Equation (2.2.5) now follows from Lemma 2.2.1.

Now $E(T_i)$ for a fixed i, $1 \leq i \leq k$, follows from (2.2.5) by a simple relabeling of distribution functions. Define the distribution functions $G_1,\ldots,G_k$ as follows: $G_1 = F_i$, $G_i = F_1$, $G_j = F_j$ for $j \neq 1,i$. Then the expected value of T_i is given by (2.2.5) with $c_{\ell 1}$ defined in terms of the distribution functions $G_1,\ldots,G_k$.

Corollary 2.2.1. (a) If $F_i(x) = F(x)$, i=1,...,k, then

(2.2.8) $$E(T_i) = n(nk+1)/2, \quad i=1,\ldots,k.$$

(b) If $F_1(x) \geq F_2(x) \geq \cdots \geq F_k(x)$ for all x, then

(2.2.9) $$E(T_1) \leq E(T_2) \leq \cdots \leq E(T_k).$$

<u>Proof</u>. (a) If $F_i(x) = F(x)$, $i=1,\ldots,k$, then $c_{ij} = n^2/2$ for all i,j. For (b),

$$E(T_{i+1}) - E(T_i) = \sum_{\ell=1}^{k} (c_{\ell i} - c_{\ell,i+1})$$

$$= n^2 \sum_{\ell=1}^{k} \int_{-\infty}^{\infty} (F_i(x) - F_{i+1}(x))dF_\ell(x)$$

$$\geq 0, \quad i = 1,\ldots,k-1.$$

From the remarks given above on relabeling of distribution functions to find $Var(T_i)$, $i=1,\ldots,k$, it suffices to find $Var(T_1)$. Let

(2.2.10) $$d_{ij\ell} = n^2 \int_{-\infty}^{\infty} [1-F_i(x)][1-F_j(x)]\,dF_\ell(x), \quad i,j,\ell=1,\ldots,k,$$

(2.2.11) $$h_{ij\ell} = n^2 \int_{-\infty}^{\infty} F_i(x)\,F_j(x)\,dF_\ell(x), \quad i,j,\ell=1,\ldots,k.$$

It can easily be shown that

(2.2.12) $$h_{ij\ell} = d_{ij\ell} - c_{j\ell} - c_{i\ell} + n^2.$$

Theorem 2.2.2. The variance of T_1 is given by

$$(2.2.13)\qquad \mathrm{Var}(T_1) = \sum_{\ell=2}^{k}[c_{\ell 1}-d_{\ell\ell 1}+(n-1)(h_{11\ell}-n^{-2}c_{\ell 1}^{2})] + \sum_{\ell=2}^{k}\sum_{s=2}^{k}(nd_{\ell s1}-n^{-1}\, c_{\ell 1}c_{s1}).$$

Proof. From (2.2.2),

$$(2.2.14)\qquad E[U^{(1)}]^2 = E[\sum_{\ell=2}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n}U_{ij}^{(1,\ell)}][\sum_{s=2}^{k}\sum_{q=1}^{n}\sum_{v=1}^{n}U_{qv}^{(1,s)}]$$

$$= \sum_{\ell,i,j} E[U_{ij}^{(1,\ell)}]^2 + \sum_{\substack{\ell,s,i,j\\ \ell\neq s}} E[U_{ij}^{(1,\ell)}U_{ij}^{(1,s)}]$$

$$+ \sum_{\substack{\ell,i,j,v\\ j\neq v}} E[U_{ij}^{(1,\ell)}U_{iv}^{(1,\ell)}] + \sum_{\substack{\ell,i,q,j\\ i\neq q}} E[U_{ij}^{(1,\ell)}U_{qj}^{(1,\ell)}]$$

$$+ \sum_{\substack{\ell,s,i,j,v\\ \ell\neq s,\, j\neq v}} E[U_{ij}^{(1,\ell)}U_{iv}^{(1,s)}] + \sum_{\substack{\ell,s,i,q,j\\ \ell\neq s,\ i\neq q}} E[U_{ij}^{(1,\ell)}U_{qj}^{(1,s)}]$$

$$+ \sum_{\substack{\ell,i,q,j,v\\ i\neq q,\ j\neq v}} E[U_{ij}^{(1,\ell)}U_{qv}^{(1,\ell)}] + \sum_{\substack{\ell,s,i,q,j,v\\ \ell\neq s,i\neq q,j\neq v}} E[U_{ij}^{(1,\ell)}U_{qv}^{(1,s)}]$$

which simplifies to

$$(2.2.15)\qquad E[U^{(1)}]^2 = \sum_{\ell=2}^{k}[c_{\ell 1}+(n-1)(d_{\ell\ell 1}+h_{11\ell})+n^{-2}(n-1)^2c_{\ell 1}^2]$$

$$+\sum_{\ell=2}^{k}\sum_{\substack{s=2\\ \ell\neq s}}^{k}[n\,d_{\ell s1}+n^{-1}(n-1)c_{\ell 1}c_{s1}].$$

Equation (2.2.13) now follows from (2.2.15) and (2.2.6) using the relations $\mathrm{Var}(T_1) = \mathrm{Var}(U^{(1)})=E[U^{(1)}]^2-E^2[U^{(1)}]$.

Corollary 2.2.2. If $F_i(x)=F(x)$, $i=1,\ldots,k$, then

$$(2.2.16)\qquad \mathrm{Var}(T_i) = n^2(k-1)(nk+1)/12,\quad i=1,\ldots,k,k\geq 1,$$

and the covariance between T_i and T_j is given by

$$(2.2.17)\qquad \mathrm{Cov}(T_i,T_j) = -n^2(nk+1)/12,\quad i\neq j,$$

and hence the correlation coefficient between T_i and T_j is

$$(2.2.18)\qquad \rho_{ij} = -(k-1)^{-1}.$$

Proof. If $F_i(x)=F(x)$ for all i, then $c_{ij}=n^2/2$, $d_{ij\ell}=n^2/3$, $h_{ij\ell}=n^2/3$ for all i,j,ℓ. Substituting these values in (2.2.13) and simplifying yields (2.2.16). Since $T_1,\ldots,T_k$ sum to a constant and are identically distributed when $F_i(x)=F(x)$, $i=1,\ldots,k$, the relation $\mathrm{Var}(\sum_{i=1}^{k} T_i)=\sum_{i=1}^{k}\mathrm{Var}(T_i)+2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\mathrm{Cov}(T_i,T_j)$ implies $-kn^2(k-1)(nk+1)/12$ $=k(k-1)\,\mathrm{Cov}(T_i,T_j)$, $i\neq j$. Equation (2.2.17) follows upon simplification.

This completes the proof.

Equations (2.2.8), (2.2.16) and (2.2.17) are well known. Assume $1 \leq r < s \leq k$. An expression for $\mathrm{Cov}(T_1,T_k)$ will now be derived in the general case. To obtain $\mathrm{Cov}(T_r,T_s)$ interchange distribution functions as indicated earlier and then apply the results of the case where $r=1$, $s=k$.

$$\text{(2.2.19)}\quad E[U^{(1)}U^{(k)}] = \sum_{\ell=2}^{k}\sum_{m=1}^{k-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n} E[U_{ij}^{(1,\ell)}(1-U_{pq}^{(m,k)})].$$

In order to use independence partition as follows:

$$\text{(2.2.20)}\quad \sum_{\ell,m,i,j,p,q} = \sum_{\substack{\ell,m,i,j,p,q\\ m\neq 1,\ell\neq k,\ell=m}} + \sum_{\substack{\ell,m,i,j,p,q\\ m\neq 1,\ell=k,\ell\neq m}} + \sum_{\substack{\ell,m,i,j,p,q\\ m\neq 1,\ell\neq k,\ell\neq m}} + \sum_{\substack{\ell,m,i,j,p,q\\ m=1,\ell\neq k,\ell\neq m}}$$

$$+ \sum_{\substack{\ell,m,i,j,p,q\\ m=1,\ell=k,\ell\neq m}}$$

$$\equiv \sum\nolimits_1 + \sum\nolimits_2 + \sum\nolimits_3 + \sum\nolimits_4 + \sum\nolimits_5$$

$$\sum\nolimits_1 = \sum_{j=p}\nolimits_1 + \sum_{j\neq p}\nolimits_1$$

$$\sum\nolimits_2 = \sum_{j\neq q}\nolimits_2 + \sum_{j=q}\nolimits_2$$

$$\sum\nolimits_4 = \sum_{i\neq p}\nolimits_4 + \sum_{i=p}\nolimits_4$$

$$\sum_5 = \sum_{i=p, j\neq q}^{5} + \sum_{i\neq p, j=q}^{5} + \sum_{i\neq p, j\neq q}^{5} .$$

These calculations yield

(2.2.21) $$E[U^{(1)}U^{(k)}]=n^3(k-2)+n^{-2}(n-1)^2\ c_{k1}c_{1k}+(n-1)(n^2-h_{kk1}-h_{11k})$$
$$+\sum_{\ell=2}^{k-1}[n(h_{1k\ell}-h_{1\ell k}-h_{k\ell 1})+n^{-1}(n-1)(c_{\ell 1}c_{\ell k}+c_{k1}c_{\ell k}+c_{1k}c_{\ell 1})]$$
$$+\sum_{\ell=2}^{k-1}\sum_{\substack{m=2\\ m\neq\ell}}^{k-1} c_{\ell 1}c_{mk}.$$

Now $Cov(T_1,T_k)=Cov(U^{(1)},U^{(k)})=E[U^{(1)}U^{(k)}]-E(U^{(1)})E(U^{(k)})$, so from (2.2.21) and (2.2.6) Theorem 2.2.3 follows.

<u>Theorem 2.2.3</u>. The covariance of T_1 and T_k is given by

(2.2.22) $$Cov(T_1,T_k) = n^3(k-2)+n^{-2}(1-2n)c_{k1}c_{1k}+(n-1)(n^2-h_{kk1}-h_{11k})$$
$$+\sum_{\ell=2}^{k-1}[n(h_{1k\ell}-h_{1\ell k}-h_{k\ell 1})-n^{-1}(c_{\ell 1}c_{\ell k}+c_{k1}c_{\ell k}+c_{1k}c_{\ell 1})].$$

It is easy to show that (2.2.22) reduces to (2.2.17) when $F_i(x) = F(x)$, $i=1,\ldots,k$.

A common sample size from each population was not crucial in any of the above derivations. Asymptotic results of this nature are available for more general scoring procedures in Puri [53]. However, his expressions do not yield the exact moments for finite sample sizes. For $k=2$, equations (2.2.5) and (2.2.13) can be obtained from Wilks ([64], pg. 461).

Suppose $F_i(x)=F(x-\theta_i)$, $i=1,\ldots,k$, where each θ_i belongs to some interval, Θ, on the real line. Let $\Delta_{ij}=\theta_j-\theta_i$. Then

$$c_{ij}=n^2\int_{-\infty}^{\infty}[1-F(x+\Delta_{ij})]dF(x),\quad d_{ij\ell}=n^2\int_{-\infty}^{\infty}[1-F(x+\Delta_{i\ell})][1-F(x+\Delta_{j\ell})]dF(x).$$

Thus, for a given continuous distribution $F(x)$, the means, variances and covariances of the rank sums depend only on the quantities Δ_{ij}, $i,j=1,\ldots,k$. Similarly, if $F_i(x)=F(x/\theta_i)$, $\theta_i>0$, $i=1,\ldots,k$, then the previous statement again applies with $\Delta_{ij}=\theta_j/\theta_i$, $i,j=1,\ldots,k$.

2.3. The Distribution of $\max\limits_{1\le j\le k} T_j-T_i$.

A. Case k=2

Suppose $k=2$ and the random variables associated with π_i have a distribution $F_i(x)$ with density $f_i(x)$, $i=1,2$. For Theorem 2.3.1 the rank sums can be based on an unequal number of observations, so assume $n_i(\geq 1)$ observations from π_i. For $d\geq 0$,

$$\text{(2.3.1)}\qquad P[\max_{j=1,2} T_j-T_2\le d] = P\{T_1\le [d+(n_1+n_2)(n_1+n_2+1)/2]/2\}\ .$$

Theorem 2.3.1 below specifies the distribution of T_1 and hence that of $\max\limits_{j=1,2} T_j-T_2$. The moment generating function of T_1 has been obtained by Dwass [20].

Theorem 2.3.1. The mass function of T_1 is given by

$$(2.3.2)\quad P[T_1=x]=\begin{cases} n_1!\int_{-\infty}^{\infty}\int_{-\infty}^{x_{n_1}}\cdots\int_{-\infty}^{x_2} P[\sum_{j=1}^{n_1}(n_1-j+1)N_j=x-n_1(n_1+1)/2]\prod_{i=1}^{n_1} f_1(x_i)dx_i, \\ \qquad x=n_1(n_1+1)/2,\ldots,n_1(n_1+2n_2+1)/2 \\ 0 \quad , \text{ otherwise} \end{cases}$$

where N_j, $j=1,\ldots,n_1$, are random variables having the joint mass function

$$(2.3.3)\qquad P[N_1=y_1,\ldots,N_{n_1}=y_{n_1}] = \binom{n_2}{y_1,\ldots,y_{n_1},y_{n_1+1}}\prod_{i=1}^{n_1+1} p_i^{y_i},$$

where $p_i=F_2(x_i)-F_2(x_{i-1})$, $i=1,\ldots,n_1+1$; $x_o=-\infty$, $x_{n_1+1}=\infty$, and $y_1,\ldots,y_{n_1+1}$ are nonnegative integers which sum to n_2.

Proof. Let $x_{1[1]}\leq\cdots\leq x_{1[n_1]}$ be the ordered observations from π_1, and let $x_{1[j]}$ have rank $R_{1[j]}$ in the combined sample. Then $R_{1[j]}$ is the number of observations not greater than $x_{1[j]}$.

$$P[T_1=x]=P[\sum_{j=1}^{n_1} R_{1[j]}=x]$$

$$=n_1!\int_{-\infty}^{\infty}\int_{-\infty}^{x_{n_1}}\cdots\int_{-\infty}^{x_2} P[\sum_{j=1}^{n_1} R_{1[j]}=x|X_{1[1]}=x_1,\ldots,X_{1[n_1]}=x_{n_1}]\cdot \prod_{i=1}^{n_1} f_1(x_i)dx_i.$$

Given the order statistics $X_{1[j]}$, $R_{1[j]}=j+N_j^*$, $j=1,\ldots,n_1$, where N_j^* is the number of observations from π_2 not greater than x_j. Thus

$$\sum_{j=1}^{n_1} R_{1[j]} = n_1(n_1+1)/2 + \sum_{j=1}^{n_1} N_j^*$$

$$= n_1(n_1+1)/2 + \sum_{j=1}^{n_1} (n_1-j+1)N_j,$$

where N_j is the number of observations from π_2 belonging to the interval $(x_{j-1}, x_j]$, $x_o = -\infty$. Thus (2.3.2) holds and (2.3.3) is easily verified.

In the case where $F_1(x)=F_2(x)$, the T_i are identically distributed if $n_1=n_2$. The appropriate d value for rule R_1 can then be obtained in the following manner. Note that obtaining this d value is equivalent to obtaining a c value for rule R_2, since these two rules are equivalent for k=2. Let $n_1=n_2=n \geq 1$.

$$(2.3.4) \qquad P[T_1 \geq \max_{j=1,2} T_j - d] = P\{T_1 \geq [n(2n+1)-d]/2\}$$

$$= P[U^{(2)} \geq (n^2-d)/2],$$

where $U^{(2)}$ is defined by (2.2.2). In this case, $U^{(2)}$ is the Mann-Whitney U-statistic. If $L(x)$ is the smallest integer not less than x, then

$$(2.3.5) \qquad P[T_1 \geq \max_{j=1,2} T_j - d] = P[U^{(2)} \geq L((n^2-d)/2)].$$

For rule R_1 to satisfy the P* condition (1.1.5), the d value must be chosen so that

(2.3.6) $$P[U^{(2)} \geq L((n^2-d)/2)] \geq P^*.$$

Thus, one seeks the greatest integer u such that

(2.3.7) $$P[U^{(2)} \leq u] \leq 1-P^*.$$

This u value is given in Milton [43] for $P^* = .90, .95, .975, .99, .995, .9975, .999, .9995$ and $n=2(1)20$, and in Owen [47] the distribution of $U^{(2)}$ is given for $n=2(1)10$. Then the desired d is the smallest integral solution of

(2.3.8) $$u = L((n^2-d)/2) - 1,$$

which is

(2.3.9) $$d = n^2-2(u+1).$$

These d solutions are given in Table 2 where they are compared with asymptotic solutions.

B. Case $k \geq 3$

Suppose $k=3$ and that the independent random variables X_{ij}, $j=1,\ldots,n_i$, $i=1,2,3$ have a continuous distribution $F(x)$. To form the rank sums, T_1, T_2 and T_3, all the observations are ordered to obtain the ranks R_{ij}. The same results are obtained if each observation in this ordered sample is replaced by an i if it came from population π_i. Now one has only to consider a sequence of length $S=n_1+n_2+n_3$ consisting of n_1 1's, n_2 2's and n_3 3's. Since the random variables are identically distributed, each of the $\binom{S}{n_1,n_2,n_3}$ different sequences are equally likely. Hence, to find $P[T_1 \geq \max_{1\leq j\leq 3} T_j - d]$, it suffices to count the number of sequences which possess the attribute $[T_2-T_1 \leq d,\ T_3-T_1 \leq d]$. The recursion formula presented here is of the same type as that given by Odeh [46] in tabulating the distribution of the maximum rank sum. Let

(2.3.10) $$S = n_1 + n_2 + n_3,$$

and define

(2.3.11) $$N(n_1,n_2,n_3|m_2,m_3) = \text{number of sequences in which } T_2-T_1 \leq m_2 \text{ and } T_3-T_1 \leq m_3.$$

Then by conditioning on the parent population of the last element in a sequence, the following recursion formula is obtained:

(2.3.12) $$N(n_1,n_2,n_3|m_2,m_3)=N(n_1-1,n_2,n_3|m_2+S,m_3+S)+N(n_1,n_2-1,n_3|m_2-S,m_3)$$
$$+N(n_1,n_2,n_3-1|m_2,m_3-S),$$

with the boundary conditions:

1) If for any $i \geq 2$, $m_i < [n_i(n_i+1)-n_1(1+2S-n_1)]/2$, then $N(n_1,n_2,n_3|m_2,m_3) = 0$.

2) If for every $i \geq 2$, $m_i \geq [n_i(1+2S-n_i)-n_1(n_1+1)]/2$, then $N(n_1,n_2,n_3|m_2,m_3) = \binom{S}{n_1,n_2,n_3}$.

3) $N(0,n_2,n_3|m_2,m_3)$ = number of sequences of n_2 2's and n_3 3's such that $\frac{1}{2}S(S+1)-m_3 \leq T_2 \leq m_2$, so

(a) if $\frac{1}{2}S(S+1)-m_3 > m_2$, $N(0,n_2,n_3|m_2,m_3) = 0$,

(b) if $m_2 < n_2(n_2+1)/2$, $N(0,n_2,n_3|m_2,m_3) = 0$,

(c) if $m_3 < n_3(n_3+1)/2$, $N(0,n_2,n_3|m_2,m_3) = 0$,

(d) if (a) through (c) do not hold, the term can be evaluated from a Mann-Whtney tables.

4) $N(n_1,0,n_3|m_2,m_3)$ = number of sequences of n_1 1's and n_3 3's such that $T_1 \geq \max\{-m_2, L(S(S+1)/4-m_3/2)\} \equiv M$, where $L(x)$ is the smallest integer not less than x, so

(a) if $M > n_1(n_1+2\,n_3+1)/2$, $N(n_1,0,n_3|m_2,m_3) = 0$,

(b) if $M \leq n_1(n_1+1)/2$, $N(n_1,0,n_3|m_2,m_3) = \binom{S}{n_1}$,

(c) if (a) and (b) fail to hold, the term can be evaluated from a Mann-Whitney tables.

5) $N(n_1,n_2,0|m_2,m_3)$ = number of sequences of n_1 1's and n_2 2's such that $T_1 \geq \max\{-m_3, L(S(S+1)/4-m_2/2)\} \equiv M$, so

(a) if $M > n_1(n_1+2n_2+1)/2$, $N(n_1,n_2,0|m_2,m_3) = 0$,

(b) if $M \leq n_1(n_1+1)/2$, $N(n_1,n_2,0|m_2,m_3) = \binom{S}{n_1}$,

(c) if (a) and (b) fail to hold, the term can be evaluated from a Mann-Whitney tables.

The following symmetry holds:

(2.3.13) $$N(n_1,n_2,n_3|m_2,m_3) = N(n_1,n_3,n_2|m_3,m_2).$$

Thus, at an ''equal n_i, equal m_i stage,'' equation (2.3.12) can be written as

(2.3.14) $N(n,n,n|m,m)=N(n-1,n,n|m+3n,m+3n)+2N(n,n-1,n|m-3n,m).$

In order to get $P[T_1 \geq \max_{1\leq j\leq 3} T_j - d]$ for values of $d \geq 0$, one uses

$$P[T_1 \geq \max_{1\leq j\leq 3} T_j - d] = N(n_1,n_2,n_3|d,d)\binom{S}{n_1,n_2,n_3}^{-1}. \tag{2.3.15}$$

A recursion formula similar to (2.3.12) can be written for an arbitrary number of populations. Equation (2.3.12) was programmed for an IBM 7094 and $N(n,n,n|m,m)$ was computed for $n=2,3,4$, $m=0,1,\ldots,2n^2$. Partial results were obtained for $n=5$. Using (2.3.15), $P[T_1 \geq \max_{1\leq j\leq 3} T_j - d]$ was then obtained to four decimal places, the fifth being rounded. These computations compose Table 1.

C. Asymptotic Theory

Let $\pi_1,\ldots,\pi_k$ be k independent populations. The associated random variables X_{ij}, $j=1,\ldots,n$, $i=1,\ldots,k$ are assumed independent and to have a continuous distribution function $F_i(x)$, $j=1,\ldots,n$. As usual, let T_i be the rank sum associated with π_i. Define

$$\mu_i = E(T_i), \quad i = 1,\ldots,k \tag{2.3.16}$$

$$\sigma_i^2 = Var(T_i), \quad i = 1,\ldots,k \tag{2.3.17}$$

$$\sigma_{ij} = Cov(T_i,T_j), \quad i,j=1,\ldots,k;\ i\neq j. \tag{2.3.18}$$

TABLE 1

$P[T_1 \geq \max_{1 \leq j \leq 3} T_j - m]$ for Independent Identically Distributed Populations with n Observations from Each

m \ n	2	3	4
0	.42222	.35714	.35405
1	.48889	.41786	.38961
2	.60000	.48095	.43169
3	.68889	.54286	.47105
4	.77778	.59762	.51180
5	.84444	.66190	.55302
6	.93333	.71786	.59417
7	.97778	.77024	.63157
8	1.00000	.81786	.67169
9		.85595	.70707
10		.88690	.74193
11		.91905	.77478
12		.94286	.80681
13		.96309	.83423
14		.97857	.86124
15		.99048	.88358
16		.99643	.90436
17		.99881	.92214
18		1.00000	.93784
19			.95053
20			.96214
21			.97097
22			.97870
23			.98488
24			.98978
25			.99348
26			.99619
27			.99781
28			.99896
29			.99954
30			.99983
31			.99994
32			1.00000

m \ n	5
22	.89929
23	.91304
24	.92520
25	.93611
26	.94596
27	.95467
28	.96213
29	.96877
30	.97443
31	.97925
32	.98339
33	.98689
34	.98976
35	.99216
36	.99411
37	.99565
38	.99689
39	.99785
40	.99855
41	.99906
42	.99942
43	.99966
44	.99980
45	.99989
46	.99995
47	.99998
48	.99999
49	.99999
50	1.00000

The quantities μ_i, σ_i^2 and σ_{ij} are given by (2.2.5), (2.2.13) and (2.2.22). Lemma 2.3.1 is given in Puri [53] with asymptotic expressions for μ_i, σ_i^2 and σ_{ij}. In the case where $F_i(x)=F(x)$, $i=1,\ldots,k$, this lemma can also be found in Dunn [18], the moments being given by (2.2.8), (2.2.16) and (2.2.17). An equal number of observations from each population is not required in either of these papers.

Lemma 2.3.1. The random vector $(\sigma_1^{-1}(T_1-\mu_1),\ldots,\sigma_k^{-1}(T_k-\mu_k))$ has a limiting normal distribution (as $n \to \infty$) with a zero mean vector and variance-covariance matrix $M=(m_{ij})$, where $m_{ii}=1$, $i=1,\ldots,k$ and $m_{ij}=\lim_{n\to\infty} \sigma_{ij}(\sigma_i\sigma_j)^{-1}$, $i,j=1,\ldots,k$, $i\neq j$.

First consider the case where $F_i(x) = F(x)$, $i=1,\ldots,k$. Let $W_i=U_i-\overline{U}$, $i=1,\ldots,k$, where U_i are independent standard normal variables and $\overline{U}$ is their mean. Then $E(W_i)=0$, $E(W_iW_j)=\delta_{ij}-k^{-1}$, where δ_{ij} is the Kronecker delta. Thus the joint distribution of $(W_1,\ldots,W_k)$ is the singular normal distribution with zero mean vector and variance-covariance matrix $\Sigma_W=(\delta_{ij}-k^{-1})$. Let

$$(2.3.19) \qquad V_i = (1-k^{-1})^{1/2}\, \sigma_i^{-1}(T_i-\mu_i)$$

$$=[T_i-n(nk+1)/2][n^2k(nk+1)/12]^{-1/2}, \quad i=1,\ldots,k.$$

Then

$$(2.3.20) \qquad E(V_i) = 0, \quad i = 1,\ldots,k$$

$$(2.3.21) \qquad E(V_iV_j) = \delta_{ij} - k^{-1}, \quad i,j=1,\ldots,k.$$

Now combining the last few remarks with Lemma 2.3.1, a statement given in Odeh [46] is obtained as the next lemma.

Lemma 2.3.2. The asymptotic joint distribution of $(V_1,\ldots,V_k)$ is the distribution of $(W_1,\ldots,W_k)$.

The next lemma follows from a result of Rubin [58] and Lemma 2.3.2.

Lemma 2.3.3. Let $g(W_1,\ldots,W_k)$ be a function such that its set of discontinuity points has probability zero under the normal distribution with zero mean vector and variance-covariance matrix Σ_W. Then the asymptotic distribution of $g(V_1,\ldots,V_k)$ is the distribution of $g(W_1,\ldots,W_k)$.

Now define

$$c=c(n,k) = n[k(nk+1)/12]^{1/2}. \tag{2.3.22}$$

Then the asymptotic distribution of $\max\limits_{1\leq j\leq k} T_j-T_k$ is given by the following theorem.

Theorem 2.3.2. Let X_{ij}, $j=1,\ldots,n$, $i=1,\ldots,k$, be independent identically distributed random variables with a continuous distribution function. Then

$$P[T_k \geq \max_{1\leq j\leq k} T_j-d] \approx \int_{-\infty}^{\infty}[\Phi(x+d/c)]^{k-1}\,\varphi(x)dx,\ d\geq 0, \tag{2.3.23}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the cumulative distribution function and density of a standard normal random variable, respectively.

Proof. For $d \geq 0$,

$$P[T_k \geq \max_{1\leq j\leq k} T_j-d] = P[\max_{1\leq j\leq k-1} V_j-V_k \leq d/c].$$

Using Lemma 2.3.2 and Lemma 2.3.3,

$$P[\max_{1\leq j\leq k} V_j - V_k \leq d/c] \approx P[\max_{1\leq j\leq k-1} W_j - W_k \leq d/c].$$

Now,

$$P[\max_{1\leq j\leq k} W_j - W_k \leq d/c] \doteq P[U_j - U_k \leq d/c,\ j = 1,\ldots,k-1]$$

$$= \int_{-\infty}^{\infty} [\Phi(x+d/c)]^{k-1} \varphi(x)dx,$$

which completes the proof.

Integrals of the type appearing on the right hand side of equation (2.3.23) have been considered by Gupta [29]. His Table I gives h values satisfying the equation

$$(2.3.24) \qquad \int_{-\infty}^{\infty} [\Phi(x+h/2)]^{k-1} \varphi(x)dx = P^*$$

for $P^* = .99, .975, .95, .90, .75$ and $k = 2(1)51$. If $\tilde{d}$ denotes the value of d based on the normal approximation, then from (2.3.24) one obtains

$$(2.3.25) \qquad \tilde{d} = hn[k(nk+1)/6]^{1/2},$$

h being the entry in Gupta's Table I corresponding to the given P^* and k.

Remarks: (1) By using (2.3.25) one can obtain an asymptotic value of d to use in rule R_1 when a slippage configuration in Ω exists (see remarks after (1.2.5)) for k=2(1)51 and for any common sample size n, n large. (2) In general $\tilde{d}$ will not be an integer. So for the solution the smallest integer not less than $\tilde{d}$, $L(\tilde{d})$, should be taken. This method was used to calculate an asymptotic value of d for k=2(1)5, n=2(1)25 and P* = .99,.975,.95,.90,.75. These results are presented in Table 2. Exact d values are given in parentheses where they are known. In most cases where the asymptotic value and exact value do not agree, the asymptotic value is larger and hence a conservative constant for the rule R_1. From the values given in this table it is seen that $-1 \leq L(\tilde{d})-d \leq 3$ for k=2 and $0 \leq L(\tilde{d})-d \leq 3$ for k=3.

Now consider the more general case where the $F_i(x)$, i=1,...,k, may not all be identical. For large n, the distribution of $\underline{T}'=(T_1,\ldots,T_k)$ is approximately a multivariate normal distribution with mean vector $\underline{\mu}_T' =(\mu_1,\ldots,\mu_k)$ and variance-covariance matrix $\Sigma_T= (\sigma_{ij})$. Let

(2.3.26) $$\underline{W} = A\underline{T},$$

where A is a (k-1) x k matrix defined by

(2.3.27) $$A = \begin{pmatrix} 1 & 0 & 0 \ldots 0 & -1 \\ 0 & 1 & 0 \ldots 0 & -1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot \ldots 1 & -1 \end{pmatrix} .$$

(2.3.28) $$W_i = T_i-T_k, \quad i = 1,\ldots,k-1.$$

TABLE 2

Asymptotic d-Solution of $P(T_1 \geq \max_{1 \leq j \leq k} T_j - d) \geq P^*$.

Exact Value is Given As (d)

k = 2

n \ P*	.99	.975	.95	.90	.75
2	7	6	5	4	2(2)
3	11	9	8(7)	6(5)	4(3)
4	17	14(14)	12(12)	9(8)	5(4)
5	23(21)	19(19)	16(15)	13(13)	7(7)
6	30(28)	25(24)	21(20)	17(16)	9(8)
7	37(35)	31(31)	26(25)	21(21)	11(11)
8	45(44)	38(36)	32(32)	25(24)	13(14)
9	53(51)	45(45)	38(37)	30(29)	16(15)
10	62(60)	52(52)	44(44)	34(34)	18(18)
11	71(69)	60(59)	51(51)	40(39)	21
12	81(80)	68(68)	57(58)	45(44)	24
13	91(89)	77(77)	65(65)	50(51)	27
14	102(100)	86(84)	72(72)	56(56)	30
15	113(111)	95(95)	80(79)	62(63)	33
16	124(122)	105(104)	88(88)	69(68)	36
17	136(133)	114(113)	96(95)	75(75)	40
18	148(146)	124(124)	104(104)	82(82)	43
19	160(157)	135(133)	113(113)	88(89)	47
20	172(170)	145(144)	122(122)	95(96)	50
21	185	156	131	102	54
22	199	168	141	110	58
23	212	179	150	117	62
24	226	191	160	125	66
25	240	203	170	133	70

For given values of k,n,p*, this table gives the asymptotic smallest integer d which satisfies $P[T_k \geq \max_{1 \leq j \leq k} T_j - d] \geq P^*$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k which are independent identically distributed. Exact d values, where known, are given in parentheses.

TABLE 2 (Continued)

k = 3

n \ P*	.99	.975	.95	.90	.75
2	10(8)	9(7)	8(7)	6(6)	4(4)
3	18(15)	15(14)	13(13)	11(11)	7(7)
4	27(25)	23(22)	20(19)	17(16)	11(11)
5	37(35)	32(31)	28(27)	23(23)	15(15)
6	48	41	36	30	19
7	60	52	45	37	24
8	73	63	55	45	29
9	87	75	65	54	35
10	101	88	76	63	40
11	117	101	87	72	46
12	133	115	99	82	53
13	149	129	112	92	59
14	167	144	125	103	66
15	185	160	138	114	73
16	203	176	152	125	81
17	222	192	167	137	88
18	242	209	181	149	96
19	262	227	197	162	104
20	283	245	212	175	112
21	304	263	228	188	121
22	326	282	244	201	130
23	349	301	261	215	138
24	371	321	278	229	148
25	395	341	296	244	157

For given values of k,n,P*, this table gives the asymptotic smallest integer d which satisfies $P[T_k \geq \max_{1\leq j\leq k} T_j - d] \geq P^*$. The rank sums T_i, i=1,...,k are based on random variables X_{ij}, j=1,...,n, i=1,...,k which are independent identically distributed.

TABLE 2 (Continued)

$k = 4$

n \ P*	.99	.975	.95	.90	.75
2	14	12	11	9	6
3	24	21	19	16	11
4	37	32	28	24	17
5	51	44	39	33	23
6	66	58	51	43	30
7	83	73	64	54	37
8	101	89	78	66	45
9	121	106	93	78	54
10	141	123	108	91	63
11	162	142	125	105	72
12	185	162	142	119	82
13	208	182	160	135	92
14	232	203	179	150	103
15	257	225	198	166	114
16	283	248	218	183	126
17	310	271	238	201	138
18	338	296	260	218	150
19	366	320	281	237	162
20	395	346	304	255	175
21	425	372	327	275	188
22	456	399	350	295	202
23	487	426	374	315	216
24	519	454	399	335	230
25	551	483	424	356	244

For given values of k,n,P*,this table gives the asymptotic smallest integer d which satisfies $P[T_k \geq \max_{1 \leq j \leq k} T_j - d] \geq P^*$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k which are independent identically distributed.

TABLE 2 (Continued)

k = 5

n \ P*	.99	.975	.95	.90	.75
2	17	15	14	12	8
3	31	27	24	21	15
4	47	41	37	31	22
5	65	57	51	43	31
6	85	75	66	57	40
7	107	94	83	71	51
8	130	115	102	86	62
9	155	137	121	103	73
10	181	160	141	120	86
11	209	184	163	139	99
12	238	209	185	158	112
13	268	236	209	178	126
14	299	263	233	198	141
15	331	292	258	220	156
16	365	322	284	242	172
17	399	352	311	265	188
18	435	383	339	289	205
19	472	415	368	313	222
20	509	449	397	338	240
21	548	482	427	363	258
22	587	517	458	389	277
23	627	553	489	416	296
24	669	589	521	443	315
25	711	626	554	471	335

For given values of k,n,P*, this table gives the asymptotic smallest integer d which satisfies $P\{T_k \geq \max_{1\leq j\leq k} T_j - d\} \geq P^*$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k which are independent identically distributed.

It then follows that for large n, $\underline{W}' = (W_1,\dots,W_{k-1})$ is approximately distributed as a multivariate normal random vector with mean vector $\mu_W = A\mu_T$ and variance-covariance matrix $\Sigma_W = A\Sigma_T A'$. These remarks are collected in the next theorem.

Theorem 2.3.3. If Σ_W is nonsingular, then for $d \geq 0$

(2.3.29) $$P(T_k \geq \max_{1\leq j\leq k} T_j - d) \simeq K \int_{-\infty}^{d}\cdots\int_{-\infty}^{d} \exp[-(\underline{w}-\mu_W)'\Sigma_W^{-1}(\underline{w}-\mu_W)/2] \prod_{i=1}^{k-1} dw_i,$$

where

(2.3.30) $$K = [(2\pi)^{k-1}|\Sigma_W|]^{-1/2} .$$

A theorem similar to Theorem 2.3.3 can be written for the expected subset size. Let S be the number of populations retained in the chosen subset by a particular rule under consideration, and let S* be the number of non-best populations retained. Then S and S* are random variables taking on values 1 to k and 0 to k-1 respectively, and their expected values provide a criterion of the efficiency of the procedure under consideration. For any ranking and selection rule R, the following relation holds:

(2.3.31) $$E(S|R) = E(S^*|R) + P(CS|R).$$

For $\nu = 1,\dots,k$, define

(2.3.32) $$\underline{W}^{\nu} = A_{\nu}\,\underline{T},$$

where A_ν is the $(k-1)xk$ matrix obtained from matrix A defined in (2.3.27) by moving column j to column j+1, $j=\nu, \nu+1,\ldots,k-1$, and replacing column ν by column k. The matrix A_k is matrix A. Thus

$$W_i^\nu = T_i - T_\nu, \quad i = 1,\ldots,k, \; i \neq \nu. \tag{2.3.33}$$

The random vector $\underline{W}^\nu$ is asymptotically distributed as a multivariate normal random vector with mean vector $\underline{\mu}_\nu = A_\nu \underline{\mu}_T$ and variance-covariance matrix $\Sigma_\nu = A_\nu \Sigma_T A_\nu'$ Theorem 2.3.4 now follows after observing $E(S|R) = \sum_{\nu=1}^{k} P[\pi_\nu \text{ is a selected population}]$.

Theorem 2.3.4. If Σ_ν is nonsingular for $\nu = 1,\ldots,k$, then

$$E(S|R_1) \approx \sum_{\nu=1}^{k} K_\nu \int_{-\infty}^{d} \cdots \int_{-\infty}^{d} \exp[-(\underline{w}^\nu - \underline{\mu}_\nu)' \Sigma_\nu^{-1} (\underline{w}^\nu - \underline{\mu}_\nu)/2] \prod_{\substack{i=1 \\ i\neq\nu}}^{k} dw_i^\nu, \tag{2.3.34}$$

where

$$K_\nu = [(2\pi)^{k-1} |\Sigma_\nu|]^{-1/2} . \tag{2.3.35}$$

Now an application of Theorem 2.3.3 will be considered for k=3. A similar application of Theorem 2.3.4 could also be developed. Suppose π_1, π_2, π_3 are three independent populations with associated random variables X_{ij}, $j=1,\ldots,n$, $i=1,2,3$. Let X_{ij}, $j=1,\ldots,n$, have the density

$$f(x-\theta_i) = \begin{cases} e^{-(x-\theta_i)} , & x > \theta_i , \quad i = 1,2,3 \\ 0 , & x \leq \theta_i \end{cases} . \tag{2.3.36}$$

Let $\Theta = [0,\infty)$ and $\theta_{[1]} = 0$. The quantities c_{ij}, $d_{ij\ell}$ and $h_{ij\ell}$ defined by (2.2.4), (2.2.10) and (2.2.11) can now be explicitly obtained as functions of $\theta_{[2]}$ and $\theta_{[3]}$, and hence so can the means, variances and covariances of T_i, i=1,2,3. Using Theorem 2.3.3 an asymptotic value for the probability of a correct selection can be obtained as follows:

$$(2.3.37) \qquad P(CS|R_1) = P[W_1 \leq d,\ W_2 \leq d]$$

$$\simeq L(\sigma_{W_1}^{-1}(d-\mu_{W_1}),\ \sigma_{W_2}^{-1}(d-\mu_{W_2}),\ \mathrm{cor}(W_1,W_2))$$

$$+[\alpha(\sigma_{W_1}^{-1}(d-\mu_{W_1})) + \alpha(\sigma_{W_2}^{-1}(d-\mu_{W_2}))]/2,$$

where

$$(2.3.38) \qquad L(p,q,r) = P(X > p,\ Y > q),$$

X,Y being bivariate normal with zero means, unit variances and covariance r, and

$$(2.3.39) \qquad \alpha(x) = \int_{-x}^{x} (2\pi)^{-1/2}\, e^{-t^2/2}\, dt.$$

The L(p,q,r) terms can be found in [45] and the $\alpha(x)$ terms in [44]. The L terms cause some difficulty in accuracy since interpolation must be used on p,q and the correlation coefficient. In the cases considered here, accuracy to three places is obtained.

Table 3 is the evaluation of the right hand side of (2.3.37) for n = 20,50 and given $\underline{\theta}$ configurations. The d values used are the asymptotic d values obtained using the method discussed in Section 2.3.C, and subject to $P(CS|R_1) \geq P^*$ when $\theta_i = 0$, i=1,2,3, $P^* = .99, .975, .95, .90, .75$. It should be pointed out that $\underline{\theta} = (0,0,0)$ has not been shown to be that $\underline{\theta}$ configuration which minimizes the probability of a correct selection over the entire parameter space Ω. However, Table 3 indicates that this may be the case.

2.4. The Distribution of $\max_{1 \leq j \leq k} T_j/T_i$

In order to actually implement rule R_2 defined by (1.2.5), the c value must be specified. This will be the major concern of this section. Assume T_i, the rank sum associated with population π_i, based on n observations from each population, and that the random variables associated with π_i have a continuous distribution $F_i(x)$, i=1,...,k. When the distribution functions are not all identical, the distribution of $\max_{1 \leq j \leq k} T_j/T_i$ for k=2 is obtained in an exact form directly from Theorem 2.3.1. For $k > 2$, asymptotic results are presented in this section.

If $F_i(x) = F(x)$, i=1,...,k, then the distribution of $\max_{1 \leq j \leq k} T_j/T_i$ is independent of the underlying distribution of the random variables. From Theorem 1.2.1 this is the configuration which minimizes the probability of a correct selection given a slippage type problem. The appropriate c value is then obtained (or approximated) from the following analysis. Asymptotic expressions for the probability of a correct selection and expected subset size will also be given.

First consider the case $F_i(x) = F(x)$, i = 1,...,k. Let W_i be as before, i.e., $W_i = U_i - \bar{U}$, i=1,...,k, where U_i are independent standard

TABLE 3

Asymptotic Value of $P(CS|R_1)$ for Three Independent Exponential Populations

$\theta_{[2]},\theta_{[3]}$ \ P^*	.99	.975	.95	.90	.75
1.0, 2.0	1.0 1.0	1.0 1.0	1.0 1.0	1.0 1.0	1.0 1.0
1.0, 1.0	.9994 .9994	.9975 .9975	.9923 .9924	.9773 .9771	.8998 .9005
.5, 1.0	1.0 1.0	1.0 1.0	1.0 1.0	1.0 1.0	.9999 1.0
.5, .5	.9984 .9984	.9946 .9946	.9862 .9863	.9656 .9653	.8777 .8785
.25, .5	1.0 1.0	.9999 1.0	.9997 1.0	.9988 .9999	.9917 .9992
.25, .25	.9970 .9971	.9914 .9916	.9801 .9807	.9544 .9557	.8564 .8627
.1, .25	.9997 1.0	.9989 .9998	.9970 .9994	.9913 .9981	.9595 .9885
.1, .1	.9949 .9955	.9862 .9878	.9701 .9737	.9360 .9429	.8180 .8356
.01, .1	.9982 .9993	.9944 .9977	.9865 .9942	.9678 .9845	.8887 .9384
.01, .01	.9909 .9912	.9771 .9777	.9533 .9547	.9062 .9084	.7602 .7641
0, .01	.9916 .9923	.9787 .9802	.9563 .9593	.9117 .9160	.7690 .7799
0, 0	.9901 .9901	.9754 .9752	.9504 .9501	.9015 .9004	.7507 .7506

Asymptotic value of $P(CS|R_1)$ for three independent exponential populations with lcoation parameters $0 = \theta_{[1]} \leq \theta_{[2]} \leq \theta_{[3]}$ for d values chosen to satisfy $P(CS|R_1) \geq P^*$ when $\theta_i = 0$, $i = 1,2,3$. Upper value is for n=20; lower value for n=50.

normal variables and $\bar{U}$ is their mean. Let V_i, $i=1,\ldots,k$, be defined by (2.3.19) and

$$(2.4.1) \qquad q = q(c,n,k) = (c-1)[3(nk+1)/k]^{1/2} .$$

Then for $c \geq 1$,

$$(2.4.2) \qquad P[\max_{1\leq j\leq k} T_j/T_k \leq c] = P[V_i \leq cV_k+q,\ i=1,\ldots,k-1].$$

From Lemma 2.3.3,

$$(2.4.3) \qquad P[V_i \leq cV_k+q,\ i=1,\ldots,k-1] \approx P[W_i \leq cW_k+q,\ i=1,\ldots,k-1].$$

Define

$$(2.4.4) \qquad Q_i = (W_i-cW_k)\,[1+c^2-(c-1)^2/k]^{-1/2},\ i=1,\ldots,k-1.$$

Then the joint distribution of $(Q_1,\ldots,Q_{k-1})$ is the multivariate normal distribution with zero mean vector and variance-covariance matrix

$$(2.4.5) \qquad \Sigma_Q = ([k\delta_{ij}+c^2k-(c-1)^2]/[k+c^2k-(c-1)^2]).$$

Now

$$(2.4.6) \qquad P[\max_{1\leq j\leq k} T_j/T_k \leq c] \approx P[Q_1 \leq u,\ldots,Q_{k-1} \leq u],$$

where

(2.4.7) $$u = u(c,n,k) = (c-1)[3(kn+1)]^{1/2}[k(1+c^2)-(c-1)^2]^{-1/2}.$$

In Gupta [29], a suitable transformation is provided which allows the representation

(2.4.8) $$P[Q_1 \leq u,\ldots,Q_{k-1} \leq u] = \int_{-\infty}^{\infty} \Phi^{k-1}[(u+\rho^{1/2}x)(1-\rho)^{-1/2}]\varphi(x)dx,$$

where

(2.4.9) $$\rho = \rho(c,k) = [c^2k-(c-1)^2][k+c^2k-(c-1)^2]^{-1}.$$

These statements now lead to the following theorem.

Theorem 2.4.1. Let X_{ij}, $j=1,\ldots,n$, $i=1,\ldots,k$, be independent identically distributed random variables with a continuous distribution function. Then

(2.4.10) $$P[\max_{1 \leq j \leq k} T_j/T_k \leq c] \approx \int_{-\infty}^{\infty} \Phi^{k-1}[(u+\rho^{1/2}x)(1-\rho)^{-1/2}]\varphi(x)dx, \quad c \geq 1$$

where u and ρ are defined by (2.4.7) and (2.4.9) respectively.

For given values of ρ, k and u the integral given in the right hand side of (2.4.10) is evaluated in [29], and these values are displayed in Table II of that reference. Equation (2.4.9) can be solved for c in terms of ρ and k. Recalling $c \geq 1$, one obtains

(2.4.11) $$c = \{\rho-1+[k(1-\rho)(\rho k-2\rho+1)]^{1/2}\}[(k-1)(1-\rho)]^{-1}.$$

Using obvious bounds,

(2.4.12) $$P[\max_{1\le j\le k} T_j/T_k \le c] = 1 \quad \text{for} \quad c \ge (2kn-n+1)(n+1)^{-1}.$$

Remark: For fixed k, $\rho = \rho(c,k)$ is continuous and strictly increasing for $1 \le c < \infty$. Also $\rho(1,k) = 1/2$ and $\lim_{c\to\infty} \rho(c,k) = 1$. Hence $c=c(\rho,k)$ is continuous and strictly increasing for $1/2 \le \rho < 1$, and $c(1/2,k)=1$, $\lim_{\rho\to 1} c(\rho,k) = \infty$.

Table 4 gives asymptotic upper and lower bounds for $P[\max_{1\le j\le k} T_j/T_k \le c]$ for selected values of c, k=2(1)5, n=2(1)10. For a fixed k, the c values are obtained from (2.4.11) by letting ρ = .5,.6, .625,2/3,.7,.75,.8,.875,.9. These bounds can then be read from [29] Table II since u is now fixed. Some exact values are also included for comparison. For k=2, these exact values are obtained from a Mann-Whitney table; for k=3, n=2 exact results were obtained by enumeration. In most cases the exact values are within the asymptotic bounds. Using the above method it is possible to construct such tables for k=2(1)13,n $\ge$ 2 and for the c values obtained from (2.4.11) for the values of ρ indicated above.

Now consider the more general case where the $F_i(x)$, i=1,...,k may not all be identical. Following the proofs of Theorems 2.3.3 and 2.3.4, set

(2.4.13) $$\underline{Y} = B\underline{T},$$

where B is the (k-1)xk matrix defined by

TABLE 4

Asymptotic Lower and Upper Bounds on $P[\max_{1\leq j\leq k} T_j/T_1 \leq c]$ for

Indicated Values of k, n and c

k = 2

n \ c	1.0	1.23607	1.30940	1.44949	1.58199	1.82843	2.16228	3.0	3.47214
2	.50000 (.667)	.65542 (.667) .69146	(.667) .69146 .72575	(.667) .75804 .78814	.78814 .81594 (.833)	(.833) .86433 .88493	(.833) .91924 .93319	.97128 .97725 (1.0)	.98214 .98610 (1.0)
3	.50000 (.500)	(.650) .65542 .69146	(.650) .72575 .75804	.78814 (.800) .81594	(.800) .84134 .86433	(.900) .90320 .91924	.94520 (.950) .95543	.98610 .98928 (1.0)	.99379 .99534 (1.0)
4	.50000 (.557)	(.657) .69146 .72575	.72575 (.757) .75804	.81594 (.829) .84134	.86433 .88493 (.900)	.93319 (.943) .94520	(.971) .97128 .97725	.99379 .99534 (1.0)	.99744 .99813 (1.0)
5	.50000 (.500)	.72575 (.726) .75804	.75804 .78814 (.790)	.84134 (.845) .86433	.88493 (.889) .90320	.94520 (.952) .95543	.98214 (.984) .98610	.99744 .99813 (1.0)	.99903 .99931 (1.0)
6	.50000 (.531)	.72575 (.758) .75804	.78814 (.803) .81594	.86433 (.880) .88493	(.910) .91924 .93319	.96407 (.968) .97128	.98610 .98928 (.992)	.99903 .99931 (1.0)	.99966 .99977 (1.0)
7	.50000 (.500)	.75804 (.772) .78814	.78814 (.809) .81594	.88493 (.896) .90320	.93319 (.936) .94520	.97128 (.973) .97725	.99180 .99379 (.994)	.99952 .99966 (1.0)	.99977 1.0 (1.0)
8	.50000 (.520)	.75804 (.779) .78814	.81594 (.836) .84134	(.903) .90320 .91924	.94520 (.948) .95543	.97725 (.981) .98214	.99534 .99653 (.997)	.99977 1.0 (1.0)	
9	.50000 (.500)	.75804 (.782) .78814	(.830) .84134 .86433	.90320 (.919) .91924	(.953) .95543 .96407	.98610 (.988) .98928	.99653 .99744 (.998)		
10	.50000 (.515)	.78814 (.803) .81594	.84134 (.860) .86433	.91924 (.928) .93319	.95543 (.962) .96407	.98928 (.991) .99180	.99813 .99865 (.999)		

For given values of k,n,c, this table gives asymptotic upper and lower bounds for $P[\max_{1\leq j\leq k} T_j/T_k \leq c]$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k, which are independent identically distributed. The exact values of this probability, where known, are given in parentheses.

TABLE 4 (Continued)

k = 3

n \ c	1.0	1.23205	1.30278	1.43649	1.56155	1.79129	2.09808	2.85410	3.27492
2	.33333 (.42222)	.47732 (.48888) .52003	.54259 (.55555) .56680	.61520 (.62222) .65487	(.66666) .69785 .73339	(.73333) .77315 .80322	(.82222) .83634 .86051	(.93333) .93705 .94870	(.95555) .95984 .96786
3	.33333	.52003 .56248	.56680 .60822	.65487 .69288	.73339 .76667	.83074 .85564	.90124 .91798	.96682 .97368	.97997 .98440
4	.33333	.56248 .60419	.60822 .64842	.72890 .76266	.79749 .82572	.87794 .89769	.93249 .94492	.98388 .98756	.99080 .99303
5	.33333	.56248 .60419	.64842 .68699	.76266 .79394	.82572 .85130	.89769 .91500	.95547 .96432	.99279 .99459	.99611 .99714
6	.33333	.60419 .64470	.68699 .72359	.76266 .79394	.85130 .87422	.93001 .94289	.97168 .97772	.99598 .99704	.99849 .99892
7	.33333	.60419 .64470	.68699 .72359	.79394 .82263	.87422 .89455	.94289 .95382	.98264 .98660	.99784 .99844	.99923 .99946
8	.33333	.64470 .68360	.72359 .75792	.82263 .84863	.89455 .91238	.95382 .96300	.98660 .98975	.99888 .99921	.99963 1.0
9	.33333	.64470 .68360	.72359 .75792	.84863 .87196	.91238 .92785	.96300 .97063	.98975 .99223	.99944 .99961	
10	.33333	.68360 .72054	.75792 .78978	.87196 .89264	.92785 .94112	.97690 .98200	.99417 .99567	.99961 1.0	

For given values of k,n,c, this table gives asymptotic upper and lower bounds for $P[\max_{1 \leq j \leq k} T_j/T_k \leq c]$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k, which are independent identically distributed. The exact values of this probability, where known, are given in parentheses.

TABLE 4 (Continued)

$k = 4$

n \ c	1.0	1.23014	1.29966	1.43050	1.55228	1.77485	2.07037	2.79361	3.19434
2	.2500	.39946 .44266	.45021 .49381	.54944 .59190	.60133 .64206	.73009 .76387	.80534 .83279	.90949 .92511	.94139 .95241
3	.2500	.44266 .48650	.49381 .53744	.59190 .63328	.68121 .71842	.79516 .82383	.85761 .87981	.95007 .95978	.96946 .97586
4	.2500	.48650 .53046	.53744 .58058	.63328 .67313	.71842 .75336	.82383 .84981	.91661 .93147	.97461 .98010	.98533 .98873
5	.2500	.48650 .53046	.58058 .62273	.67313 .71106	.78582 .81563	.87309 .89372	.93147 .94420	.98812 .99095	.99352 .99516
6	.2500	.53046 .57400	.58058 .62273	.71106 .74676	.81563 .84269	.89372 .91180	.95498 .96402	.99317 .99489	.99737 .99809
7	.2500	.53046 .57400	.62273 .66343	.74676 .77996	.84269 .86698	.92748 .94092	.97151 .97765	.99622 .99722	.99863 .99902
8	.2500	.57400 .61660	.66343 .70225	.77996 .81049	.86698 .88854	.94092 .95231	.97765 .98263	.99798 .99855	.99931 .99952
9	.2500	.57400 .61660	.66343 .70225	.81049 .83824	.86698 .88854	.95231 .96187	.98663 .98981	.99897 .99927	.99952 1.0
10	.2500	.61660 .65780	.70225 .73886	.81049 .83824	.88854 .90746	.96187 .96980	.98981 .99231	.99949 1.0	

For given values of k,n,c, this table gives asymptotic upper and lower bounds for $P[\max_{1\le j\le k} T_j/T_k \le c]$. The rank sums T_i, $i=1,\ldots,k$, are based on random variables X_{ij}, $j=1,\ldots,n$, $i=1,\ldots,k$, which are independent identically distributed.

TABLE 4 (Continued)

k = 5

n \ c	1.0	1.22902	1.29785	1.42705	1.54699	1.76556	2.05489	2.76040	3.15037
2	.20000	.34859 .39105	.40059 .44419	.50403 .54770	.56006 .60244	.66074 .69892	.78264 .81229	.87975 .89918	.92078 .93483
3	.20000	.39105 .43480	.44419 .48843	.54770 .59076	.64361 .68315	.73503 .76877	.83930 .86362	.94366 .95442	.96563 .97273
4	.20000	.43480 .47931	.48843 .53276	.59076 .63272	.68315 .72067	.79995 .82845	.88529 .90438	.97098 .97716	.98331 .98712
5	.20000	.43480 .47931	.53276 .57665	.63272 .67311	.72067 .75586	.85419 .87719	.92102 .93536	.98220 .98625	.99015 .99254
6	.20000	.47931 .52403	.53276 .57665	.67311 .71154	.75586 .78849	.87719 .89751	.94759 .95790	.98948 .99203	.99585 .99694
7	.20000	.47931 .52403	.57665 .61956	.71154 .74766	.78849 .81839	.89751 .91525	.95790 .96650	.99402 .99555	.99777 .99839
8	.20000	.52403 .56840	.61956 .66100	.74766 .78122	.81839 .84547	.91525 .93058	.97359 .97939	.99673 .99761	.99885 .99919
9	.20000	.52403 .56840	.61956 .66100	.74766 .78122	.84547 .86972	.94367 .95472	.97939 .98406	.99828 .99877	.99943 1.0
10	.20000	.56840 .61187	.66100 .70053	.78122 .81202	.86972 .89119	.95472 .96396	.98780 .99074	.99913 .99939	

For given values of k,n,c, this table gives asymptotic upper and lower bounds for $P[\max_{1 \leq j \leq k} T_j/T_k \leq c]$. The rank sums T_i, i=1,...,k, are based on random variables X_{ij}, j=1,...,n, i=1,...,k, which are independent identically distributed.

$$(2.4.14)\qquad B = \begin{pmatrix} 1 & 0 & \cdots & 0 & -c \\ 0 & 1 & \cdots & 0 & -c \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c \end{pmatrix} .$$

For large n, $\underline{Y}' = (Y_1,\ldots,Y_{k-1})$ is approximately distributed as a multivariate normal random vector with mean vector $\mu_Y = B\mu_T$ and variance-covariance matrix $\Sigma_Y = B\Sigma_T B'$.

Theorem 2.4.2. If Σ_Y is nonsingular, then for $c \geq 1$

$$(2.4.15)\qquad P(T_k \geq c^{-1} \max_{1\leq j\leq k} T_j) \approx L\int_{-\infty}^{0}\cdots\int_{-\infty}^{0} \exp[-(\underline{y}-\mu_Y)'\Sigma_Y^{-1}(\underline{y}-\mu_Y)/2] \prod_{i=1}^{k-1} dy_i,$$

where

$$(2.4.16)\qquad L = [(2\pi)^{k-1} |\Sigma_Y|]^{-1/2} .$$

To obtain an expression for the expected subset size, define for $\nu = 1,\ldots,k$,

$$(2.4.17)\qquad \underline{Y}^{\nu} = B_{\nu}\, \underline{T},$$

where B_{ν} is the (k-1)xk matrix obtained from matrix B defined in (2.4.14) by moving column j to column j+1, $j=\nu, \nu+1,\ldots,k-1$, and replacing column ν by column k. The matrix B_k is matrix B. Let $\mu_{\nu} = B_{\nu}\mu_T$ and $\Sigma_{\nu} = B_{\nu}\Sigma_T B_{\nu}'$.

Theorem 2.4.3. If Σ_{ν} is nonsingular for $\nu=1,\ldots,k$, then

$$(2.4.18)\qquad E(S|R_2) \approx \sum_{\nu=1}^{k} L_\nu \int_{-\infty}^{0}\cdots\int_{-\infty}^{0} \exp[-(\underline{y}^\nu-\underline{\mu}_\nu)'\Sigma_\nu^{-1}(\underline{y}^\nu-\underline{\mu}_\nu)/2] \prod_{i=1}^{k} dy_i^\nu ,$$

where

$$(2.4.19)\qquad L_\nu = [(2\pi)^{k-1} |\Sigma_\nu|]^{-1/2}.$$

2.5 The Distribution of max Z_j/Z_i

A. Introduction

Suppose the populations $\pi_1,\ldots,\pi_k$ are independent and the random variables X_{ij}, $j=1,\ldots,n_i$, $i=1,\ldots,k$, are independent identically distributed with a continuous distribution function $F(x)$. Let

$$(2.5.1)\qquad Z_i = \prod_{j=1}^{n_i} R_{ij}, \quad i=1,\ldots,k.$$

Then

$$(2.5.2)\qquad \prod_{j=1}^{k} Z_j = \left(\sum_{j=1}^{k} n_j\right)!.$$

The objective of this section is to obtain the distribution of the statistic $\max_{1\le j\le k} Z_j/Z_1$ when $n_i=n$, $i=1,\ldots,k$. This statistic arises from consideration of the following ranking and selection procedure:

$$(2.5.3)\qquad R_Z\text{: Select } \pi_i \text{ iff } Z_i \geq c_Z^{-1} \max_{1\le j\le k} Z_j, \quad c_Z \geq 1.$$

If the quantity $\log(R_{ij})$ is associated with the random variable X_{ij} and $T_i' = \sum_{j=1}^{n} \log(R_{ij})$, then rule R_Z can be written in the form of rule R_1, i.e.,

$$(2.5.4) \qquad R_Z\text{: Select } \pi_i \text{ iff } T_i' \geq \max_{1\leq j\leq k} T_j' - d', \quad d' \geq 0.$$

It should be noted that a theorem similar to Theorem 1.2.1 can be established for rule R_Z. The recursion formulas given here are similar to the one developed in Section 2.3.B.

B. Case k = 2

Suppose k = 2, and define

$$(2.5.5) \qquad N(n_1,n_2|m) = \text{number of sequences in which } Z_2 \leq mZ_1.$$

Let $S = n_1+n_2$. Then by conditioning on the parent population of the largest observation, the following recursion formula is obtained:

$$(2.5.6) \qquad N(n_1,n_2|m) = N(n_1-1,n_2|Sm)+N(n_1,n_2-1|mS^{-1}),$$

with the following boundary conditions:

1) If $(n_1+n_2)!\ m < (n_2!)^2$, then $N(n_1,n_2|m) = 0$.

2) If $(n_1+n_2)! \leq m(n_1!)^2$, then $N(n_1,n_2|m) = \binom{S}{n_1}$.

Then for $m \geq 1$,

(2.5.7) $$P[\max_{1\leq j\leq k} Z_j/Z_1 \leq m] = P[Z_2 \leq m\, Z_1]$$

$$= N(n_1,n_2|m)\binom{S}{n_1}^{-1} .$$

From (2.5.2),

(2.5.8) $$P[Z_2 \leq m\, Z_1] = P[Z_2 \leq (m(n_1+n_2)!)^{1/2}].$$

C. Case $k \geq 3$

Suppose $k = 3$ and let $S = n_1+n_2+n_3$. Define

(2.5.9) $N(n_1,n_2,n_3|m_2,m_3)$ = number of sequences in which $Z_2 \leq m_2 Z_1$ and $Z_3 \leq m_3\, Z_1$.

Then

(2.5.10) $$N(n_1,n_2,n_3|m_2,m_3) = N(n_1-1,n_2,n_3|Sm_2,Sm_3)$$

$$+ N(n_1,n_2-1,n_3|m_2S^{-1},m_3)+N(n_1,n_2,n_3-1|m_2,m_3S^{-1})$$

with the boundary conditions:

1) If for any $i \geq 2$, $S!\, m_i < (n_i)!\, (n_2+n_3)!$, then $N(n_1,n_2,n_3|m_2,m_3)=0$

2) If for every $i \geq 2$, $S! \leq m_i(n_1)!\, (\sum_{j\neq i} n_j)!$, then

$$N(n_1,n_2,n_3|m_2,m_3) = \binom{S}{n_1,n_2,n_3} .$$

3) $N(0,n_2,n_3|m_2,m_3)$ = number of sequences such that $(n_2+n_3)!\, m_3^{-1} \leq Z_2 \leq m_2$, so:

(a) if $m_2 < (n_2+n_3)!\ m_3^{-1}$, then $N(0,n_2,n_3|m_2,m_3) = 0$

(b) If (a) does not hold then this term can be evaluated using (2.5.6).

4) $N(n_1,0,n_3|m_2,m_3)$ = number of sequences such that $Z_1 \geq \max\{m_2^{-1},[m_3^{-1}(n_1+n_3)!]^{1/2}\} \equiv M$, so

(a) if $M \leq n_1!$, then $N(n_1,0,n_3|m_2,m_3) = \binom{S}{n_1}$,

(b) if $(n_1+n_3)! < n_3!\ M$, then $N(n_1,0,n_3|m_2,m_3) = 0$.

(c) if (a) and (b) fail to hold then this term can be evaluated using (2.5.6).

5) $N(n_1,n_2,0|m_2,m_3)$ is handled with conditions similar to those in (4).

The following symmetry holds:

$$N(n_1,n_2,n_3|m_2,m_3) = N(n_1,n_3,n_2|m_3,m_2). \tag{2.5.11}$$

Thus, at an ''equal n_i, equal m_i stage,'' equation (2.5.10) can be written as

$$N(n,n,n|m,m) = N(n-1,n,n|Sm,Sm)+2N(n,n-1,n|mS^{-1},m). \tag{2.5.12}$$

In order to get $P[\max_{1\leq j\leq k} Z_j/Z_1 \leq m]$ for $m \geq 1$ and k=3, one uses

$$P[\max_{1\leq j\leq k} Z_j/Z_1 \leq m] = N(n_1,n_2,n_3|m,m)\binom{S}{n_1,n_2,n_3}^{-1}. \tag{2.5.13}$$

For a given $k \geq 4$ definitions and formulas similar to (2.5.9) and (2.5.10) can be written with boundary conditions involving values from the $k - 1$ case.

CHAPTER III

TWO PARAMETRIC RULES AND SOME COMPARISONS

3.1. Introduction

The goal of this chapter is to compare some performance characteristics of the distribution-free procedures R_1 and R_2 with those of possible competing procedures, as well as to compare rule R_1 with rule R_2. In Section 3.2, the asymptotic relative efficiency (ARE) of rules R_1 and R_2 relative to a normal means procedure is computed assuming the two given populations have a normal distribution with the same known variance. In Section 3.3 the ARE of rules R_1 and R_2 relative to Gupta's gamma procedure is computed assuming the two given populations have the gamma distribution differing only in the scale parameter. In Section 3.4 the k populations are assumed to have a gamma distribution differing only in the location (or guaranteed life time) parameter. Two parametric selection rules are then developed: one based on the sample means and one based on the minimum observation from each population. Section 3.5 contains some exact numerical comparisons between the distribution-free procedures R_1 and R_2 and those developed in the previous section. The last section of this chapter contains some discussion of the previous results, conjectures, further problems to be investigated, etc.

3.2. Asymptotic Relative Efficiency (ARE) of the Rules R_1 and R_2 Relative to a Normal Means Procedure

Suppose π_1 and π_2 are two independent normal populations with a common variance unity. Let the random variables associated with $\pi_{(1)}$ have mean 0 and those associated with $\pi_{(2)}$ have mean $\theta(\geq 0)$. A sample of size n is drawn from each of the two populations. The ARE of procedures R_1 and R (to be defined below) will be calculated. Based on X_{ij}, $j = 1,\ldots,n$, $i=1,2$, let T_i and $\bar{X}_i$ be the rank sum and sample mean, respectively, from π_i, $i=1,2$. Then the procedures are:

(3.2.1) $$R_1\text{: Select } \pi_i \text{ iff } T_i \geq \max_{j=1,2} T_j - d, \quad d \geq 0,$$

(3.2.2) $$R\text{: Select } \pi_i \text{ iff } \bar{X}_i \geq \max_{j=1,2} \bar{X}_j - b, \quad b \geq 0.$$

The constants d and b are chosen so that the probability of a correct selection is bounded below by a given number P*, $1/2 < P^* < 1$, for all θ, i.e.,

(3.2.3) $$\inf_{\theta \geq 0} P(\pi_{(2)} \text{ is selected}) \geq P^*.$$

Procedure R has been investigated by Gupta [30]. First consider rule R_1, and let S* denote, as before, the number of non-best populations in the selected subset. Here, of course, S* is either 0 or 1. For convenience replace d in (3.2.1) by nd. Then

$$(3.2.4)\qquad E(S^*|R_1) = P(T_{(1)} \geq T_{(2)} - nd)$$

$$= P\{\sigma^{-1}(T_{(1)}-\mu) \geq -\sigma^{-1}[\mu-n(2n+1-d)/2]\}$$

$$\simeq \Phi\{\sigma^{-1}[\mu-n(2n+1-d)/2]\},$$

where

$$(3.2.5)\qquad \mu = E(T_{(1)}) = n(3n+1)/2-n^2\ \Phi(\theta 2^{-1/2}),$$

$$(3.2.6)\qquad \sigma^2 = \mathrm{Var}(T_{(1)})$$

$$= n^2[\Phi(\theta 2^{-1/2})+2(n-1)\int_{-\infty}^{\infty}\Phi^2(x+\theta)\varphi(x)dx-(2n-1)\ \Phi^2(\theta 2^{-1/2})].$$

These moments are given by (2.2.5) and (2.2.13). Now set the right hand side of (3.2.4) equal to $\epsilon > 0$ and obtain

$$(3.2.7)\qquad \mu-n(2n+1-d)/2 = \sigma\ \Phi^{-1}(\epsilon).$$

From Theorem 1.2.1 the appropriate d value is obtained from (3.2.3) when $\theta = 0$. Equation (2.3.25) provides a large sample solution for d, namely,

$$(3.2.8)\qquad d \simeq h_1\ n^{1/2}\ ,$$

where h_1 is independent of n and θ. (Actually $h_1 = h(2/3)^{1/2}$, where h is the appropriate value obtained from Gupta [29]). Putting (3.2.5), (3.2.6) and (3.2.8) into (3.2.7) and simplifying yields

$$(3.2.9) \quad n+h_1 n^{1/2}-2n\,\Phi(\theta 2^{-1/2}) = 2\,\Phi^{-1}(\epsilon)[\Phi(\theta 2^{-1/2})+2(n-1)\int_{-\infty}^{\infty}\Phi^2(x+\theta)\varphi(x)dx-(2n-1)\Phi^2(\theta 2^{-1/2})]^{1/2},$$

or upon rearrangement,

$$(3.2.10) \quad n(1-2\Phi(\theta 2^{-1/2}))+h_1 n^{1/2}=2\Phi^{-1}(\epsilon)(2nB^2(\theta)+R(\theta))^{1/2},$$

where

$$(3.2.11) \quad B^2(\theta) = \int_{-\infty}^{\infty}\Phi^2(x+\theta)\,\varphi(x)dx-\Phi^2(\theta 2^{-1/2}),$$

$$(3.2.12) \quad R(\theta) = \Phi(\theta 2^{-1/2})-2\int_{-\infty}^{\infty}\Phi^2\,(x+\theta)\,\varphi(x)dx+\Phi^2(\theta 2^{-1/2}).$$

For large n, the $R(\theta)$ term in (3.2.10) can be ignored and then that equation simplifies to

$$(3.2.13) \quad n^{1/2} \approx [2^{3/2}\,\Phi^{-1}(\epsilon)\,B(\theta)-h_1][1-2\Phi(\theta 2^{-1/2})]^{-1}.$$

Thus,

$$(3.2.14) \quad n_{R_1}(\epsilon) \approx [2^{3/2}\,\Phi^{-1}(\epsilon)\,B(\theta)-h_1]^2\,[1-2\Phi(\theta 2^{-1/2})]^{-2}.$$

Now consider rule R.

$$(3.2.15)\qquad E(S^*|R) = P[\bar{X}_{(1)} \geq \bar{X}_{(2)} - b]$$

$$= \Phi[(b-\theta)(n/2)^{1/2}].$$

Again, b is obtained from (3.2.3) when $\theta = 0$.

$$(3.2.16)\qquad b = h_1(3/n)^{1/2}.$$

Setting the right hand side of (3.2.15) equal to ϵ and using (3.2.16) yields

$$(3.2.17)\qquad n_R(\epsilon) = [(3^{1/2}\, h_1 - 2^{1/2}\, \Phi^{-1}(\epsilon))/\theta]^2.$$

The asymptotic relative efficiency of R_1 relative to R is defined to be

$$(3.2.18)\qquad ARE(R_1,R;\theta) = \lim_{\epsilon \downarrow 0}[n_R(\epsilon)/n_{R_1}(\epsilon)].$$

From equations (3.2.14) and (3.2.17),

$$(3.2.19)\qquad ARE(R_1,R;\theta) = \{[2\Phi(\theta 2^{-1/2})-1]/2\theta B(\theta)\}^2.$$

If θ is allowed to decrease to 0, then

$$(3.2.20)\qquad \lim_{\theta \downarrow 0} ARE(R_1,R;\theta) = 3/\pi = .9549.$$

Recall that for k=2, rule R_2 has the form:

(3.2.21) $$R_2\text{: Select } \pi_i \text{ iff } T_i \geq c^{-1} \max_{j=1,2} T_j, \quad c \geq 1$$

In this case rules R_1 and R_2 are equivalent, so equations (3.2.19) and (3.2.20) remain valid if R_1 is replaced by R_2.

3.3 Asymptotic Relative Efficiency of the Rules R_1 and R_2 Relative to Gupta's Gamma Procedure

Let π_1, π_2 be two independent exponential populations with associated independent random variables X_{ij}, $j=1,\ldots,n$, $i=1,2$. The density function of X_{ij} is

(3.3.1) $$f_i(x) = \begin{cases} \theta_i^{-1} e^{-x/\theta_i} & , \quad x > 0 \quad , \quad i = 1,2, \\ 0 & , \quad x \leq 0 \end{cases}$$

where $1 = \theta_{[1]} \leq \theta_{[2]} = \theta$. The ARE of procedures R_2 (and hence R_1) and R' (to be defined below) will be calculated. Procedure R_2 is given by (3.2.21) and R' by

(3.3.2) $$R'\text{: Select } \pi_i \text{ iff } \bar{X}_i \geq b^{-1} \max_{j=1,2} \bar{X}_j, \quad b \geq 1.$$

The constants c and b are chosen so that

(3.3.3) $$\inf_{\theta \geq 1} P(\pi_{(2)} \text{ is selected}) \geq P^*.$$

Procedure R' has been studied by Gupta [28]. Now first consider rule R_2. For convenience replace c by c^{-1} in (3.2.21). Then

$$(3.3.4)\qquad E(S^*|R_2) = P(T_{(1)} \geq cT_{(2)})$$

$$\approx \Phi\{\sigma^{-1}[\mu-cn(2n+1)(c+1)^{-1}]\},$$

where

$$(3.3.5)\qquad \mu = E(T_{(1)}) = 2^{-1}n(3n+1)-n^2\,\theta(1+\theta)^{-1},$$

$$(3.3.6)\quad \sigma^2 = \mathrm{Var}(T_{(1)})$$

$$= n^2\{\theta(1+\theta)^{-1}+(n-1)[1-2(1+\theta)^{-1}+(2\theta+1)^{-1}+\theta(2+\theta)^{-1}]$$
$$-(2n-1)\,\theta^2(1+\theta)^{-2}\}.$$

Now set the right hand side of (3.3.4) equal to $\epsilon > 0$ and obtain

$$(3.3.7)\qquad \mu-cn(2n+1)(c+1)^{-1} = \sigma\,\Phi^{-1}(\epsilon).$$

From Theorem 1.2.1, c is obtained from (3.3.3) when $\theta = 1$. Through Theorem 2.4.1 a large sample solution for c can be obtained as

$$(3.3.8)\qquad c \approx [(2n+1)^{1/2}-q][(2n+1)^{1/2}+q]^{-1},$$

where q is independent of n, θ and ϵ. (Actually $q = 3^{-1/2}\,\Phi^{-1}(P^*)$). Putting (3.3.5), (3.3.6) and (3.3.8) into (3.3.7) and simplifying yields

(3.3.9) $$2^{-1}q(2n+1)^{1/2}+n[2^{-1}-\theta(1+\theta)^{-1}]=\Phi^{-1}(\epsilon)[nB^2(\theta)+R(\theta)]^{1/2},$$

where

(3.3.10) $$B^2(\theta) = 1-2(1+\theta)^{-1}+(2\theta+1)^{-1}+\theta(2+\theta)^{-1}-2\theta^2(1+\theta)^{-2},$$

(3.3.11) $$R(\theta)=\theta^2(1+\theta)^{-2}+\theta(1+\theta)^{-1}-\theta(2+\theta)^{-1}-(2\theta+1)^{-1}+2(1+\theta)^{-1}-1.$$

For large n, the $R(\theta)$ term in (3.3.9) can be ignored. That equation then simplifies to:

(3.3.12) $$n_{R_2}(\epsilon) \approx 4(\theta+1)^2(\theta-1)^{-2}[q2^{-1/2}-\Phi^{-1}(\epsilon)\ B(\theta)]^2.$$

Now consider rule R'. From Barlow and Gupta [3] (see their equation (4.21) with $\delta = \theta^{-1}$),

(3.3.13) $$n_{R'}(\epsilon) = 2(\log\theta)^{-2}\ [\Phi^{-1}(\epsilon)-3^{1/2}q]^2.$$

Using (3.3.12) and (3.3.13),

(3.3.14) $$\begin{aligned} ARE(R_2,R';\theta) &= \lim_{\epsilon\downarrow 0}[n_{R'}(\epsilon)/n_{R_2}(\epsilon)] \\ &= [(\theta-1)/4(\theta+1)\ B(\theta)\log\theta]^2 . \end{aligned}$$

Letting θ decrease to 1 yields

(3.3.15) $$\lim_{\theta\downarrow 1} ARE(R_2,R';\theta) = 3/4 .$$

Again equations (3.3.14) and (3.3.15) remain valid if R_2 is replaced by R_1.

3.4. Two Parametric Procedures for the Location Parameter of Gamma Distributions

A. A Procedure Based on Sample Means

Let $\pi_1,\dots,\pi_k$ be k independent populations whose associated random variables X_{ij}, $j=1,\dots,n$, $i=1,\dots,k$, are independent and have a gamma distribution with density given by

$$(3.4.1)\quad f(x-\theta_i) = \begin{cases} [\lambda/\Gamma(r)][(x-\theta_i)\lambda]^{r-1} e^{-\lambda(x-\theta_i)} & , \ x \geq \theta_i \\ \\ 0 & , \ x < \theta_i , \end{cases}$$

with common parameters $r(>0)$ and $\lambda(>0)$ which are assumed known. In life time studies and reliability work the parameter θ is often called the ''guaranteed life time''. Without loss of generality λ may be assumed equal to unity.

The ''best'' population is that one (or tagged one) with the longest guaranteed life, i.e., $\pi_{(k)}$. It is desired to choose a subset of the given k populations which will contain the best population with a given probability. A selection procedure based on the sample means can be stated as

$$(3.4.2)\qquad R_M\text{: Select } \pi_i \text{ iff } \bar{X}_i \geq \max_{1\leq j\leq k} \bar{X}_j - d, \quad d \geq 0.$$

The nonnegative constant d is chosen so that

$$(3.4.3) \qquad \inf_{\Omega} P(CS|R_M) = P^*,$$

where P^* is a given constant $(k^{-1} < P^* < 1)$ and

$$(3.4.4) \qquad \Omega = \{\underline{\theta} = (\theta_1,\ldots,\theta_k):\ \theta_i \geq 0,\ i = 1,\ldots,k\}.$$

The distribution of $\bar{X}_i$ is given by

$$(3.4.5)\ P(\bar{X}_i \leq x) = \begin{cases} [n/\Gamma(nr)]\int_{\theta_i}^{x} [n(t-\theta_i)]^{nr-1} e^{-n(t-\theta_i)} dt, & x \geq \theta_i \\ 0, & x < \theta_i \end{cases}$$

$$\equiv G_{\theta_i}(x) .$$

Let $G_o(x) = G(x)$, and $\bar{X}_{(i)}$ denote the (unknown) sample mean which comes from $\pi_{(i)}$, that population with parameter $\theta_{[i]}$. Then for $i=1,\ldots,k$,

$$(3.4.6) \qquad P(\pi_{(i)} \text{ is selected}) = P(\bar{X}_{(i)} \geq \max_{1\leq j\leq k} \bar{X}_{(j)} - d)$$

$$= \int_a^{\infty} \prod_{\substack{j=1 \\ j\neq i}}^{k} G(x+d+\theta_{[i]}-\theta_{[j]})\ dG(x),$$

where

$$a = \max(0, \max_{j \neq i} (\theta_{[j]} - \theta_{[i]} - d)). \tag{3.4.7}$$

In particular,

$$P(CS|R_M) = \int_0^\infty \prod_{j=1}^{k-1} G(x+d+\theta_{[k]} - \theta_{[j]}) \, dG(x). \tag{3.4.8}$$

Thus,

$$\inf_\Omega P(CS|R_M) = \int_0^\infty [G(x+d)]^{k-1} \, dG(x), \tag{3.4.9}$$

and this equation is used to determine d.

Consider the case where nr is a positive integer. Then

$$G(x) = 1 - \sum_{j=0}^{nr-1} e^{-nx} [(nx)^j/j!]. \tag{3.4.10}$$

Thus

$$G(x+d) = \sum_{j=0}^{nr} a_j, \tag{3.4.11}$$

where

$$a_j = \begin{cases} -e^{-n(x+d)}[n^j(x+d)^j/j!] , & j=0,1,\ldots,nr-1 \\ 1 , & j=nr \end{cases} \tag{3.4.12}$$

$$[G(x+d)]^{k-1} = (\sum_{j=0}^{nr} a_j)^{k-1} \tag{3.4.13}$$

$$= \sum_{\substack{k_o,\ldots,k_{nr} \\ \Sigma k_i = k-1}} A e^{-nx(k-1-k_{nr})} [n(x+d)]^{\ell},$$

where

(3.4.14) $$A = \binom{k-1}{k_o,\ldots,k_{nr}} (-e^{-nd})^{k-1-k_{nr}} \prod_{j=0}^{nr-1} (j!)^{-k_j},$$

(3.4.15) $$\ell = \sum_{j=0}^{nr-1} jk_j .$$

From (3.4.9),

(3.4.16) $$\inf_{\Omega} P(CS|R_M) = [n/\Gamma(nr)] \sum_{\substack{k_o,\ldots,k_{nr} \\ \Sigma k_i = k-1}} A \int_o^{\infty} e^{-nx(k-k_{nr})} [n(x+d)]^{\ell} (nx)^{nr-1} dx.$$

Expanding $[n(x+d)]^{\ell}$ and performing the integration yields

Lemma 3.4.1. If nr is a positive integer, then

(3.4.17) $$\inf_{\Omega} P(CS|R_M) = [1/\Gamma(nr)] \sum_{\substack{k_o,\ldots,k_{nr} \\ \Sigma k_i = k-1}} A \sum_{j=o}^{\ell} \binom{\ell}{j} (nd)^{\ell-j} (k-k_{nr})^{-(j+nr)} \Gamma(j+nr),$$

where A and ℓ are defined by (3.4.14) and (3.4.15).

As a special case of (3.4.17), note that if $nr=1$ then $\ell=0$ and (3.4.17) simplifies to

(3.4.18) $$\inf_{\Omega} P(CS|R_M) = k^{-1} \sum_{j=0}^{k-1} \binom{k}{j+1}(-e^{-nd})^j.$$

B. A Procedure Based on Minimum Observations

Let $X_{i[1]} = \min_{1 \le j \le n} X_{ij}$, $i=1,\ldots,k$, and define the selection rule R_m to be

(3.4.19) $$R_m\text{: Select } \pi_i \text{ iff } X_{i[1]} \ge \max_{1 \le j \le k} X_{j[1]} - b, \quad b \ge 0,$$

where b is chosen to satisfy the basic P* requirement. Barr and Rizvi [5] have studied procedures of this form. Let

(3.4.20) $$F(x-\theta_i) = \begin{cases} \int_{\theta_i}^{x} f(t-\theta_i)dt, & x \ge \theta_i \\ 0, & x < \theta_i \end{cases}, \quad i = 1,\ldots,k.$$

The distribution of $X_{i[1]}$ is given by

(3.4.21) $$P(X_{i[1]} \le x) = \begin{cases} 1-[1-F(x-\theta_i)]^n, & x \ge \theta_i \\ 0, & x < \theta_i \end{cases}$$

$$\equiv H_{\theta_i}(x).$$

Let $H_o(x)=H(x)$, and $X_{(i)[1]}$ denote the (unknown) minimum observation which is taken from population $\pi_{(i)}$. Then for $i=1,\ldots,k$,

$$(3.4.22)\quad P(\pi_{(i)} \text{ is selected}) = P(X_{(i)[1]} \geq \max_{1\leq j\leq k} X_{(j)[1]} - b)$$

$$= \int_a^\infty \prod_{\substack{j=1\\ j\neq i}}^{k} H(x+b+\theta_{[i]}-\theta_{[j]})\, dH(x),$$

where

$$(3.4.23)\qquad a = \max(0, \max_{j\neq i}(\theta_{[j]}-\theta_{[i]}-b)).$$

In particular,

$$(3.4.24)\qquad P(CS|R_m) = \int_0^\infty \prod_{j=1}^{k-1} H(x+b+\theta_{[k]}-\theta_{[j]})\, dH(x).$$

Thus

$$(3.4.25)\qquad \inf_\Omega P(CS|R_m) = \int_0^\infty [H(x+b)]^{k-1}\, dH(x),$$

and this equation is used to determine b.

Consider the case where r is a positive integer. The next lemma is obtained in the same way Lemma 3.4.1 was obtained.

Lemma 3.4.2. If r is a positive integer, then

$$(3.4.26)\qquad \inf_\Omega P(CS|R_m) = [k\Gamma(r)]^{-1} \sum_{j=0}^{k-1} \binom{k}{j}(-e^{-nb})^{k-1-j} \sum_{\substack{a_0,\ldots,a_{r-1}\\ \Sigma a_i=n(k-1-j)}} B\ .$$

$$\cdot \sum_{\substack{c_o,\dots,c_{r-1}\\ \Sigma c_i = n-1}} D \cdot \sum_{\ell=0}^{\mu} \binom{\mu}{\ell} b^{\mu-\ell} [n(k-j)]^{1-\ell-v-r}\, \Gamma(\ell+v+r),$$

where

(3.4.27) $$B = \binom{n(k-1-j)}{a_o,\dots,a_{r-1}} \prod_{i=0}^{r-1} (i!)^{-a_i},$$

(3.4.28) $$D = \binom{n-1}{c_o,\dots,c_{r-1}} \prod_{i=0}^{r-1} (i!)^{-c_i},$$

(3.4.29) $$\mu = \sum_{i=0}^{r-1} i\, a_i,$$

(3.4.30) $$v = \sum_{i=0}^{r-1} i c_i \;.$$

As a special case of (3.4.26), note that if $r=1$ then

(3.4.31) $$\inf_{\Omega} P(CS|R_m) = k^{-1} \sum_{j=0}^{k-1} \binom{k}{j} (-e^{-nb})^{k-1-j}.$$

3.5. Some Exact Comparisons

Sections 3.2 and 3.3 compared the distribution-free rules R_1 and R_2 with certain parametric procedures assuming a particular underlying distribution of the random variables X_{ij}. The comparison was in terms of relative asymptotic efficiency, i.e., a large sample comparison. This section contains some exact numerical comparisons between the four selection rules R_1, R_2, R_M and R_m. These rules are defined by (1.2.4), (1.2.5), (3.4.2) and (3.4.19), respectively. Suppose π_1, π_2, π_3 are three independent populations and X_{ij}, j=1,2, i=1,2,3, are the associated independent random variables. Let X_{ij}, j=1,2, have the distribution

$$(3.5.1) \qquad F(x-\theta_i) = \begin{cases} 1-e^{-(x-\theta_i)} & , \ x \geq \theta_i \\ 0 & , \ x < \theta_i , \end{cases}$$

where $\Theta = [0,\infty)$, $\theta_{[1]} = 0 \leq \theta_{[2]} = \theta_{[3]} = \theta$. For each of the four rules under consideration the probability of a correct selection is minimized when $\theta = 0$, and it is this configuration which is used to determine the appropriate constants needed to implement these rules for a given P*. We consider P* = .6 and P* = 14/15. For these choices the infimum of the probability of a correct selection is equal to P* for all four rules, and the appropriate constants for the rules are:

P*	R_1	R_2	R_M	R_m
.6	2	4/3	0.5928	0.3719
14/15	6	11/4	1.8668	1.3425

The expected size of the selected subset for each of the four rules and for both values of P* is computed for θ = 0(.1)1.5. For rules R_M and R_m this quantity can be obtained using (3.4.6) and (3.4.22) respectively; for the procedures R_1 and R_2, an enumeration method yields

$$(3.5.2) \qquad E(S|R_1) = \begin{cases} (25 + 4e^{-\theta}-2e^{-2\theta})/15, & P^* = .6 \\ \\ (35 + 14e^{-\theta}-7e^{-2\theta})/15, & P^* = 14/15, \end{cases}$$

$$(3.5.3) \qquad E(S|R_2) = \begin{cases} (25 + 4e^{-\theta}-2e^{-2\theta})/15, & P^* = .6 \\ \\ (10 + 8e^{-\theta}-4e^{-2\theta})/5, & P^* = 14/15. \end{cases}$$

These computations are found in Table 5 and are now summarized:

1) Between the two parametric procedures, the rule based on the minimum observations, R_m, is better in the sense that $E(S|R_m) \leq E(S|R_M)$, $P^* = .6$, $14/15$, $\theta = 0(.1)1.5$. The inequality is strict except when $\theta = 0$.

2) The two distribution-free procedures, R_1 and R_2, perform equally well for $P^* = .6$, i.e., $E(S|R_1) = E(S|R_2)$ for all values of θ. For $P^* = 14/15$, rule R_2 is better since $E(S|R_2) \leq E(S|R_1)$ for all θ, equality holding only when $\theta = 0$.

3) For $P^* = .6$, both parametric rules perform better than the distribution-free procedures. However, for $P^* = 14/15$, there are values of θ for which $E(S|R_1) < E(S|R_m)$, i.e., both distribution-free procedures perform better than the rule R_m (and hence R_M). There are also points at which R_m performs better than R_2 (and hence R_1); R_2 performs better than R_M; and R_M performs better than R_1.

3.6. Discussion and Related Problems

The two distribution-free selection rules, R_1 and R_2, are based upon statistics which can be looked upon as generalizations of the

TABLE 5

$E(S)$ for Rules R_1, R_2, R_M and R_m

θ	$E(S\|R_1)$	$E(S\|R_2)$	$E(S\|R_M)$	$E(S\|R_m)$
0	1.80000 2.80000	1.80000 2.80000	1.80000 2.80000	1.80000 2.80000
.1	1.79879 2.79577	1.79879 2.79276	1.79663 2.79837	1.79066 2.79746
.2	1.79562 2.78467	1.79562 2.77371	1.78726 2.79347	1.76557 2.78982
.3	1.79104 2.76865	1.79104 2.74626	1.77311 2.78529	1.73046 2.77690
.4	1.78551 2.74928	1.78551 2.71305	1.75555 2.77379	1.69392 2.75839
.5	1.77936 2.72775	1.77936 2.67615	1.73595 2.75887	1.66324 2.73385
.6	1.77286 2.70500	1.77286 2.63714	1.71550 2.74047	1.63813 2.70278
.7	1.76621 2.68173	1.76621 2.59726	1.69516 2.71849	1.61757 2.66460
.8	1.75957 2.65849	1.75957 2.55741	1.67554 2.69287	1.60073 2.61881
.9	1.75305 2.63566	1.75305 2.51827	1.65706 2.66360	1.58695 2.56510
1.0	1.74672 2.61353	1.74672 2.48034	1.63995 2.63075	1.57566 2.50360
1.1	1.74066 2.59231	1.74066 2.44395	1.62432 2.59447	1.56642 2.43525
1.2	1.73489 2.57211	1.73489 2.40934	1.61019 2.55507	1.55886 2.36239
1.3	1.72944 2.55304	1.72944 2.37663	1.59754 2.51300	1.55267 2.28967
1.4	1.72432 2.53511	1.72432 2.34591	1.58629 2.46890	1.54760 2.22496
1.5	1.71953 2.51835	1.71953 2.31718	1.57635 2.42356	1.54345 2.17182

$E(S)$ for rules R_1, R_2, R_M and R_m given n=2 and k=3 independent exponential populations with location parameters $0 = \theta_{[1]} \leq \theta_{[2]} = \theta_{[3]} = \theta$. Upper value is for $P^* = .6$, lower for $P^* = 14/15$.

Mann-Whitney U-statistic (or Wilcoxon rank sum statistic) to k populations with continuous random variables. Both rules have been suggested for selecting a subset of the given populations which will contain the stochastically largest one with a specified probability. It is assumed that each population is characterized by a single parameter which stochastically orders the k populations. There is, as yet, no fixed rule determining which of these two rules should be used in a given situation. One criterion could be simply an analogy with established means procedures, i.e., use rule R_1 when the distribution functions differ by a location parameter and use rule R_2 when the difference is in a scale parameter. However, if one desires to minimize the expected size of the selected subset, then this analogy leads to an incorrect choice of rules as seen in Section 3.5. Other reasons can be given for preferring rule R_1 to rule R_2. For independent identically distributed populations, the distribution of the statistic $\max_{1\leq j\leq k} T_j - T_k$ is better tabulated than that of $\max_{1\leq j\leq k} T_j/T_k$. Also rule R_1 is symmetric in the sense that the corresponding rule for selecting a subset containing the stochastically smallest population (see (1.1.6)) yields a statistic, $T_k - \min_{1\leq j\leq k} T_j$, which is stochastically equivalent to the statistic $\max_{1\leq j\leq k} T_j - T_k$ when the populations are identically distributed and equal samples are taken from each. The same is not true for rule R_2. There are many open problems here. For which distributions is the rule $R_1(R_2)$ the best rule to use from the class $R_1(G)$ $(R_2(G))$ in the sense of minimizing expected subset size? For which distributions does rule R_1 perform better than rule R_2 and vice versa? Certainly Monte Carlo techniques could contribute a great deal in these areas.

Two problems arise from consideration of the probability of a correct selection. The first is to characterize that class of distribution functions for which the infimum of the probability of a correct selection, using rules R_1 and R_2, takes place when the populations are identically distributed. This has been verified in some specific examples, and from Theorem 1.2.1 the result is true for two populations. However, it is not true in general as is demonstrated by Rizvi and Woodworth [57]. The same problem can also be posed for the more general class of rules, $R_1(G)$ and $R_2(G)$. The major difficulty in this respect is that the statistics involved (rank sums) are dependent from population to population, a difficulty encountered by Gupta and Nagel [33] in working with multinomial cell frequencies. Secondly, it would be desirable to characterize the class of distributions which yield the probability of a correct selection using rule R_1 (or R_2) as a nondecreasing function in the quantities $\theta_{[k]} - \theta_{[i]}$, $i=1,\ldots,k-1$. In the first chapter an example is given where this is not the case. It should be noted this example involved a distribution whose density lacked the MLR property and did not have the real line for its support.

Tabulating the distribution of the statistic $\max_{1 \leq j \leq k} T_j - T_k$ as was done in Section 2.3 requires a large amount of computer time for relatively small values of k and n. For two and three populations, the asymptotic results obtained from Theorem 2.3.2 (and Theorem 2.4.1) are in very close agreement with exact results for small sample sizes. For a larger number of populations it would be desirable to know how large a sample is required to obtain a ''reasonable'' approximation using these methods.

Finally, this work can be related to the work of Rizvi and Sobel [56] and Barlow and Gupta [3]. These two papers are concerned with selecting a subset of populations to contain the population with the largest (smallest) quantile of a given order $\alpha (0 < \alpha < 1)$ provided there exists a stochastically largest population. The rules R_1 and R_2 can be looked upon as selection procedures for the population possessing the largest quantiles of every order. In other words, no fixed quantile need be specified, since the distributions are assumed to be stochastically ordered.

BIBLIOGRAPHY

1. Alam, K. and Rizvi, M.H. (1966). Selection from multivariate normal populations, Ann. Inst. Statist. Math. Tokyo (18), 307-318.

2. Bahadur, R.R. (1950). On the problem in the theory of k populations, Ann. Math. Statist. (21), 362-375.

3. Barlow, R.E. and Gupta, S.S. (1967). Selection procedures for restricted families of probability distributions. Department of Statistics, Purdue University, mimeograph series No. 132.

4. Barlow, R.E., Gupta, S.S., and Panchapakesan, S. (1967). On the distribution of the maximum and minimum of ratios of order statistics. Department of Statistics, Purdue University, mimeograph series No. 133.

5. Barr, D.R. and Rizvi, M.H. (1966). Ranking and selection problems of uniform distributions, Trabajos de Estadistica (17), 15-31.

6. Barron, A.M. (1968). A class of sequential multiple decision procedures. Department of Statistics, Purdue University, mimeograph series No. 169.

7. Bartlett, N.S. and Govindarajulu, Z. (1965). Some distribution-free statistics and their application to the selection problem. Unpublished paper. Abstract in Ann. Math. Statist. (36), 1597-1598.

8. Bechhofer, R.E. (1954). A single sample multiple decision procedure for ranking means of normal populations with known variances, Ann. Math. Statist. (25), 16-39.

9. Bechhofer, R.E. and Sobel, M. (1954). A single sample decision procedure for ranking variances of normal populations, Ann. Math. Statist. (25), 273-289.

10. Bechhofer, R.E., Dunnett, C.W. and Sobel, M. (1954). A two sample multiple decision procedure for ranking means of normal populations with a common unknown variance, Biometrika (41), 170-176.

11. Bechhofer, R.E. (1958). A sequential multiple decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. Biometrics (14), 408-429.

12. Bechhofer, R.E. and Blumenthal, S. (1962). A sequential multiple-decision procedure selecting the best one of several normal populations with a common unknown variance II: Monte Carlo sampling results and new computing formulae, Biometrics (18), 52-67.

13. Bechhofer, R.E., Kiefer, J. and Sobel, M. Sequential identification and ranking procedures, University of Chicago Press, 1967.

14. Bell, C.B. and Doksum, K.A. (1965). Some new distribution-free statistics, Ann. Math. Statist. (36), 203-214.

15. Chambers, M.L. and Jarratt, P. (1964). Use of double sampling for selecting the best population, Biometrika (51), 49-64.

16. Deely, J.J. (1965). Multiple decision procedures from an empirical Bayes approach. Department of Statistics, Purdue University, mimeograph series No. 45.

17. Deely, J.J. and Gupta, S.S. (1968). On the properties of subset selection procedures, Sankhya Series A (30), 1-14.

18. Dunn, O.J. (1964). Multiple comparisons using rank sums, Technometrics (6), 241-252.

19. Dunnett, C.W. (1960). On selecting the largest of k normal population means, J. Roy. Stat. Soc. Series B (22), 1-40.

20. Dwass, M. (1957). On the distribution of ranks and of certain rank order statistics, Ann. Math. Statist. (28), 424-431.

21. Eaton, M.L. (1967). Some optimum properties of ranking procedures, Ann. Math. Statist. (38), 124-137.

22. Gnanadesikan, M. (1966). Some selection and ranking procedures for multivariate normal populations, Ph.D. thesis, Department of Statistics, Purdue University.

23. Gnanadesikan, M. and Gupta, S.S. (1968). Selection procedures for multivariate normal distributions in terms of measures of dispersion. Paper submitted for publication.

24. Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Institute of Statistics, University of North Carolina, Chapel Hill, N.C., mimeograph series No. 150.

25. Gupta, S.S. and Sobel, M. (1958). On selecting a subset which contains all populations better than a standard, Ann. Math. Statist. (29), 235-244.

26. Gupta, S.S. and Sobel, M. (1960). Selecting a subset containing the best of several binomial populations. Contributions to Probability and Statistics, Ch. XX, Stanford University Press.

27. Gupta, S.S. and Sobel, M. (1962). On selecting a subset containing the population with the smallest variance, Biometrika (49), 495-507.

28. Gupta, S.S. (1963). On a selection and ranking procedure for gamma populations, Ann. Inst. Statist. Math. Tokyo (14), 199-216.

29. Gupta, S.S. (1963). Probability integrals of multivariate normal and multivariate t, Ann. of Math. Statist. (34), 792-828.

30. Gupta, S.S. (1965). On some multiple decision rules, Technometrics (7), 225-245.

31. Gupta, S.S. (1965). Selection and ranking procedures and order statistics for the binomial distributions. Classical and Contagious Discrete Distributions, Statistical Publishing Society, Ch. 6, 1965.

32. Gupta, S.S. (1966). On some selection and ranking procedures for multivariate normal populations using distance functions. Multivariate Analysis, Academic Press, 1966.

33. Gupta, S.S. and Nagel, K. (1967). On selection and ranking procedures and order statistics from the multinomial distribution, Sankhya Series B (29), 1-34.

34. Gupta, S.S. and Panchapakesan, S. (1968). Some selection and ranking procedures for multivariate normal populations. To appear in the Proceedings of the Second Symposium on Multivariate Analysis, Academic Press.

35. Gupta, S.S. and Studden, W.J. (1968). On some selection and ranking procedures with applications to multivariate populations. To appear in the S.N. Roy Memorial Volume, University of North Carolina Press.

36. Guttman, I. and Tiao, G.C. (1964). A Bayesian approach to some best population problems, Ann. Math. Statist. (35), 825-835.

37. Hajek, J. (1967). Asymptotic normality of linear rank statistics under alternatives, Proc. Nat. Acad. Sci. U.S.A. (57), 19-20.

38. Lehmann, E.L. Testing Statistical Hypotheses, John Wiley, 1959.

39. Lehmann, E.L. (1961). Some model I problems of selection, Ann. Math. Statist. (32), 990-1012.

40. Lehmann, E.L. (1963). A class of selection procedures based on ranks, Math. Annalem (150), 268-275.

41. Mahamunulu, D.M. (1967). Some fixed-sample ranking and selection problems, Ann. Math. Statist. (38), 1079-1091.

42. Mann, H.B. and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Statist. (18), 50-60.

43. Milton, R.C. (1964). An extended table of critical values for the Mann-Whitney (Wilcoxon) two-sample statistic, J. Amer. Stat. Assn. (59), 925-934.

44. National Bureau of Standards. Tables of normal probability functions, Applied Mathematics Series 23, U.S. Government Printing Office, 1953.

45. National Bureau of Standards. Tables of the bivariate normal distribution and related functions, Applied Mathematics Series 50. U.S. Government Printing Office, 1959.

46. Odeh, R.E. (1967). The distribution of the maximum sum of ranks, Technometrics (9), 271-278.

47. Owen, D.B. Handbook of Statistical Tables, Addison-Wesley Publishing Company, Inc., 1962.

48. Patterson, D.W. (1966). Contributions to the theory of ranking and selection problems, Ph.D. Thesis, State University of Rutgers, New Brunswick, New Jersey.

49. Paulson, E. (1962). A sequential procedure for comparing several experimental categories with a standard or control, Ann. Math. Statist. (33), 438-443.

50. Paulson, E. (1963). A sequential decision procedure for choosing one of k hypothesis concerning the unknown mean of a normal distribution, Ann. Math. Statist. (34), 549-554.

51. Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations, Ann. Math. Statist. (35), 174-180.

52. Paulson, E. (1967). Sequential procedures for selection of the best one of several binomial populations, Ann. Math. Statist. (38), 117-123.

53. Puri, M.L. (1964). Asymptotic efficiency of a class of c-sample tests, Ann. Math. Statist. (35), 102-121.

54. Puri, P.S. and Puri, M.L. (1967). Selection procedures based on ranks: scale parameter case. Department of Statistics, Purdue University, mimeograph series No. 105.

55. Rizvi, M.H. (1963). Ranking and selection problems of normal populations using the absolute values of their means; fixed sample size case. Department of Statistics, University of Minnesota, technical report No. 31.

56. Rizvi, M.H. and Sobel, M. (1967). Nonparametric procedures for selecting a subset containing the population with the largest α-quantile, Ann. Math. Statist. (38), 1788-1803.

57. Rizvi, M.H. and Woodworth, G.G. (1968). On selection procedures based on ranks: counterexamples concerning the least favorable configuration. Department of Operations Research and Department of Statistics, Stanford University, technical report No. 114.

58. Rubin, H. (1961). Limit theorems for transformed variables. Department of Statistics, Michigan State University, technical report No. 85.

59. Seal, K.C. (1955). On a class of decision procedures for ranking means of normal populations, Ann. Math. Statist. (26), 387-398.

60. Seal, K.C. (1958). On ranking parameters of scale in type III populations, J. Amer. Stat. Assn. (53), 164-175.

61. Sobel, M. and Huyett, M.M. (1957). Selecting the best one of several binomial populations, Bell System Tech. J. (36), 537-576.

62. Studden, W.J. (1967). On selecting a subset of k populations containing the best, Ann. Math. Statist. (38), 1072-1078.

63. Whitney, D.R. (1951). A bivariate extension of the U statistic, Ann. Math. Statist. (22), 274-282.

64. Wilks, S.S. Mathematical Statistics, John Wiley, 1962.