BEST POPULATIONS AND TOLERANCE REGIONS[1][2]

By Irwin Guttman

(Received Feb. 21, 1961)

1. Introduction and summary

We consider a collection of populations $\Pi=(\Pi_1, \cdots, \Pi_k)$ defined over the sample space $\mathfrak{X}(\mathfrak{A})$, where $\mathfrak{A}$ is the σ-algebra of subsets of $\mathfrak{X}$. We suppose there is a class of probability measures defined over $\mathfrak{X}(\mathfrak{A})$ which we designate by $\{P_x^\theta/\theta \in \Omega\}$. Denote the distribution function of Π_m by $P_x^{\theta_m}$, where $\theta_m \in \Omega$.

Now let $\mathfrak{b}_j=\int_A dP_x^{\theta_j}$, $\theta_j \in \Omega$ and $A \in \mathfrak{A}$. $\mathfrak{b}_j$ is called the coverage of the set A. We now make the following

DEFINITION 1.1. A collection of populations contains a best population re the set of interest $A \in \mathfrak{A}$ if and only if there exists an ordering of the $\mathfrak{b}_j$ such that

$$\mathfrak{b}_{[k]} > \mathfrak{b}_{[k-1]} \geqq \mathfrak{b}_{[k-2]} \geqq \cdots \geqq \mathfrak{b}_{[1]} .$$

That is, the best population is one that gives largest coverage to the set $A \in \mathfrak{A}$.

Now it very often happens that a statistician is confronted with k populations, θ_i, $i=1, \cdots, k$, unknown, and it is desirable to know, or find, or pick the "best" population (best in the sense of definition 1.1). Because of the uncertainty involved, the statistician usually settles for a procedure which will select a subset of Π in such a way that the "best" population is included in the subset with probability at least as large as a predetermined number, say P^*. (This is the philosophy of [1] and [2]). If such a procedure selects the best population, we call it a correct selection (CS), and we wish the procedure to be such that the $\Pr(CS) \geqq P^*$.

If in addition, the procedure used is independent of $(\theta_1, \cdots, \theta_k)$, the unknown parameters involved, then we say that the procedure is parameter-free.

We examine the problem of setting parameter-free procedures for collections of normal distributions (section 2) and single exponential distributions (section 3), where A is the interval $(-\infty, a) \in R'$ and "a"

(1) Work supported by a summer grant from the National Research Council of Canada.
(2) Delivered as part of an invited address to the 1960 Annual Meeting of the Institute of Mathematical Statistics, August, 1960.

is a constant that is known and specified beforehand.

2. Normal populations

Suppose we consider a collection of populations $\Pi=(\Pi_1, \cdots, \Pi_k)$ where Π_i is distributed by $N(\mu_i, \sigma_i^2)$. We assume that there is a best population, that is, a population which has the largest value of

$$\int_{-\infty}^{a} dN(\mu, \sigma^2)=\int_{-\infty}^{(a-\mu)/\sigma} dN(0,1) . \tag{2.1}$$

Now we know that $\int_{-\infty}^{t} dN(0,1)$ is a monotone increasing function of t. Hence the problem of selecting the best population is the selection of that population with

$$\text{the largest value of } \frac{a-\mu}{\sigma} \tag{2.2}$$

or

$$\text{the least value of } \frac{\mu-a}{\sigma} . \tag{2.3}$$

This problem splits itself into various cases. To restate, we wish to pick a subset of the k populations (based on independent samples of size n independent observations from each population) in such a way that the probability of a correct selection, $\Pr(CS)\geqq P^*$. We now state the procedures and give the accompanying analysis for the various cases.

Case 2.1: μ's unknown and variable; σ_i^2 known, $\sigma_i^2=\sigma^2$, $i=1, \cdots, k$.

Examining the criterion of bestness for normal populations, that is, (2.3), we see that under condition of case 2.1, a population is best if its mean is least. We assume that for $(\mu_1, \cdots, \mu_k)$, then, that there exists a best population, that is, there is a reordering of the μ's into

$$\mu_{[1]}<\mu_{[2]}\leqq\cdots\leqq\mu_{[k]} . \tag{2.4}$$

Let a sample of n observations be taken independently from each population, and let $\bar{X}_i$ denote the sample mean of the observations X_{ij}, $j=1, \cdots, n$ from Π_i. We adopt the following

Procedure. Retain population Π_i in the subset if

$$\bar{x}_i<\bar{x}_{(1)}+d_1 \tag{2.5}$$

where $\bar{x}_{(1)}$ is the smallest of the k sample means $\bar{x}_i$, and d_1 is a constant chosen to make the probability of a correct selection at least equal to a predetermined number, P^*. We now state the following

THEOREM 2.1: *Procedure* (2.5) *is parameter-free.*

PROOF: We must show that there exists a unique d_1 such that

(i) $\Pr(CS) \geqq P^*$ for procedure (2.5), and

(ii) d_1 is independent of $(\mu_1, \cdots, \mu_k)$.

Now the $\Pr(CS) = \Pr(\bar{X} \leqq \bar{X}_{(1)} + d_1)$ where $\bar{X}$ is the sample mean computed from the best population; that is, $\Pr(CS)$

$$= \Pr(\bar{X}_{(1)} \geqq \bar{X} - d_1)$$

$$= \int_{-\infty}^{\infty} \prod_{i=2}^{k} (1 - G(\bar{x} - d_1, \mu_{[i]}\sigma^2)) dG(\bar{x};\ \mu_{[1]}, \sigma^2)$$

where

$$G(t;\ \mu, \sigma^2) = \int_{-\infty}^{t} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp -\frac{n}{2}\left(\frac{x-\mu}{\sigma}\right)^2 dx.$$

Hence we have that

$$\Pr(CS) = \left(\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\right)^k \int_{-\infty}^{\infty} \int_{\bar{x}-d_1}^{\infty} \cdots \int_{\bar{x}-d_1}^{\infty} \exp -\frac{n}{2\sigma^2} \sum_{2}^{k} (\bar{x}_i - \mu_{[i]})^2$$

$$\cdot \exp -\frac{n}{2\sigma^2} (\bar{x} - \mu_{[1]})^2 \, d\bar{x}_2 \cdots d\bar{x}_k d\bar{x}$$

$$= \left(\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\right)^k \int_{-\infty}^{\infty} \int_{\bar{x}-d_1-\mu[k]}^{\infty} \cdots \int_{\bar{x}-d_1-\mu[2]}^{\infty} \exp -\frac{n}{2\sigma^2} \sum_{2}^{k} t_i^2$$

$$\cdot \exp -\frac{n}{2\sigma^2} (\bar{x} - \mu_{[1]})^2 dt_2 \cdots dt_k d\bar{x}\ .$$

We let $\bar{x} - \mu_{[1]} = t_1$, and we then have that the

$$\Pr(CS) = \left(\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\right)^k \int_{-\infty}^{\infty} \int_{t_1-d_1+\mu[1]-\mu[k]}^{\infty} \cdots \int_{t_1-d_1+\mu[1]-\mu[2]}^{\infty}$$

$$\exp -\frac{n}{2\sigma^2} \sum_{1}^{k} t_i^2 \, dt_2 \cdots dt_k dt_1 = H_{d_1}(\mu_{[1]} - \mu_{[k]}, \cdots, \mu_{[1]} - \mu_{[2]}).$$

An examination of H_{d_1} shows it is a monotone decreasing function in its arguments. Further, if we fix $\mu_{[1]}$, and bearing in mind (2.4), we note that H_{d_1} is minimized for a choice of $\mu_{[2]}$ if we set $\mu_{[2]}$ so close to $\mu_{[1]}$ that for all purposes $\mu_{[2]} = \mu_{[1]}$. Similarly, H_{d_1} is minimized over a choice of $\mu_{[3]}$ if we set $\mu_{[3]} = \mu_{[2]} = \mu_{[1]}$, and finally, H_{d_1} is minimized if

$$\mu_{[1]} = \mu_{[2]} = \cdots = \mu_{[k]}\ .$$

That is, the minimum value of H_{d_1} is $H_{d_1}(0, \cdots, 0)$. Now $H_{d_1}(0,\cdots, 0)$ when regarded as a function of d_1, is continuous and monotone increasing. Hence if we let

$$P^* = H_{d_1}(0,\cdots,0) \tag{2.7}$$

we may solve for d_1 and obtain a unique d_1 which satisfies (2.7), and because $H_{d_1}(0,\cdots, 0)$ is the minimum value of H, then for the "true configuration" (2.4), we have

$$\Pr(CS) \geqq P^*$$

where d_1 is determined from (2.7), and is thus independent of $(\mu_{[1]},\cdots, \mu_{[k]})$. Hence, the theorem is proved.

Case 2.2: μ's unknown and variable; σ^2's known and variable.

Let $\delta_i' = (\mu_i - a)/\sigma_i$, and denote the ordered δ_i' (under the assumption that there is a best population in Π) by

$$\delta'_{[1]} < \delta'_{[2]} \leqq \delta'_{[3]} \leqq \cdots \leqq \delta'_{[k]}$$

We seek to establish a procedure that will choose a subset of Π which contains that population that has $\delta_i' = \delta'_{[1]}$. Let a sample of n independent observations be taken independently from each population, and let $\bar{X}_i$ be the sample mean of the observations taken from Π_i. We let

$$z_i' = \frac{\bar{x}_i - a}{\sigma_i}$$

and denote the ordered z_i' *by*

$$z'_{(1)} < z'_{(2)} < \cdots < z'_{(k)} .$$

We adopt the following

Procedure. Retain population Π_i in the subset if

$$z_i' < z'_{(1)} + d_2 \tag{2.8}$$

where d_2 is a constant chosen to make the $\Pr(CS) \geqq P^*$. We now state and prove the following

THEOREM 2.2: *Procedure* (2.8) *is parameter-free.*

PROOF: We have that the $\Pr(CS) = \Pr(z' < z'_{(1)} + d_2)$ where z' is computed from the population with $\delta' = \delta'_{[1]}$. Now the

$$\begin{aligned}\Pr(CS) &= \Pr(\sqrt{n}\, z' < \sqrt{n}\, z'_{(1)} + \sqrt{n}\, d_2) \\ &= \Pr(z < z_{(1)} + d_2')\end{aligned}$$

where we let $\sqrt{n}\, z'=z$, $\sqrt{n}\, z'_{(1)}=z_{(1)}$ and $\sqrt{n}\, d_2=d_2'$, and we will let $\sqrt{n}\, \delta'_i=\delta_i$.

Hence the

$$\Pr(CS)=\Pr(z_{(1)}>z-d_2')$$
$$=\int_{-\infty}^{\infty}\prod_{i=2}^{k}(1-G^{(1)}(z-d_2';\,\delta_{[i]}))\;dG^{(1)}(z,\,\delta_{[1]})$$

where

$$G^{(1)}(t;\,\delta)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}(x-\delta)^2\,dx;$$

that is,

$$\Pr(CS)=\int_{-\infty}^{\infty}\int_{z-d_2'}^{\infty}\cdots\int_{z-d_2'}^{\infty}\left[\prod_{2}^{k}\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}(z_i-\delta_{[i]})^2\right]$$
$$\cdot\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}(z-\delta_{[1]})^2dz_2\cdots dz_k\,dz$$

$$=\int_{-\infty}^{\infty}\int_{z-d_2'-\delta_{[k]}}^{\infty}\cdots\int_{z-d_2'-\delta_{[2]}}^{\infty}\left[\prod_{2}^{k}\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}t_i^2\right]$$
$$\cdot\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}(z-\delta_{[1]})^2dt_2\cdots dt_k dz\,.$$

Now set $z-\delta_{[1]}=t_1$, that is, $z=t_{(1)}+\delta_{[1]}$. Then the

$$\Pr(CS)=\int_{-\infty}^{\infty}\int_{t_1-d_2'+\delta_{[1]}-\delta_{[k]}}^{\infty}\cdots$$
$$\cdot\int_{t_1-d_2'+\delta_{[1]}-\delta_{[2]}}^{\infty}\left[\prod_{1}^{k}\frac{1}{\sqrt{2\pi}}\exp-\frac{1}{2}t_i^2\right]dt_2\cdots dt_k dt_1$$
$$=H'_{d_2'}(\delta_{[1]}-\delta_{[k]},\cdots,\,\delta_{[1]}-\delta_{[2]})\,.$$

As in Theorem 2.1, it can be shown that the minimum value of H'_{d_2} is $H'_{d_2'}(0,0,\cdots,0)$ and that there exists a unique d_2' for which

$$P^*=H'_{d_2'}(0,\cdots\cdot,0)\,.$$

Hence the theorem is proved, and we may use procedure (2.8) to pick a subset containing the best population with confidence at least P^*, and where $d_2'=\sqrt{n}\;d_2$.

Case 2.3: μ's unknown and variable, σ_i^2 unknown, $\sigma_i^2\equiv\sigma^2$.

It is clear from the criterion (2.3) that here again we wish to retain in our subset that population with the smallest μ. However, we do not know the common value σ^2 of the σ_i^2, and so we use as an estimate the pooled sample variance S^2, where

$$S^2 = \frac{(n-1)s_1^2 + \cdots + (n-1)s_k^2}{k(n-1)} = \frac{1}{k} \sum_{i=1}^{k} s_i^2 \tag{2.9}$$

where

$$S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2, \qquad j=1,\cdots,n \quad \text{and} \quad i=1,\cdots,k.$$

$k(n-1)(S^2/\sigma^2)$ is of course a χ^2-variable with $k(n-1)=\gamma$ degrees of freedom. For this case, we use the following

Procedure. Retain population Π_i if

$$\bar{x}_i \leqq \bar{x}_{(1)} + d_3 S \tag{2.10}$$

where, as usual $\bar{x}_{(1)}$ is the smallest of the $\bar{x}_i$, and d_3 is a constant chosen to make the $\Pr(CS) \geqq P^*$. We now state the following:

THEOREM 2.3: *Procedure* (2.10) *is parameter-free.*

PROOF: We have that the $\Pr(CS) = \Pr(\bar{x} \leqq \bar{x}_{(1)} + d_3 S)$ where $\bar{x}$ is the sample mean computed from $N(\mu_{[1]}, \sigma^2)$. Now the

$$\begin{aligned} \Pr(CS) &= \Pr(\bar{x}_{(1)} \geqq \bar{x} - d_3 S) \\ &= \Pr(\bar{x}_{(1)} - \mu_{[1]} \geqq \bar{x} - \mu_{[1]} - d_3 S) \\ &= \Pr\left(\frac{\bar{x}_{(1)} - \mu_{[1]}}{S} > \frac{\bar{x} - \mu_{[1]}}{S} - d_3\right) \\ &= \int_{-\infty}^{\infty} \left[\prod_{j=2}^{k} (1 - T(t' - d_3; \delta_{[j]})\right] dT(t'; \delta=0) \end{aligned} \tag{2.11}$$

where $t' = (\bar{x} - \mu_{[1]})/S$ and is a Student $t/\sqrt{n}$ variable with $\gamma = k(n-1)$ degrees of freedom,

$$\delta_{[j]} = \frac{\sqrt{n}}{\sigma} [\mu_{[j]} - \mu_{[1]}],$$

$$1 - T(t' - d_3; \delta_{[j]}) = \int_{t'-d_3}^{\infty} f(t_j'; \delta_{[j]}) dt_j',$$

and $f(t_j', \delta_{[j]})$ is the probability density function of the noncentral $t/\sqrt{n}$ variable, noncentrality parameter $\delta_{[j]}$ with $\gamma = k(n-1)$ degrees of freedom, and $T(t', \delta=0)$ is the Student $t/\sqrt{n}$ distribution with γ degrees of freedom given by

$$\int_{-\infty}^{t'} \frac{\sqrt{n}}{\sqrt{\pi\gamma}} \frac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)} \frac{1}{\{1+(n\nu^2/\gamma)\}^{(\gamma+1)/2}} d\nu \ .$$

Since we have the ordering of the μ's, viz

$$\mu_{[1]}<\mu_{[2]}\leqq \cdots \leqq \mu_{[k]}$$

note that this induces an ordering of the δ's, viz

(2.12) $$0<\delta_{[2]}\leqq\delta_{[3]}\cdot \cdots \leqq \delta_{[k]} \ .$$

Now it is well known that $1-T(\omega, \delta)$ is an increasing function of its non-centrality parameter δ. To see this, let X denote an $N(0,1)$ variable. Then by definition of a non-central $t/\sqrt{n}$ variable, we have that

$$1-T(\omega;\ \delta)=\Pr\left(\frac{1}{\sqrt{n}}\frac{X+\delta}{S'}\geqq\omega\right)=\Pr\left(\frac{X+\delta}{S}\geqq\sqrt{n}\,\omega\right)$$

$$=\Pr\left(X\geqq\sqrt{n}\,\omega S-\delta\right) .$$

As δ increases, the region $[(X,S)|X\geqq\sqrt{n}\,\omega S-\delta]$ expands, that is, more and more of the probability measure over the half plane $[(X,S)|-\infty<X<\infty,\ 0<S<\infty]$ is included, and hence $1-T(\omega, \delta)$ increases as δ increases. Now noting the definition of the $\delta_{[j]}$, $j=2,\cdots, k$ and the condition (2.12), we see that the quantity (2.11) attains its minimum if the $\delta_{[j]}$ are zero. (The $\delta_{[j]}$ are never negative since $\mu_{[j]}>\mu_{[1]}$). That is, the minimum value of (2.11), is

$$\int_{-\infty}^{\infty}\prod_{j=2}^{k}(1-T(t'-d_3;\ \delta=0))\ dT(t';\ \delta=0)$$

where $T(t;\ \delta=0)$ is given above. Note that this is a continuous function of d_3, and monotone increasing in d_3, and hence by similar arguments to the above theorems, this theorem is proved, and we can always use procedure (2.10) to select a subset containing the best population under Case 2.3, with confidence at least P^*.

Case 2.4: μ's known, with $\mu_i\equiv\mu$, $i=1,\cdots, k$; σ^2's unknown and variable.

We discuss the case $\mu>a$. Because we are interested in the population with the least value of $(\mu-a)/\sigma_i$, Case 2.4, and the assumption $\mu>a$ implies that we are looking for that population with the largest of the σ_i. Suppose the ordered σ's are

(2.13) $$\sigma^2_{[1]}\leqq\sigma^2_{[2]}\leqq\cdots\leqq\sigma^2_{[k-1]}<\sigma^2_{[k]}$$

that is, there exists a best population in the sense of Definition 1.1. Suppose again that independent samples of n independent observations, X_{ij} are taken, where $i=1,\cdots,k$; $j=1,\cdots,n$. Let

$$v_i^2=\frac{1}{n}\sum_{j=1}^{n}(X_{ij}-\mu)^2 \,. \tag{2.14}$$

Let $v^2_{(1)}<\cdots<v^2_{(k)}$ be the ordered v_i^2's. We use the following

Procedure. Retain Π_i in the subset if

$$v_i^2\geqq d_4 v^2_{(k)} \tag{2.15}$$

where d_4 is a constant such that $0<d_4<1$, and is chosen so that the $\Pr(CS)\geqq P^*$. Again, we may state the following

THEOREM 2.4: *Procedure* (2.15) *is parameter-free.*

PROOF: We have that the $\Pr(CS)=\Pr(v^2\geqq d_4 v^2_{(k)})$ where v^2 is the sample variance defined in (2.14) computed from the best population. That is, the

$$\Pr(CS)=\Pr\left(v^2_{(k)}\leqq\frac{v^2}{d_4}\right)$$
$$=\int_0^\infty\left[\prod_{i=1}^{k-1}C\left(\frac{v^2}{d_4};\ \sigma^2_{[i]}\right)\right]dC(v^2;\ \sigma^2_{[k]})$$

where

$$C(v^2;\ \sigma^2_{[i]})=\int_0^{v^2}\frac{1}{\Gamma(n/2)}\frac{n^{n/2}}{(2\sigma_1^2)^{n/2}}\exp\frac{-nv_i^2}{2\sigma^2_{[i]}}(v_i^2)^{(n/2)-1}\ dv_i^2 \,.$$

Hence we may see that the

$$\Pr(CS)=\int_0^\infty\int_0^{(\omega_k^2/d4)(\sigma^2_{[k]}/\sigma^2_{[k-1]})}\cdots\int_0^{(\omega_k^2/d_4)(\sigma^2_{[k]}/\sigma^2_{[1]}}\cdots\left[\prod_{i=1}^{k}\frac{n^{n/2}}{\Gamma(n/2)2^{n/2}}(\omega_i^2)^{(n/2)-1}\right.$$
$$\left.\cdot\exp\frac{-n\omega_i^2}{2}\right]d\omega_1^2\cdots d\omega_{k-1}^2 d\omega_k^2$$
$$=K_{d_4}\left(\frac{\sigma^2_{[k]}}{\sigma^2_{[k-1]}},\ \cdots,\ \frac{\sigma^2_{[k]}}{\sigma^2_{[1]}}\right).$$

Now K_{d_4} is a monotone increasing function in its arguments, subject to (2.13). It is obvious that the minimum value of K_{d_4} is $K_{d_4}(1,\cdots,1)$ and hence if we set $K_{d_4}(1,\cdots,1)=P^*$ we may find a unique d_4 which makes

$$\Pr(CS)\geqq P^*$$

and the procedure (2.15) is parameter-free.

(It should be pointed out that if one analyzes the case $\mu<a$, best population is that population with the least σ^2, and that one may verify that the

Procedure. Retain Π_i if

$$v_i^2 \leqq d_4' v_{(1)}^2 \,, \tag{2.15}$$

is parameter-free, where $v_{(1)}^2$ is the smallest of the v_i^2's defined in (2.14), and d_4' is a constant chosen to make the $\Pr(CS) \geqq P^*$. In fact it can be shown that the

$$\Pr(CS)=\int_0^\infty \int_{(\omega_1^2/d'_4)(\sigma_{[1]}^2/\sigma_{[k]}^2}^\infty \cdots \int_{(\omega_1^2\sigma_{[1]}^2)/d_4'\sigma_{[2]}^2)}^\infty \left[\prod_{i=1}^k \frac{n^{n/2}}{2^{n/2}\Gamma(n/2)} (\omega_i^2)^{(n/2)-1} \cdot \exp \frac{-n\omega_i^2}{2}\right] d\omega_2^2 \cdots d\omega_k^2 d\omega_1^2 = K'_{d_4'}\left(\frac{\sigma_{[1]}^2}{\sigma_{[k]}^2}, \cdots, \frac{\sigma_{[1]}^2}{\sigma_{[2]}^2}\right)$$

where under the assumption of the existence of a best population in we have

$$\sigma_{[1]}^2 < \sigma_{[2]}^2 \leqq \cdots \leqq \sigma_{[k]}^2$$

and hence that the

$$\Pr(CS) \geqq K'_{d_4'}(1, \cdots, 1) \,.$$

On setting the right hand member of the above inequality to P^*, we obtain a unique d_4' satisfying $K'_{d_4'}(1, \cdots, 1) = P^*$, and hence (2.15) is parameter-free).

Case 2.5: μ's known, variable; σ's unknown and variable.

Again, let us assume that we have a collection of normal populations Π_i, and that they are distributed by the $N(\mu_i, \sigma_i^2)$ distribution. Bearing in mind the condition of Case 2.5, and that we seek to find that population with least $(\mu_i - a)/\sigma_i$, it is readily seen that this case splits into the following three cases.

Case 2.5 (a) All μ_i known and less than a
(b) All μ_i known and greater than a
(c) All μ_i known, with $\mu_{[1]} < \mu_{[2]} < \cdots \mu_{[k_1]} < a$, and $a < \mu_{[k_1+1]} < \cdots < \mu_{[k]}$ where $1 < k_1 < k$.

The case (2.5a) will be readily seen to be symmetric and analogous to case (2.5b). Further, if for a normal distribution, the population mean is such that $\mu > a$, then the coverage of $(-\infty, a)$ is less than $\frac{1}{2}$. That is, for the case (2.5c), we can disregard those populations Π_j with $a < \mu_j$, and formulate a procedure for selecting the best population out of the remaining k_1 populations (that have their means $\mu < a$). Of course, this will be the same solution for case (2.5a). Note that $k_1 > 1$ for if $k_1 = 1$, then automatically we know the best population.

We now discuss, then, the problem of finding the best normal population of a collection $\Pi = (\Pi_1, \cdots, \Pi_k)$ of normal populations, where the means are known, and $\mu_i < a$, $i = 1, \cdots, k$, and where best population implies, as we have seen, the population with the largest $(a-\mu_i)/\sigma_i$. That is, we wish to select a subset of Π in such a way that the population with the smallest $\sigma_i/(a-\mu_i)$ is retained in our subset, with probability of this correct selection at least P^*.

Using the notation of the previous cases, let

$$v_i^2 = \frac{1}{n} \sum_{j=1}^{n} (X_{ij} - \mu_i)^2 \qquad i = 1, \cdots, k$$

be the unbiased estimate of σ_i^2. Let

$$q_i = \frac{v_i}{a - \mu_i} \tag{2.16}$$

and denote the ordered q_i's by

$$q_{(1)} < q_{(2)} < \cdots < q_{(k)}$$

We now state the following

Procedure. Retain Π_i if

$$q_i \leqq d_5 q_{(1)} \tag{2.17}$$

where d_5 is a constant chosen to make the $\Pr(CS) \geqq P^*$, and is such that $1 < d_5$. We now prove the following

THEOREM 2.5: *Produre* (2.17) *is parameter-free.*

PROOF: We have that the $\Pr(CS) = \Pr(q \leqq d_5 q_{(1)})$, where q denotes that q_i which is computed from the population having smallest $\sigma_i/a-\mu_i$, that is, the best population.

Let $\delta_i = \dfrac{\sigma_i}{a-\mu_i}$ and denote the ranked δ's by

$$\delta_{[1]} < \delta_{[2]} \leqq \delta_{[3]} \leqq \cdots \leqq \delta_{[k]} \,. \tag{2.18}$$

Then we have that

$$\Pr(CS)=P\left(q_{(1)}\geqq\frac{q}{d_5}\right)$$
$$=\int_0^\infty\left[\prod_{i=2}^{k}\left(1-M\left(\frac{q}{d_5};\ \delta_{[i]}\right)\right)\right]dM(q;\ \delta_{[1]})$$

where $$1-M(q/d_5;\ \delta_{[i]})=\int_{q/d_5}^\infty\frac{1}{\delta_{[i]}2^{(n/2)-1}\Gamma\left(\frac{n}{2}\right)}\left(\frac{q_i}{\delta_{[i]}}\right)^{n-1}e^{-\delta_i^2/2\delta_{[i]}^2}dq_i\,.$$

By using procedures similar to the above, we may verify that the

$$\Pr(CS)=\int_0^\infty\int_{(\omega_1/d_5)(\delta_{[1]}/\delta_{[k]})}^\infty\cdots\int_{(\omega_1/d_5)(\delta_{[1]}/\delta_{[2]})}^\infty\left[\prod_{i=1}^{k}\frac{e^{-\omega_i^2/2}\omega_i^{n-1}}{2^{(n/2)-1}\Gamma\left(\frac{n}{2}\right)}\right]d\omega_2\cdots d\omega_k\,d\omega_1$$
$$=U_{d_5}\left(\frac{\delta_{[1]}}{\delta_{[k]}},\cdots,\frac{\delta_{[1]}}{\delta_{[2]}}\right).$$

An examination of U_{d_5} shows it is a monotone decreasing function of its arguments, subject to (2.18). It is easy to see that U_{d_5} has minimum value

$$U_{d_5}(1,\cdots,1)$$

that is, the minimum value occurs when

$$\delta_{[1]}=\delta_{[2]}=\cdots=\delta_{[k]}\,.$$

Note that $U_{d_5}(1,\cdots,1)$ is a continuous and monotone increasing function of d_5. Hence there exists a unique d_5 which satisfies

$$U_{d_5}(1,\cdots,1)=P^*$$

and for this unique d_5,

$$\Pr(CS)\geqq U_{d_5}(1,\cdots,1)=P^*$$

that is, procedure (2.17) is parameter-free.

(It should be pointed out that for the case (2.5*b*), that under the assumption of the existence of a best population, we wish to retain the population with largest value of $\delta'_i=\sigma_i/(\mu_i-a)$. We let $q_i'=v_i/(\mu_i-a)$ and it can be verified that the

Procedure. Retain population Π_i if

(2.19) $$q_i'>d_5'\,q'_{(k)}$$

where $q'_{(k)} = \max_{i=1}^{k} q'_i$, and d'_5 is a constant, $0 < d'_5 < 1$, such that the $\Pr(CS) \geqq P^*$, is parameter-free. In fact, the

$$\Pr(CS) = \int_0^\infty \int_0^{(\omega_k/d'_5)(\delta'_{[k]}/\delta'_{[k-1]})} \cdots \int_0^{(\omega_k/d'_s)/(\delta'_{[k]}/(\delta'_{[1]}} \left[\prod_{i=1}^{k} \frac{e^{(-\omega_i^2/2)} \omega_i^{n-1}}{2^{(n/2)-1} \Gamma(n/2)} \right]$$

$$\cdot d\omega_i \cdots d\omega_{k-1} d\omega_k$$

$$= U'_{d'_5}\left(\frac{\delta'_{[k]}}{\delta'_{[k-1]}}, \cdots, \frac{\delta'_{[k]}}{\delta'_{[1]}} \right).$$

It is easy to see that the minimum value of U'_{d_5} is $U'_{d_5}(1, \cdots, 1)$ and setting this equal to the desired P^*, gives a unique d'_5, independent of the δ'_i and hence (2.19) is parameter-free.

3. Exponential populations

We turn now to the situation where our collection Π of k populations is exponentially distributed, that is, the probability density function of the ith population Π_i is

$$(3.1) \qquad \begin{cases} \dfrac{1}{\sigma_i} \exp - \dfrac{1}{\sigma_i}(x - \mu_i) & x \geqq \mu_i, \ \sigma_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

We again assume that the set of interest is $A = (-\infty, a)$, where "a" is a known constant. As in the normal cases discussed above, we assume in the sequel that there always exist a best population in the sense of Definition 1.1 in the collection. Again, as in the normal cases, the best population is that with the largest value of $(a - \mu)/\sigma$. We discuss the following cases.

Case 3.1: μ's known, $\mu_i \equiv \mu$, $i = 1, \cdots, k$; σ_i's unknown and variable.

If the known value of μ is such that $\mu < a$, it is clear that the best population is that with least σ. If $\mu > a$, the exponential distribution whose density is defined in (3.1), gives zero coverage to $(-\infty, a)$ and hence there would not be a best population re this set of interest in the collection Π, contrary to assumption, and we thus disregard this problem.

We now assume, then, that $\mu < a$, and let k independent samples of n independent observations be taken, and let $Y_{ij} = X_{ij} - \mu$. The Y_{ij} have density functions

$$(3.2)\qquad \begin{cases} \dfrac{1}{\sigma}e^{-y/\sigma}, & y\geqq 0,\ \sigma>0 \\ 0 & \text{otherwise}. \end{cases}$$

We adopt the following

Procedure. Retain Π_i if

$$(3.3)\qquad \bar{y}_i<f_1\,\bar{y}_{(1)}$$

where $\bar{y}_i=n^{-1}\sum_{j=1}^{n} Y_{ij}$, $\bar{y}_{(1)}$ is the smallest of the $\bar{y}_i$, and f_1 is a constant chosen to make the $\Pr(CS)\geqq P^*$.

THEOREM 3.1: *Procedure* (3.3) *is parameter-free.*

PROOF: It is straightforward to show that the

$$\Pr(CS)=\int_0^\infty\int_{(t_1/f_1)(\sigma_{[1]}/\sigma_{[k]})}^\infty\cdots\int_{(t_1/f_1)(\sigma_{[1]}/\sigma_{[2]})}^\infty\left[\prod_{i=1}^{k}\frac{n^n}{\Gamma(n)}e^{-nt_i}\,t_i^{n-1}\right]dt_2\cdots dt_k\,dt_1$$
$$=V_{f_1}\left[\frac{\sigma_{[1]}}{\sigma_{[k]}},\cdots,\frac{\sigma_{[1]}}{\sigma_{[2]}}\right].$$

An examination of V_{f_1} shows it is a monotone decreasing function of its arguments, subject to

$$(3.4)\qquad \sigma_{[1]}<\sigma_{[2]}\leqq\cdots\leqq\sigma_{[k]}\,.$$

Hence the $\Pr(CS)\geqq V_{f_1}(1,\cdots,1)$. But V_{f_1} is a monotone increasing function and continuous in f_1, and thus there exists a uniqne f_1 such that $V_{f_1}(1,\cdots,1)=P^*$, and which is clearly independent of the parameters $\sigma_{[i]}$, that is, (3.3) is parameter-free and its use enables us to make a correct selection with $\Pr(CS)P\geqq P^*$.

Case 3.2: μ's unknown and variable, $\sigma_i\equiv\sigma$, $i=1,\cdots,k$ and known.

We assume that there is a best population, that is, there is at least one of the k Π_i having $\mu_i<a$.

Let $t_i=x^i_{(1)}=\min_{j=1}^{n} x_{ij}$, and let $t_{(1)},\cdots,t_{(k)}$ be the ordered t_i's. Note that the best population is the one with the least μ. Hence we adopt the following

Procedure. Retain Π_i if

$$(3.5)\qquad t_i\leqq t_{(1)}+f_2$$

where f_2 is a constant chosen to make the $\Pr(CS)\geqq P^*$.

THEOREM 3.2: *Procedure* (3.5) *is parameter-free.*

PROOF: It is straight forward to verify that the

$$\Pr(CS)=\int_0^\infty \int_{\omega_1-f_2+\mu_{[1]}-\mu_{[k]}}^\infty \cdots \int_{\omega_1-f_2+\mu_{[1]}-\mu_{[2]}}^\infty \left[\prod_{i=1}^k \frac{n}{\sigma} e^{-n\omega_i/\sigma}\right] d\omega_2 \cdots d\omega_k\, d\omega_1$$
$$= W_{f_2}(\mu_{[1]}-\mu_{[k]}, \cdots, \mu_{[1]}-\mu_{[2]})$$

where $\mu_{[1]}<\mu_{[2]}\leqq\cdots\leqq\mu_{[k]}$. Hence the $\Pr(CS)\geqq W_{f_2}(0,\cdots,0)$ and if we set $W_{f_2}(0,\cdots,0)=P^*$, there exists a unique f_2 satisfying this latter equation, independent of the μ's, and hence parameter-free, with the $\Pr(CS)\geqq P^*$.

Case 3.3: μ's unknown, variable; σ_i's known and variable.

We again assume that there is a best population, that is, at least one of the $k\ \Pi_i$ have $\mu_i<a$. Let $\delta_i=(\mu_i-a)/\sigma_i$ and let the ordered δ's be denoted by

$$(3.6) \qquad \delta_{[1]}<\delta_{[2]}\leqq\delta_{[3]}\leqq\cdots\leqq\delta_{[k]}\ .$$

Clearly we wish to select the population with its $\delta=\delta_{[1]}$. Now let

$$X^i_{(1)}=\min_{j=1}^{n} X_{ij} \qquad i=1,\cdots,k$$

let $Z_i=(X^i_{(1)}-a)/\sigma_i$. We adopt the following

Procedure. Retain Π_i if

$$(3.7) \qquad Z_i\leqq Z_{(1)}+f_3$$

where $Z_{(1)}$ is the smallest of the Z_i and f_3 is a constant chosen to make the $\Pr(CS)\geqq P^*$. We now state the following

THEOREM 3.3: *Procedure* (3.7) *is parameter-free.*

PROOF: Using the same analysis as in the previous cases, it is readily verified that the

$$\Pr(CS)=\int_0^\infty \int_{\omega_1-f_3+\delta_{[1]}-\delta_{[k]}}^\infty \cdots \int_{\omega_1-f_3+\delta_{[1]}-\delta_{[2]}}^\infty \left[\prod_1^k n\, e^{-n\omega_i}\right] d\omega_2 \cdots d\omega_k\, d\omega_1$$
$$=L_{f_3}(\delta_{[1]}-\delta_{[k]}, \cdots, \delta_{[1]}-\delta_{[2]})$$

where the δ's are subject to (3.6). The minimum value of L_{f_3} is $L_{f_3}(0,\cdots,0)$ and if we set $L_{f_3}(0,\cdots,0)=P^*$, there exists a unique f_3 satisfying this latter equation and independent of the μ_i and σ_i; that is, procedure (3.7) is parameter-free and is such that the $\Pr(CS)\geqq P^*$.

Case 3.4: μ's known and variable, σ's unknown and variable.

This case splits itself into the following cases;

Case 3.4(a) All μ_i known and such that $\mu_i<a$, $i=1, \cdots, k$.

Case 3.4(b) All μ_i known and such that $\mu_i>a$, $i=1, \cdots, k$.

Case 3.4(c) All μ_i known with $\mu_{[1]}<\cdots<\mu_{[k_1]}<a$ and

$a<\mu_{[k_1+1]}<\cdots<\mu_{[k]}$, where $1<k_1<k$.

Case (3.4b) can obviously be disregarded since for the exponential distribution as defined by (3.1), the coverage of $(-\infty, a)$ is zero if $\mu_i>a$. Case (3.4c), then, is such that we can immediately disregard the $k-k_1$ populations which are such that their $\mu_i>a$, and use a procedure to find the best population of the remaining k_1 populations which are such that their $\mu_i<a$. Of course, this is case (3.4a) with k_1 replaced by k. Note that $k_1>1$, for if $k_1=1$, we automatically know the best population. We therefore formulate a procedure for case (3.4a), which can be used if case (3.4c) obtains.

Let k independent samples of n independent observations be taken, and let $(X^i_{(1)}, \cdots, X^i_{(n)})$ denote the n ordered observations from population Π_i.

Define $S_i=(n-1)^{-1}\sum_{j=2}^{n}(X^i_{(j)}-X^i_{(1)})$. To restate, we wish to find that population with the largest value of $(a-\mu_i)/\sigma_i$, where the μ_i are known and less than a, and thus we wish to find the population with the least value of $\delta_i=\sigma_i/(a-\mu_i)$. We let

$$\delta_{[1]}<\delta_{[2]}\leqq\cdots\leqq\delta_{[k]} \tag{3.8}$$

denote the ordered δ_i's.

Let $z_i=s_i/(a-\mu_i)$ and let $z_{(1)}<z_{(2)}<\cdots<z_{(k)}$ denote the ordered z_i's. We now formulate the following

Procedure. Retain Π_i if

$$z_i\leqq f_4\, z_{(1)} \tag{3.9}$$

where f_4 is a constant chosen to make the $\Pr(CS)\geqq P^*$.

THEOREM 3.4: *Procedure* (3.9) *is parameter-free.*

PROOF: Since the probability density function of z_i is given by

$$\frac{(n-1)^{n-1}}{\delta_i^{n-1}\Gamma(n-1)}\, z_i^{n-2}\exp\{-(n-1)z_i/\delta_i\}\ dz_i$$

it is easy to see that the $\Pr(CS)$ is given by

$$\Pr(CS)=\int_0^\infty \int_{(\omega_1/f_4)(\delta_{[1]}/\delta_{[k]})}^\infty \cdots \int_{(\omega_1/f_4)(\delta_{[1]}/\delta_{[2]})}^\infty \left[\prod_{i=1}^k \frac{(n-1)^{n-1}}{\Gamma(n-1)} \omega_i^{n-2} e^{-(n-1)\omega_i}\right]$$

$$\cdot d\omega_2 \cdots d\omega_k \, d\omega_1$$

$$=M_{f_4}\left(\frac{\delta_{[1]}}{\delta_{[k]}}, \cdots, \frac{\delta_{[1]}}{\delta_{[2]}}\right)$$

where $\delta_{[1]}<\delta_{[2]}\leqq \cdots \leqq \delta_{[k]}$. M_{f_4} is a monotone decreasing function in its arguments, and hence

$$\Pr(CS) \geqq M_{f_4}(1, \cdots, 1) .$$

If we set $M_{f_4}(1, \cdots, 1)=P^*$, and because the function $M_{f_4}(1, \cdots, 1)$ we see that there exists a unique f_4 satisfying this last equation, and is independent of the parameters δ_i, that is, the procedure (3.9) is parameter-free and such that the $\Pr(CS) \geqq P^*$.

Case 3.5: μ's unknown and variable; σ_i unknown, $\sigma_i=\sigma$.

Before analyzing this case, we discuss an analogue of the Student-t variable, to be denoted by the symbol U_v, and called the central U-variable with v degrees of freedom. We denote the exponential distribution by $E(\mu_i, \sigma_i)$, whose density function is given by expression (3.1).

Now let Y be a random variable which is distributed as a $\gamma(v)/v$ variables, that is, Y has the density function

$$(3.10) \qquad \begin{cases} \dfrac{V^v}{\Gamma(v)} y^{v-1} e^{-vy} \, dy & y \geqq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Further, let W be an $E(0,1)$ variable, and suppose that W and Y are independent. Define $U_v = W/Y$, and it is easy to see that the distribution of U has the density function given by

$$(3.11) \qquad \begin{cases} \dfrac{du}{[1+(U/v)]^{v+1}} & \textit{if } U>0 \\ 0 & \text{otherwise.} \end{cases}$$

We define $U_v^1=(W+\delta)/Y$, to be called the non-central U variable, noncentrality parameter δ, with v degrees of freedom. Although we do not derive its density, we note that its "anti-cumulative,"

$$(3.12) \qquad 1-H(C;\ \delta)=\Pr(U_v^1 \geqq C)$$

is an increasing function of δ, for this is the

$$\Pr(W > CY - \delta)$$

and as δ increases, more and more of the probability measure over the region $\{(W, Y) | 0 < W, Y < \infty\}$ is included.

Now suppose we take k independent samples of n observations and let $(X^i_{(1)}, \cdots, X^i_{(n)})$ be the ordered observations from Π_i.

Let $S_i = (n-1)^{-1} \sum_{j=2}^{n} (X^i_{(j)} - X^i_{(1)})$. Now it is known that if sampling from $E(\mu_i, \sigma_i)$, that $X^i_{(1)}$ and S_i are independent (and sufficient for μ_i, σ_i). Further $n(X^i_{(1)} - \mu_i)/\sigma_i$ has the $E(0,1)$ distribution and S_i is a $\gamma(n-1)/n-1$ variable.

For the case being considered, we have $\sigma_i = \sigma$, but σ is unknown. We will therefore use the pooled estimate

$$S = \frac{(n-1)S_1 + \cdots + (n-1)S_k}{k(n-1)} = \frac{1}{k} \sum_{i=1}^{k} S_i \tag{3.13}$$

and it is easy to see that S is a $\sigma\{\gamma(k(n-1))/k(n-1)\}$ variable.

Now, we wish to find the population with least $(\mu_i{}^{-a})/\sigma$, that is, with least μ_i. We assume, of course, that there is at least one population with $\mu_i < a$.

Let $t_i = nX^i_{(1)}$, and denote the ordered t's by

$$t_{(1)} < t_{(2)} < \cdots < t_{(k)}. \tag{3.14}$$

We now adopt the following

Procedure. Retain Π_i if

$$t_i < t_{(1)} + f_5 S \tag{3.15}$$

where f_5 is a constant chosen to make the $\Pr(CS) \geqq P^*$. We now state the following

THEOREM 3.5: *Procedure* (3.15) *is parameter-free.*

PROOF: We have that the

$$\begin{aligned}\Pr(CS) &= \Pr(t_{(1)} \geqq t - f_5 S) \\ &= \Pr\left(\frac{t_{(1)} - n\mu_{[1]}}{S} \geqq \frac{t - n\mu_{[1]}}{S} - f_5\right) \\ &= \Pr(U^1_{(1)} > U - f_5)\end{aligned}$$

where t is that t_i computed from the population having $\mu = \mu_{[1]}$,

U is a central U variable with $k(n-1)$ degrees of freedom, and

$U^1_{(1)}$ is a non-central U^1 variable with $k(n-1)$ degrees of freedom, and non-centrality parameter

$$\delta_{[i]}=n\left(\frac{\mu_{[i]}-\mu_{[1]}}{\sigma}\right)$$

where $i\neq 1$. Hence the

$$\Pr(CS)=\int_0^\infty \left[\prod_2^k (1-H(U-f_5;\ \delta_{[i]}))\right] dG(U)$$

where $G(U)$ is the distribution function of (3.11) with v put equal to $k(n-1)$, and $1-H(C;\ \delta)$ is given by (3.12). Now we have that $1-H$ is an increasing function in δ, and the $\Pr(CS)$ depends on a product of the $(k-1)$ function, $1-H(U-f_5;\ \delta_{[i]})$, where

$$0<\delta_{[2]}<\cdots<\delta_{[k]}.$$

Therefore the $\Pr(CS)$ is minimized if $\delta_{[2]}=\cdots=\delta_{[k]}=0$, and we have that the

$$\begin{aligned}\Pr(CS)&\geqq\int_0^\infty \left[\prod_2^k (1-G(U-f_5))\right] dG(U)\\ &=\int_0^\infty\int_{u-f_5}^\infty\cdots\int_{u-f_5}^\infty\left[\prod_1^k\left(1+\frac{U}{k(n-1)}\right)^{-[k(n-1)+1]}\right]dU_2\cdots dU_k\,dU.\end{aligned}$$

The last expression is a monotone increasing and continuous function of f_5, and if we set it equal to P^*, there is a unique f_5 satisfying the resulting equation, and which is independent of the parameters. That is, (3.15) is parameter-free and such that the $\Pr(CS)\geqq P^*$.

4. Acknowledgements

Thanks are due to Dr. S. S. Gupta, Bell Telephone Laboratories, for bringing to the author's attention the approach of the work in [1] and [2], and to Dr. D. A. Freedman, Princeton University, for stimulating and helpful discussion.

McGill University

References

[1] S. S. Gupta, "On a decision rule for a problem in ranking means," *Mimeograph Series*, No. 150 (1956), Institute of Statistics,University of North Carolina.

[2] Shanti S. Gupta and Milton Sobel, "On selecting a subset which contains all populations better than a standard," *Ann. Math. Stat.*, Vol. 29 (1958), p. 235.

[3] E, L, Lehmann, "Problems of selection" unpublished manuscript.