ON A PROBLEM IN THE THEORY OF k POPULATIONS[1]

By Raghu Raj Bahadur

University of North Carolina

1. Summary. In two recent papers, Paulson [1] and Mosteller [2] have called attention to several unsolved problems in k-sample theory. A problem which is typical of the ones considered in this paper is as follows.

Let $\pi_1, \pi_2, \cdots, \pi_k$ be a set of normal populations, π_i having an unknown mean m_i and variance σ^2, $G(x, \theta_i)$ being the distribution function which characterizes π_i. Samples of equal size are drawn from each population, $\bar{\bar{X}}_i$ being the sample means, and S^2 the estimate of σ^2 obtained. The problem is to construct a suitable decision rule $d = d(\{\bar{\bar{X}}_i\}; S^2)$ to select one or more populations, the object being to minimize the expected value of the random distribution function

$$G(x \mid s(d)) = \sum_{i=1}^{k} Z_i(d) \cdot G(x, \theta_i) \Big/ \sum_{i=1}^{k} Z_i(d),$$

where $Z_i(d) = 1$ if π_i is selected by d, and $= 0$ otherwise. It is shown that under the restriction of impartial decision, the rule d_k = "Always select only the population corresponding to the greatest $\bar{\bar{X}}_i$" cannot be improved, no matter what x or the true parameter values may be. It follows (i) that d_k is the uniformly best decision rule in the class of impartial decision rules for all weight functions of type

$$W = \max_i \{m_i\} - \left(\sum_{i=1}^{k} z_i m_i \Big/ \sum_{i=1}^{k} z_i\right),$$

and (ii) that the customary F and t tests of analysis of variance are not relevant to the problem.

This result is an application of Theorem 1 which applies to a number of similar problems concerning k populations, especially when the populations admit sufficient statistics for their parameters. Two examples of statistical applications are given in Section 6.

2. Introduction. It has been recognized for some time that the classical theory of statistical inference does not provide direct answers to many problems which are of great interest in the applications. One of them, which arises in the theory of samples from k populations, is what Mosteller has called "the problem of the greatest one." The word "population" is used here for a process, $\pi(\theta)$ say, which generates independent random variables $X_1, X_2, \cdots$, each X having the same distribution function $P(X \leq x) = G(x, \theta)$ say, and a set of X's

[1] This paper is based on a thesis submitted to the Department of Mathematical Statistics, University of North Carolina, in partial fulfilment of the requirements for the Ph. D. degree. This work was sponsored by the Office of Naval Research.

which have been generated by π is called a sample from the population. We shall describe the problem, as also the formulation adopted in the following section, in terms of two special cases. These cases occur when the k given populations $\pi_1, \pi_2, \cdots, \pi_k$ are such that π_i is characterized by the distribution function $G(x, \theta_i) = h\left(\frac{x - b_i}{c_i}\right)$, $\theta_i = (b_i, c_i)$, $c_i > 0$, $i = 1, 2, \cdots, k$, where $h(x)$ is an absolutely continuous non-decreasing function with $h(-\infty) = 0$, $h(+\infty) = 1$. Such sets of populations appear frequently in statistical theory and practice, a given set of normal, or rectangular, or gamma type populations being familiar instances.

CASE 1. Let $X_{ij}, j = 1, 2, \cdots, n$ be a sample from the population π_i, $i = 1, 2, \cdots, k$ where π_i is characterized by the distribution function $h\left(\frac{x - b_i}{c}\right)$, b_i being unknown, and suppose that the statistician is asked to select the population which he thinks has the greatest b_i, but is allowed to select more than one population if (as a consequence, say, of "insignificant" outcomes of tests of differences between populations) he does not feel confident enough to select only one. This situation will occur if, for example, the X_{ij}'s are observed yields in an agricultural experiment in which each of k varieties has been replaced n times, the yield with variety π_i being normally distributed with unknown mean m_i and variance σ^2, and the statistician is asked to recommend one or more varieties for general use. (Cf. Example 1 in Section 6.)

CASE 2. Suppose now that the X_{ij}'s are samples from populations π_i characterized by distribution functions $h\left(\frac{x - b}{c_i}\right)$, $c_i > 0$ unknown, $i = 1, 2, \cdots, k$, and the statistician is asked to select the population which he thinks has the greatest $1/c_i$, but is allowed to select more than one population.[2] This situation will occur if, for instance, the π_i are factories producing an article having a numerical quality characteristic X, $h\left(\frac{x - b}{c_i}\right)$ being the distribution function of X in the product of π_i, and the statistician is required to assign production to one or more factories, the object being to obtain product of stable quality, b being the standard characteristic.

It is clear that the usual statistical theory, which confines itself to estimation of parameters θ_i and testing of hypotheses of the kind $H_0(b_i = \text{constant})$, is inadequate to deal with problems of this sort, where a definite course of action is required of the statistician. It is hardly necessary to add that selection is an important problem in the applications, and the testing of hypotheses is often an indirect attempt to justify selection. In accordance with Wald's formulation of

[2] There is no essential difference between the problem of the greatest one and the problem of the least one. In order to avoid trivial complications, the terminology of the former will be used wherever possible.

the problem of statistical inference,[3] we proceed to consider explicitly the purpose of selection and the "loss" involved in making any particular selection.

3. A class of weight functions. Let $\pi_1, \pi_2, \cdots, \pi_k$ be a given set of populations, π_i being characterized by the distribution function $G(x, \theta_i)$, and let us denote any particular selection, say s, by indicator variables $z_1, z_2, \cdots, z_k$ where $z_i = 1$ if π_i is selected and $= 0$ otherwise. Since any meaningful selection must concern itself with the random variables generated by the populations selected, consider the function $G(x \mid s) = \sum_{i=1}^k z_i G(x, \theta_i) / \sum_{i=1}^k z_i$. $G(x \mid s)$ is a distribution function, and provides a logical and direct overall picture of the effect of making the selection s, since no distinction is made between the populations selected. In immediate generalization, we define a "selection" s to be a vector, $s = (p_1, p_2, \cdots, p_k)$ with $p_i \geq 0$, $\sum_{i=1}^k p_i = 1$, and put $G(x \mid s) = \sum_{i=1}^k p_i G(x, \theta_i)$. Roughly speaking, $G(x \mid s)$ is the distribution function which characterizes the mixed population obtained if sampling rates $p_1, p_2, \cdots, p_k$ are assigned to $\pi_1, \pi_2, \cdots, \pi_k$ respectively, $p_r = 0$ corresponding to rejection of π_r . Henceforth, a selection vector will be called a decision.

Now, if each of the $G(x, \theta_i)$'s were known, an appropriate decision s could be chosen without resort to sampling. If not, the statistician must construct (in advance) and use an s-valued function of the sample values. Such a function, say d, is called a statistical decision function or decision rule. The decision s according to d, say $s(d) = (p_1(d), p_2(d), \cdots, p_k(d))$, is in general a random vector, so that for any fixed x, $G(x \mid s(d))$ is a random variable. Consider the distribution function $H(x \mid d) = E[G(x \mid s(d))] = \sum_{i=1}^k G(x, \theta_i) E[p_i(d)]$, where E denotes the expectation operator. It represents the average overall effect of using the decision rule d, and so affords a reasonable description of the performance of d. Clearly, the problem is to construct d in such a way that $H(x \mid d)$ has desirable properties.

The "desirable properties" will depend, of course, on the particular problem being considered. Returning to our two cases, denote the arbitrary but given set of all possible parameter points $\omega = (\theta_1, \theta_2, \cdots, \theta_k)$ by Ω, and let D be a given class of decision rules $d = d(\{X_{ij}\})$. Then, in Case 1 we wish to choose $d^* \in D$ such that $H(x \mid d^*) = \inf_{d \in D} H(x \mid d)$ for every x and every $\omega \in \Omega$. In Case 2, we wish to choose d^* so that for every x and every ω we have $H(x \mid d^*) = \inf_{d \in D} H(x \mid d)$ whenever $x < b$, and $= \sup_{d \in D} H(x \mid d)$ whenever $x > b$.

These requirements are very strong, and in general no such d^* will exist without heavy restrictions on Ω and on D. (Cf. however the corollary to Theorem 1. It will be found that in a number of cases no restrictions on Ω are required provided that D is the class defined there.) For some purposes, it may be sufficient to consider functionals of $H(x \mid d)$. The functionals which are most useful in the applications are the moments. Thus, one may wish to find d^* such that $\alpha(d^*) = \sup_{d \in D} \alpha(d)$, where $\alpha(d) = \int_{-\infty}^{+\infty} g(x) dH(x \mid d)$, $g(x)$ being some appropriate function.

[3] See, for example, [3], Chapter VI.

For example, in Case 1 we may take $g(x) = x$. Then $\alpha(d)$ is the mean of a random variable having $H(x \mid d)$ for its distribution function, and constructing a suitable d to maximize $\alpha(d)$ is "the problem of the greatest mean." Again, in Case 2 we may take $g(x) = -(x - b)^2$, and in that case maximizing $\alpha(d)$ would be "the problem of the smallest variance."[4]

In terms of mixtures of distributions, $H(x \mid d)$ is the mixture of $G(x \mid s)$ with respect to δ, where δ is the probability measure induced by the decision rule d on the class of Borel sets in the space of all possible decisions s. It follows by the use of Theorem 5 in [4], or otherwise directly, that maximizing $\alpha(d)$ is equivalent to maximizing the expected value (δ) of $\sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} g(x)dG(x, \theta_i)$. Writing $g_i = \int_{-\infty}^{+\infty} g(x)dG(x, \theta_i)$, one may say that the object is to construct d in such a way that the expected value (δ) of the "weight function"

$$W(\omega, s) = \max_i \{g_i\} - \sum_{i=1}^{k} p_i g_i$$

is minimized for every ω. W represents the "loss" incurred by choosing the decision s when the true parameter point is ω. It will be seen that W defined according to (A) in Section 5 includes essentially all weight functions which are likely to be of interest in the type of problem considered in this paper.

We have so far not emphasized the obvious fact that the probability measure δ which is induced by d on the space of decisions will in general depend on the unknown parameter point ω. Therefore, the expected value (δ) of W is to be written as $E[W(\omega, s(d)) \mid \omega] = r(d \mid \omega)$ say. Following the usual terminology, we shall call $r(d \mid \omega)$ the risk function of the rule d, and shall say that $d^* \in D$ is the uniformly best rule in the class D if $r(d^* \mid \omega) = \inf_{d \in D} r(d \mid \omega)$ for all $\omega \in \Omega$.

4. A class of decision rules. The class of decision rules to which we shall confine ourself is rather limited, and may be described as follows, with reference to the previous sections:

(i) Given independent random variables $\{X_{ij}\}, j = 1, 2, \cdots, n; i = 1, 2, \cdots, k$ from the k populations π_i, let

$$X_i = \phi(X_{i1}, X_{i2}, \cdots, X_{in}),\ i = 1, 2, \cdots, k \text{ and } Y = \psi(\{X_{ij}\}),$$

where $X_1, X_2, \cdots, X_k$; Y is an independent set, and the X_i's have frequency functions. The choice of ϕ and ψ will depend upon particular cases: in Case 1, $X_1, \cdots, X_k$; Y will be statistics relevant to the estimation of

[4] An unpublished theorem of Herbert Robbins insures that if a d^* satisfies the strong requirements of the preceding paragraph, it will also maximize all functionals $\alpha(d)$ corresponding to such functions $g(x)$.

$b_1, b_2, \cdots, b_k; c$ respectively, and in Case 2 they will be relevant to $c_1, c_2, \cdots, c_k; b$.[5]

(ii) Given the statistics $\{X_i\}; Y$, $D(\phi; \psi)$ is the class of all impartial decision rules which are based on them. A decision rule $d = d(\{X_i\}; Y)$ is said to be impartial if it has the following structure. Let $X_{(1)} < X_{(2)} < \cdots < X_{(k)}$ be the ordered X_i's. Then d defines non-negative random variables $\lambda_j(X_{(1)}, X_{(2)}, \cdots, X_{(k)}; Y)$, $j = 1, 2, \cdots, k$ such that $\sum_{j=1}^{k} \lambda_j \equiv 1$, and λ_j is the proportion $p(d)$ which is assigned by d to the π corresponding to $X_{(j)}$. We use the term "impartial" for such decision rules because they determine the proportions $[\lambda_1, \lambda_2, \cdots, \lambda_k]$ without regard to which X belongs to which population, and then assign these proportions in strict order of the X_i's.

We shall specify the intuitively plausible class of impartial decision rules for the important normal cases, and give a few instances of such rules.

Suppose first that the X_{ij}'s are from normal populations having means m_i and a common variance σ^2, and that we are interested in the problem of the greatest mean. D is then the class of all impartial decision rules which are based on the statistics

$$X_i = \bar{X}_i = \sum_{j=1}^{n} X_{ij}/n, \qquad i = 1, 2, \cdots, k;$$

$$Y = S^2 = \sum_{j=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2/k\,(n-1).$$

The numerical factors are of no importance, and may be omitted (Cf. footnote 4. See also Example 2 in Section 6, where such factors have been omitted for convenience). A rather simple member of D is the rule $[\lambda_{k-1} \equiv 1/3, \lambda_k \equiv 2/3]$ i.e. "Always assign the proportion 2/3 to the population which has the greatest $\bar{X}_i$, and the proportion 1/3 to the population with the second greatest." In using this rule although the λ_j's remain constant from sample to sample, the decision $s(d)$ is a random vector. In general, however, the λ_j's will themselves be random variables. This is the case if, for instance, one insists on utilising the standard test of differences between populations, and uses the impartial rule "Perform the F test of $H_0(m_i = \text{constant})$ at the five per cent level. If H_0 is rejected, assign the proportion 1 to the population which has the greatest $\bar{X}_i$. If not, assign equal proportions to all populations for which $\bar{X}_i > \sum_{i=1}^{k} \bar{X}_i/k$, and zero proportions to the rest." Another type of impartial decision rule according to which the λ_j's are random variables will be described at the end of Example 1 in the next section. Now, it is (intuitively) clear that if the sample size n is indefinitely large, the rule $[\lambda_k \equiv 1]$, i.e., "Always assign the proportion 1 to the population

[5] It is unnecessary to specify here the exact relation between the statistics and the parameters: (a) the definition of the parameter which determines a distribution function $G(x, \theta)$ is more or less arbitrary, e.g., instead of writing $\theta = (b, c)$ we may write $\theta = (b^3/c, \cosh c)$, and (b) $D(\phi_1; \psi_1) = D(\phi_2; \psi_2)$, provided that $\phi_2 = f(\phi_1)$, $\psi_2 = g(\psi_1)$, where $f(x), g(x)$ are strictly monotonic functions. It will be seen that Theorem 1 is invariant under such transformations of parameters and/or of statistics.

with the greatest $\bar{X}_i$", cannot be improved, no matter what the true parameter values may be. Our main result (Theorem 1) asserts that the statement is in fact valid for any n, provided that one restricts oneself to the class of impartial decision rules.

In a similar way, if the X_{ij}'s are from normal populations having a common mean m and variances σ_i^2, D would be the class of all impartial decision rules which are based on the statistics

$$X_i = S_i^2 = \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2/n - 1, \qquad i = 1, 2, \cdots, k;$$

$$Y = \sum_{i=1}^{k} \sum_{j=1}^{n} X_{ij}/kn,$$

and analogous remarks will apply to this case.

It should be observed that in a given case the appropriate statistics $\{X_i\}$; Y may not be as obvious as in the case of populations like the normal which admit sufficient statistics for their parameters. This real difficulty is not to be confused with the ambiguities mentioned in footnote 4. Furthermore, given the X_i's there may not exist $Y = \psi(\{X_{ij}\})$ which is independent of the X_i's: we shall then assume, without invalidating our result, that the parameter which Y is supposed to estimate is known. Theorem 1 becomes operative only after such questions have been resolved.

5. The uniformly best decision rule. It is convenient to define here some terms which will be used subsequently without further explanation. All functions are assumed to be Borel measurable. Sets will be denoted by curly brackets: thus $\{f = c\}$ is the set on which $f = c$ holds, and $\{a_i\}$ is the set of all a_i in question. "Measure" will refer to ordinary Lebesgue measure in the xy plane.

DEFINITION 1. Given k independent random variables X_i, $i = 1, 2, \cdots, k$, such that each X has a frequency function, let $X_{(j)}$, $j = 1, 2, \cdots, k$, be the ordered set, $X_{(j)}$ being the jth X_i in ascending order of magnitude. Then $A_{ij} = \{X_i = X_{(j)}\}$, and a_{ij} is the characteristic function of the set A_{ij}, that is, $a_{ij} = 1$ for any point of A_{ij} and $= 0$ elsewhere.

Since the X_i's have a joint distribution which is absolutely continuous, the sets A_{ij} are well defined with probability one. Clearly, we have $\sum_{i=1}^{k} a_{ij} = 1$ for every j and $\sum_{j=1}^{k} a_{ij} = 1$ for every i, with probability one.

DEFINITION 2. Let $\beta = (b_1, b_2, \cdots, b_k)$ be a vector of real numbers b_i, and $\phi = (f_1, f_2, \cdots, f_k)$ a vector of real-valued functions $f_i(x)$ defined for every real x. We shall say that $\phi \in T(\beta)$ if for any $r, s = 1, 2, \cdots, k$ for which $b_r \leq b_s$, the set $\{f_r(x)f_s(y) < f_r(y)\, f_s(x), x < y\}$ is of measure zero.

We require the following

LEMMA. *Suppose that $X_1, X_2, \cdots, X_k$; Y are independent random variables, X_i having a frequency function $f_i(x)$ and that $\phi = (f_1, f_2, \cdots, f_k) \in T(\beta)$, where $\beta = (b_1, b_2, \cdots, b_k)$ with*

$$b_1 \leq b_2 \leq \cdots \leq b_k . \tag{1}$$

Then, for any non-negative random variable $\lambda = \lambda(X_{(1)}, X_{(2)}, \cdots, X_{(k)}; Y)$ *and any* $p, q, m = 1, 2, \cdots, k$ *with* $p \leq q$, *we have*

$$\sum_{i=m}^{k} E(\lambda a_{ip}) \leq \sum_{i=m}^{k} E(\lambda a_{iq}). \tag{2}$$

PROOF. Since (2) holds trivially if $p = q$ or if $m = 1$ suppose that $p < q$ and $m \geq 2$. Writing $B(m, j) = \left\{\sum_{i=m}^{k} a_{ij} = 1\right\} = \sum_{i=m}^{k} A_{ij}$, (2) is equivalent to $\int_{B(m,q)} \lambda \, dP \geq \int_{B(m,p)} \lambda \, dP$, and hence to

$$\int_{B(m,q)B'(m,p)} \lambda \, dP \geq \int_{B(m,p)B'(m,q)} \lambda \, dP, \tag{3}$$

where B' denotes the complement of B, and P the probability measure in $(x_1, x_2, \cdots, x_k; y)$ space.

For any permutation $i_1 i_2 \cdots i_k$ of $123 \cdots k$, define $O(i_1 i_2 \cdots i_k) = A_{i_1 1} A_{i_2 2} \cdots A_{i_k k}$. Clearly, the O's corresponding to different permutations are disjoint and each of the sets $B(m, q)B'(m, p)$ and $B(m, p)B'(m, q)$ is the set-theoretic sum of certain O's. Now, it is easy to see that

$$\begin{aligned} O \subset B(m, q)B'(m, p) \text{ if and only if } &\begin{cases} i_p = 1, \text{ or } 2, \cdots, \text{ or } m-1, \text{ and} \\ i_q = m, \text{ or } m+1, \cdots, \text{ or } k. \end{cases} \\ O^* \subset B(m, p)B'(m, q) \text{ if and only if } &\begin{cases} i_p^* = m, \text{ or } m+1, \cdots, \text{ or } k, \text{ and} \\ i_q^* = 1, \text{ or } 2, \cdots, \text{ or } m-1. \end{cases} \end{aligned} \tag{4}$$

Hence a one-one correspondence between subsets $O(i_1 \cdots i_k)$ of $B(m, q)B'(m, p)$ and subsets $O^* = O(i_1^* \cdots i_k^*)$ of $B(m, p)B'(m, q)$ exists through interchange of the pth and qth elements of the defining permutations, the other elements remaining the same. It will be sufficient to prove that if O and O^* are any pair of corresponding subsets, the integral of λ over O is greater than or equal to its integral over O^*, for then (3) will follow by addition.

It is clear that for any O,

$$\begin{aligned} \int_{O(i_1 i_2 \cdots i_k)} \lambda \, dP &= \int_{\{x_{i_1} < x_{i_2} < \cdots < x_{i_k}\}} \lambda(x_{i_1}, x_{i_2}, \cdots, x_{i_k}; y) \\ &\qquad \cdot \left[\prod_{i=1}^{k} f_i(x_i) \, dx_i\right] dF(y) \\ &= \int_R \lambda(t_1, t_2, \cdots, t_k; y) \left[\prod_{r=1}^{k} f_{i_r}(t_r) \, dt_r\right] dF(y), \end{aligned} \tag{5}$$

where R is the domain $\{t_1 < t_2 < \cdots < t_k\}$ and $F(y)$ is the distribution function of Y. Let O and O^* be any pair of corresponding subsets. It follows from (5) that

$$\int_O \lambda \, dP - \int_{O^*} \lambda \, dP = \int_R Q \left[\prod_{r \neq p, q} f_{i_r}(t_r)\right] \prod_{r=1}^{k} dt_r \, dF(y),$$

where

$$Q = \lambda(t_1, t_2, \cdots k_k; y)[f_{i_p}(t_p)f_{i_q}(t_q) - f_{i_q}(t_p)f_{i_p}(t_q)]. \tag{6}$$

From (4) and (1) we have $b_{i_p} \leq b_{i_q}$. Since $p < q$ implies that $t_q < t_q$ over R, and $\phi \in T(\beta)$, it follows that the expression in square brackets in (6) is (except perhaps for a set of measure zero) non-negative over R. Since λ is also non-negative, it follows that Q is non-negative over R, and the Lemma is proved.

We shall now state and prove the main result. Note that the statistic Y is not necessarily real-valued.

THEOREM 1. *Suppose that*

(A). *Ω is a given set of points $\omega = (\theta_1, \theta_2, \cdots, \theta_k)$. $\beta(\omega) = (b_1, b_2, \cdots, b_k)$ and $\gamma(\omega) = (g_1, g_2, \cdots, g_k)$ are defined for every ω such that $b_p \leq b_q$ implies $g_p \leq g_q$ for every $p, q = 1, 2, \cdots, k$.*

Given an $s = (p_1, p_2, \cdots, p_k)$ with $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$,

$$W(\omega, s) = \max_i \{g_i\} - \sum_{i=1}^{k} p_i g_i.$$

(B). *$X_1, X_2, \cdots, X_k$; Y are independent random variables, each X_i having a frequency function $f(x, \theta_i) = f_i(x)$ say, and $\phi(\omega) = (f_1, f_2, \cdots, f_k)$.*

(C). *D is the class of all decision rules d such that*

$d = d(X_{(1)}, X_{(2)}, \cdots, X_{(k)}; Y) = [\lambda_1, \lambda_2, \cdots, \lambda_k]$, $\lambda_j \geq 0$, $\sum_{j=1}^{k} \lambda_j \equiv 1$, and $s(d)$ $= (p_1(d), p_2(d), \cdots, p_k(d))$ where $p_i(d) = \sum_{j=1}^{k} \lambda_j a_{ij}$, $i = 1, 2, \cdots, k$.

Given $d \in D$, $r(d \mid \omega) = E[W(\omega, s(d)) \mid \omega]$.

(D). *For every ω, $\phi \in T(\beta)$.*[6]

Then, for every ω, $r(d_1 \mid \omega) = \sup_{d \in D} r(d \mid \omega)$ and $r(d_k \mid \omega) = \inf_{d \in D} r(d \mid \omega)$, where $d_1 \equiv [1, 0, 0, \cdots, 0]$ and $d_k \equiv [0, 0, \cdots, 0, 1]$.

COROLLARY. *Suppose that π_i, $i = 1, 2, \cdots, k$ are populations characterized by distribution functions $G(x, \theta_i) = h\left(\frac{x - b_i}{c_i}\right)$, $c_i > 0$. For any fixed x, let*

$$G(x \mid \omega, s) = \sum_{i=1}^{k} p_i G(x, \theta_i), \quad \text{and} \quad H(x \mid d, \omega) = E[G(x \mid \omega, s(d)) \mid \omega].$$

CASE 1. *If for every ω,* (i) $c_1 = c_2 = \cdots = c_k$,

(ii) $\phi \in T(\beta)$, *where* $\beta = (b_1, b_2, \cdots, b_k)$,

then, for every ω,

$$H(x \mid d_k, \omega) = \inf_{d \in D} H(x \mid d, \omega).$$

CASE 2. *If for every ω,* (i) $b_1 = b_2 = \cdots = b_k = b(\omega)$, *say,*

(ii) $\phi \in T(\beta)$, where $\beta = (c_1, c_2, \cdots, c_k)$,

---

[6] Note that $\phi \in T(-\beta)$ is equivalent to $\phi^* \in T(\beta)$, where $\phi^* = (f_1^*, f_2^*, \cdots, f_k^*)$, and f_i^* is the frequency function of $X_i^* = -X_i$.

then, for every ω,

$$H(x \mid d_1, \omega) = \begin{cases} \inf_{d \in D} H(x \mid d, \omega) & \text{if } x < b(\omega), \\ \sup_{d \in D} H(x \mid d, \omega) & \text{if } x > b(\omega). \end{cases}$$

PROOF. Choose and fix an arbitrary $\omega \in \Omega$. Without loss of generality we may assume the notation to be so chosen (by simultaneous interchanges of indices i in each of $\{\theta_i\}$, $\{b_i\}$, $\{g_i\}$, $\{p_i\}$, $\{X_i\}$, $\{f_i\}$, and $\{a_{ij}\}$, $j = 1, 2, \cdots, k$) that (1) holds. It then follows that $g_1 \leq g_2 \leq \cdots \leq g_k$ and we write

$$g_i = g_1 + h_1 + h_2 + \cdots + h_i, \qquad h_i \geq 0, \qquad i = 1, 2, \cdots, k. \tag{7}$$

Choose and fix an arbitrary member of the class of impartial decision rules, say $d = [\lambda_1, \lambda_2, \cdots, \lambda_k]$. We have

$$r(d \mid \omega) = \max_i \{g_i\} - \sum_{i,j=1}^{k} g_i E(\lambda_j a_{ij}). \tag{8}$$

Now

$$\begin{aligned} \sum_{i,j=1}^{k} g_i E(\lambda_j a_{ij}) &= \sum_{i,j=1}^{k} (g_1 + h_1 + \cdots + h_i) E(\lambda_j a_{ij}) \\ &= g_1 + \sum_{m,j=1}^{k} \left[\sum_{i=m}^{k} E(\lambda_j a_{ij}) \right] h_m. \end{aligned} \tag{9}$$

Since $\lambda_j = \lambda_j(X_{(1)}, X_{(2)}, \cdots, X_{(k)}; Y) \geq 0$, it follows from the Lemma that

$$\sum_{i=m}^{k} E(\lambda_j a_{ij}) \leq \sum_{i=m}^{k} E(\lambda_j a_{ik}) \quad \text{for every } m \text{ and every } j, \tag{10}$$

by writing $\lambda = \lambda_j$, $p = j$, and $q = k$ in (2). By using (7), (9) and (10) it follows that

$$\begin{aligned} \sum_{i,j=1}^{k} g_i E(\lambda_j a_{ij}) &\leq g_1 + \sum_{m,j=1}^{k} \left[\sum_{i=m}^{k} E(\lambda_j a_{ik}) \right] h_m \\ &= g_1 + \sum_{m=1}^{k} \sum_{i=m}^{k} h_m E(a_{ik}) \\ &= \sum_{i=1}^{k} g_i E(a_{ik}). \end{aligned} \tag{11}$$

Therefore, by (8) and (11),

$$r(d \mid \omega) \geq \max_i \{g_i\} - \sum_{i=1}^{k} g_i E(a_{ik}) = r(d_k \mid \omega), \tag{12}$$

by definition of d_k. The inequality $r(d \mid \omega) \leq r(d_1 \mid \omega)$ follows from (8) and (9) by a similar use of the Lemma. Since both $d \in D$ and $\omega \in \Omega$ are arbitrary, this completes the proof of Theorem 1.

The verification of the corollary is as follows. Choose and fix an arbitrary x and write $h\left(\frac{x - b_i}{c_i}\right) = t_i(\omega)$.

CASE 1. Let $\gamma(\omega) = (1 - t_1 , 1 - t_2 , \cdots , 1 - t_k)$. Then $r(d \mid \omega) = H(x \mid d, \omega) - \min_i \{t_i\}$, and it follows from the Theorem that $H(x|d_1,\omega) = \sup_{d\epsilon D} H(x \mid d, \omega)$ and $H(x \mid d_k , \omega) = \inf_{d\epsilon D} H(x \mid d, \omega)$, for all ω.

CASE 2. Let
$$\gamma(\omega) = \begin{cases} (t_1 , t_2 , \cdots , t_k) & \text{if } b(\omega) > x, \\ (1 - t_1 , 1 - t_2 , \cdots , 1 - t_k) & \text{otherwise.} \end{cases}$$

Then we have
$$r(d \mid \omega) = \begin{cases} \max_i\{t_i\} - H(x \mid d, \omega) & \text{if } b(\omega) > x, \\ H(x \mid d, \omega) - \min_i \{t_i\} & \text{otherwise,} \end{cases}$$

so that
$$H(x \mid d_1 , \omega) = \begin{cases} \inf_{d\epsilon D} H(x \mid d, \omega) & \text{if } b(\omega) > x, \\ \sup_{d\epsilon D} H(x \mid d, \omega) & \text{otherwise,} \end{cases}$$

and conversely for $H(x \mid d_k , \omega)$, for all ω.

The preceding proofs suggest that perhaps (D) is not a necessary condition, but the following theorem for the case of two populations shows that it is indispensable if Theorem 1 is to hold in general.

THEOREM 2. *Suppose that (A), (B), and (C) hold with $k = 2$ and θ_1 , θ_2 real-valued, that the set Ω of points $\omega = (\theta_1 , \theta_2)$ is denumerable, that $\beta(\omega) = \omega$, that $g_1 \neq g_2$ for any ω, and that Y is a fixed constant. Let $\mu(\omega) = \min_i \{\theta_i\}$, $\nu(\omega) = \max_i \{\theta_i\}$, and defining the sets*

$$R(\omega) = \{f(t_1 , \mu) f(t_2 , \nu) < f(t_1 , \nu) f(t_2 , \mu), \quad t_1 < t_2\},$$
$$S(\omega) = \{f(t_1 , \mu) f(t_2 , \nu) > f(t_1 , \nu) f(t_2 , \mu), \quad t_1 < t_2\}$$

in the t_1, t_2-plane, put

$$R^*(t_1 , t_2) = \sum_\omega R(\omega),$$
$$S^*(t_1 , t_2) = \sum_\omega S(\omega).$$

Then a uniformly best decision rule in the class D exists if and only if the set R^*S^* is of measure zero. Subject to existence, the uniformly best rule, say d^*, may be defined as

$$d^* = \begin{cases} [1, 0] \text{ if } (X_{(1)} , X_{(2)}) \epsilon R^*, \\ [0, 1] \text{ otherwise.} \end{cases}$$

The proof is quite simple, and will not be given. It is clear that under the hypotheses of this theorem, the conclusion of Theorem 1 is valid if and only if the set R^* is of measure zero, that is, if and only if condition (D) holds.

6. Examples and discussion. We begin with two applications of Theorem 1.

EXAMPLE 1. Suppose that grain is to be raised on a given area, say A, of land. k varieties, $\pi_1, \pi_2, \cdots, \pi_k$ say, are available, the yields per unit area being normally distributed with unknown means m_i and a common variance σ^2, also unknown. A preliminary field experiment (in which n plots of unit area were assigned to each variety) has been carried out, and $\{X_{ij}\}$, $j = 1, 2, \cdots, n$; $i = 1, 2, \cdots, k$ is the set of independent plot-yields obtained. The statistician is asked to suggest how the available land should be divided between the k varieties, the object being to make the total expected yield as large as possible.[7]

Suppose that an area Ap_i is assigned to π_i, $i = 1, 2, \cdots, k$, with $\sum_{i=1}^{k} p_i = 1$. Then the expected total yield is $\sum_{i=1}^{k} Ap_i m_i$. Our object is to choose the set $(p_1, p_2, \cdots, p_k) = s$ so as to minimize the "loss"

$$W(\omega, s) = \max_i \{Am_i\} - \sum_{i=1}^{k} Am_i p_i .$$

Since the m_i's are unknown, one must construct an appropriate s-valued function of the X_{ij}'s, say d, and set $s(d) = d(\{X_{ij}\})$. The expected "loss" in using this procedure is given by $E[(\omega, s(d)) \mid \omega] = r(d \mid \omega)$, and the problem is to construct a d which makes $r(d \mid \omega)$ as small as possible. (See (A) and (C). Here we have set $\theta_i = (m_i, \sigma)$, $\omega = (\theta_1, \theta_2, \cdots, \theta_k)$, $\beta(\omega) = (m_2, m_3, \cdots, m_k)$ and $\gamma(\omega) = (Am_1, Am_2, \cdots, Am_k)$).

Let $\bar{X}_i = \sum_{j=1}^{n} X_{ij}/n$, $i = 1, 2, \cdots, k$ and $S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2/k(n-1)$.

Since $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_k$; S^2 is a set of sufficient statistics, it is easy to see by taking conditional expectations that corresponding to any decision rule based on the X_{ij}'s, there exists one defined in terms of the $\bar{X}_i$'s and S^2 alone such that the risk functions r of the two are identically equal for all possible values of the unknown parameters. Clearly, one may confine oneself to decision rules of the type $d = s(\{\bar{X}_i\}; S^2)$. Now, the frequency function of $\bar{X}_i$ is $f_i(x) = (n/2\pi\sigma^2)^{1/2} \cdot \exp[-n(x - m_i)^2/2\sigma^2]$, and it is readily seen that $m_r \leq m_s$ and $x < y$ imply $f_r(x)f_s(y) \geq f_s(x)f_r(y)$. It follows that in the class of all impartial procedures which are based on $\{\bar{X}_i\}$; S^2, the uniformly best procedure is to assign the whole area A to the variety with the greatest observed yield. (Note that by the corollary to Theorem 1, a much stronger result than the one required here holds. Cf. footnote 3.)

Although Paulson did not set up a weight function in his discussion of the selection problem for the present case of samples of equal size from k normal populations having unknown means and a common variance, also unknown, he

[7] A double expectation is involved: the expected consequence of a given decision, and the expected decision in using a particular decision rule. The argument given is justified since it is assumed that the random variables generated by the π's subsequent to decision are independent of the random variables on which decision is based. Cf. Section 3. This remark applies to Example 2 also.

gave a class $\{d_c\}$ of decision rules and evaluated some probabilities ($P(G_1)$ and P_s^*. [1], pp. 96–97) which suggest that some of the applications he had in mind are similar to the one given here. In our notation, the rule d_c is defined as follows for any given $c \geq 0$.

$$d_c = [\lambda_1, \lambda_2, \cdots, \lambda_k], \qquad \text{where} \qquad \lambda_j = \left(Z_j \Big/ \sum_{j=1}^{k} Z_j\right), \qquad j = 1, 2, \cdots, k$$

$$\text{with} \qquad Z_j = \begin{cases} 1 & \text{if } X_{(k)} - c(S/\sqrt{n}) \leq X_{(j)} \leq X_{(k)}, \\ 0 & \text{otherwise.} \end{cases}$$

EXAMPLE 2. Suppose that a manufactured article has a numerical characteristic x, and a given article is "defective" if it has an $x < a$ and "acceptable" otherwise, where a is some constant. A consumer requires a large number (N) of articles, which can be supplied by each one of k manufacturers π_i, $i = 1, 2, \cdots, k$. The characteristic (say length) of articles produced by π_i is known to have a rectangular distribution with range from b to $b + c_i$, but the c_i's are not known. As a preliminary step, the consumer has obtained samples of ν articles from each manufacturer, and finds the corresponding lengths to be X_{ij}, $j = 1, 2, \cdots, \nu$; $i = 1, 2, \cdots, k$. The statistician is asked to suggest how the consumer should order a total of N articles from the k manufacturers.

If $a \leq b$, the number of defective articles received by the consumer will be zero no matter how the order is placed. Suppose therefore that $a > b$. Then, if n_i articles are ordered from π_i with $\sum_{i=1}^{k} n_i = N$, the expected number of defectives equals $N - \sum_{i=1}^{k} (n_i/N) \cdot g_i$, where $g_i = g(c_i)$ and $g(t)$ is given by

$$g(t) = \begin{cases} N\left(1 - \dfrac{a-b}{t}\right) & \text{if } t \geq a - b, \\ 0 & \text{otherwise.} \end{cases}$$

Writing $\beta(\omega) = (c_1, c_2, \cdots, c_k)$, $\gamma(\omega) = (g_1, g_2, \cdots, g_k)$, it is clear that the expected number of defectives is of the form $W(\omega, s) + h(\omega)$, where $h(\omega)$ is independent of $s = (n_1/N, n_2/N, \cdots, n_k/N)$, and W is defined as in (A).

We have now to consider what statistics X_i should be used to construct decision rules. Evidently, we are concerned with a "problem of the greatest c_i."

(a). Assuming $\nu > 1$, let $X_i = \max_j \{X_{ij}\} - \min_j \{X_{ij}\}$. Since the frequency function of X_i is $f_i(x) = \nu(\nu - 1)c_i^{-\nu}(c_i - x)x^{\nu-2}$ if $0 < x < c_i$ and zero elsewhere, it is a simple matter to show that $c_r \leq c_s$, $x < y$ imply $f_r(x)f_s(y) \geq f_s(x)f_r(y)$. It follows that in the class of all impartial rules which are based on the sample ranges, the uniformly best rule is to order all the N articles from the manufacturer with the greatest sample range.

(b). It may be objected that since the lower end points of all the distributions are the same, the use of sample ranges to construct decision rules is not particularly appropriate. Suppose therefore that one takes the statistics $X_i^* = \max_j \{X_{ij}\} - b$. The frequency function of X_i^* is $f_i^*(x) = \nu c_i^{-\nu} x^{\nu-1}$ for $0 < x < c_i$ and $= 0$ elsewhere, and as before, condition (D) holds. Hence the uniformly

best impartial procedure in this class is to order all the N articles from the manufacturer who supplied the article with the greatest length in the whole sample of $k\nu$ articles.

It is important to observe that the uniformly best procedures according to (a) and (b) are not identical, and choosing between them is outside the scope of Theorem 1. Note also that the statistics X_i^* are sufficient for the c_i's. Therefore, corresponding to any decision rule there exists a decision rule which is defined in terms of the X_i^*'s and has the same risk function. In particular, there exists a decision rule in class (b) which is equivalent to the uniformly best impartial rule in class (a). It would be interesting to know whether this equivalent rule is also an impartial one.

The two examples given above are purely illustrative, and the reader will readily construct others in which the statistician is faced with similar problems of decision. The second example does not, strictly speaking, belong to Case 2, and the reader is urged to consider some specific instances of this Case. There are various modifications of "the problem of the greatest one" which may be indicated here very briefly. These modifications are introduced by placing restrictions on the class of possible decisions. For example, in Example 1 the statistician may be required to select two or more varieties, and to assign proportions of the land to the varieties which he selects in such a way that no variety takes more than two-thirds of the available land. In that case, the uniformly best procedure (in the class of all impartial procedures which are based on the $\bar{X}_i$'s and S^2) would be to assign two-thirds of the land to the variety with the greatest observed mean yield, and the remainder to the variety with the next greatest. The proof is a slight elaboration of the proof of Theorem 1 and is left to the reader. Again, in Example 2 the consumer may wish to obtain all the articles which he requires from some one manufacturer. In that case, assuming that an impartial selection rule based on the X_i^*'s is to be used, it follows trivially from the case considered previously that the uniformly best procedure is to select the manufacturer with the greatest X_i^* . This is intuitively obvious, but the obvious requires proof (i.e. verification of (D)), as may be seen by turning to Example 3.

The intuitive notion referred to above is one which is employed quite frequently in practice. It may be described as follows. Let X_1 and X_2 be independent and similar estimates of unknown parameters m_1 and m_2 , and suppose that in a given instance we have $X_1 > X_2$. "Then it is more reasonable to suppose that $m_1 > m_2$ than to suppose that $m_1 < m_2$." Theorem 2 shows that this notion is well-founded if and only if condition (D) is satisfied, with $\beta = (m_1 , m_2)$. The condition states essentially that "the likelihood of the greater estimate corresponding to the greater parameter is always $\geq$ the likelihood of the contrary event," and it should be observed that X_1 , X_2 being "good" estimates (e.g. maximum likelihood estimates) does not ensure that this will be the case. The following application of Theorem 2 is an illustration of these remarks.

EXAMPLE 3. Suppose that π_i , $i = 1, 2$ are Cauchy-type populations having medians m_i , and that the set Ω of possible points $\omega = (m_1 , m_2)$ consists of just

the two points $\omega_1 = (1, -1)$ and $\omega_2 = (-1, 1)$. X_1 and X_2 are single observations from the two populations, and the statistician is required to decide which population has the greater median.

Here it would be reasonable for the statistician to use a decision rule, say d^*, which minimizes $r(d \mid \omega) = P(\text{incorrect decision} \mid \omega, d)$, where "$\pi_1$ has the greater median" and "π_2 has the greater median" are the two possible decisions. That this risk function is included in the scheme described by (A) and (C) may be seen as follows. Let the only admissible values of s be (1, 0) and (0, 1), corresponding to the decisions "$m_1 > m_2$" and "$m_1 < m_2$" respectively, and setting $\beta(\omega) = (m_1, m_2)$, define $\gamma(\omega_1) = (1, 0)$, $\gamma(\omega_2) = (0, 1)$. Then for any d such that $s(d)$ equals (1, 0) or (0, 1) only, the expected value of W is for either ω the probability of error in using the rule d.

Now, if $d = d(X_{(1)}, X_{(2)}) = [\lambda_1, \lambda_2]$ is any impartial decision rule, it will equal either [1, 0] or [0, 1], corresponding to the decisions "the population with the greater X has the smaller median" and "the population with the greater X has the greater median" respectively. Since the frequency function of X_i is $f_i(x) = 1/\pi[1 + (x - m_i)]^2$, a little calculation shows that in the class of impartial decision rules a uniformly best one exists, and is given by

$$d^* = \begin{cases} [1, 0] & \text{if } X_{(1)}X_{(2)} > 2, \\ [0, 1] & \text{otherwise.} \end{cases}$$

In conclusion, we remind the reader that although the weight function W defined according to (A) is general enough to include all problems of the type considered in this paper, the sampling scheme as also the class of decision rules to which our results apply is very limited. We have (i) assumed that the samples from the k populations are all of the same size, and (ii) given no objective criterion for choosing appropriate statistics, and no justification for the use of impartial decision rules based on these "appropriate statistics." In view of the applications, it would be of interest to extend the general argument of this paper to the numerous situations where Theorem 1 does not apply or is otherwise unsuitable.

The problem of selection was suggested to the author by Professor Hotelling. The author would like to acknowledge his indebtedness also to Professor Robbins. This paper could not have been written without his constant encouragement and helpful suggestions.

REFERENCES

[1] EDWARD PAULSON, "A multiple decision procedure for certain problems in analysis of variance," *Annals Math. Stat.*, Vol. 20 (1949), pp. 95–98.

[2] FREDERICK MOSTELLER, "A k-sample slippage test for an extreme population," *Annals Math. Stat.*, Vol. 19 (1948), pp. 58–65.

[3] ABRAHAM WALD, *On the Principles of Statistical Inference*, Notre Dame Mathematical Lectures, No. 1, (1942), Notre Dame, Indiana.

[4] HERBERT ROBBINS, "Mixture of distributions," *Annals Math. Stat.*, Vol. 19 (1948), pp. 360–369.