AN OPTIMUM SLIPPAGE TEST FOR THE VARIANCES OF k NORMAL DISTRIBUTIONS

By Donald R. Truax

University of Washington

1. Summary and introduction. In many practical problems, the experimenter is faced with the task of deciding if the variability within several classes is uniform throughout the classes, or if not, which class exhibits the greatest amount of variability. This type of problem arises when the data relate to several processes, to the same process at different times, to several different products, or to the same products from different sources. If the variability is not uniform throughout the classes, then misleading results would be obtained in comparing the classes in other respects. If the experimenter expects the variability to be uniform throughout the different classes, and if the variability is large in a particular class, he will consider the situation to be "out of control" and take measures to locate the source of the large variability.

The problem we will consider here is that of comparing the variances of k populations, $\Pi_1, \Pi_2, \cdots, \Pi_k$, on the basis of n observations $x_{i1}, x_{i2}, \cdots, x_{in}$ from the ith population. We will assume that these observations are normally and independently distributed with unknown mean m_i and unknown standard deviation σ_i for $i = 1, 2, \cdots, k$. Our problem is to find a statistical procedure which will, on the basis of these observations, decide if all the populations have equal variances, and if not, which has the largest variance. We would like the procedure to be in some sense "optimum." We will say that our procedure is optimum if, subject to certain restrictions, it maximizes the probability of making the correct decision. A similar problem dealing with the means of several normal distributions has been studied by Paulson [1].

Let D_0 be the decision that all k variances are equal, and let D_j be the decision that D_0 is false and $\sigma_j^2 = \max(\sigma_1^2, \cdots, \sigma_k^2)$ for $j = 1, 2, \cdots, k$. Our problem now is to find a statistical procedure for selecting one of these $k + 1$ decisions.

Let $x_{i\alpha}$ denote the αth observation from the ith population, and let $\bar{x}_i = \sum_{\alpha=1}^{n} x_{i\alpha}/n$. Let $s_i^2 = \sum_{\alpha=1}^{n}(x_{i\alpha} - \bar{x})^2/(n - 1)$ denote the unbiased estimate of the variance of the ith population. We will say that Π_i has "slipped to the right" if $\sigma_1^2 = \cdots = \sigma_{i-1}^2 = \sigma_{i+1}^2 = \cdots = \sigma_k^2$ and $\sigma_i^2 = \lambda^2\sigma_1^2$ where $|\lambda| > 1$. In our first formulation of the problem we will want to find a statistical procedure which will select one of the $k + 1$ decisions $D_0, D_1, \cdots, D_k$ so that (a) when all the variances are equal, D_0 should be selected with probability $1 - \alpha$, where α is a small positive number fixed prior to the experiment.

Since the class of possible decision procedures seems to be too large to admit an optimum solution we will impose the following restrictions which seem to be reasonable: (b) the procedure should be symmetric, that is, the probability of selecting D_i when $\sigma_1^2 = \cdots = \sigma_{i-1}^2 = \sigma_{i+1}^2 = \cdots = \sigma_k^2$ and $\sigma_i^2 = \lambda^2\sigma_1^2$ should be

Received 12/15/52.

the same for all i; (c) the procedure should be invariant if all the observations are multiplied by the same positive constant; and (d) the procedure should be invariant if some constant b_i is added to all the observations in the ith population. We will now reformulate the problem as follows. We want a statistical procedure for selecting one of the $k + 1$ decisions $D_0, D_1, \cdots, D_k$ which, subject to conditions (a), (b), (c), and (d) will maximize the probability of making the correct decision when one of the populations has slipped to the right. We shall prove that the optimum solution is the following:

$$\text{if } s_M^2 \Big/ \sum_{i=1}^{k} s_i^2 > L_\alpha \text{ select } D_M;$$

$$\text{if } s_M^2 \Big/ \sum_{i=1}^{k} s_i^2 \leqq L_\alpha \text{ select } D_0,$$

where M denotes the population yielding the largest sample variance. L_α is a constant whose value is determined by restriction (a). This statistic has been suggested, on intuitive grounds, by Cochran [2], and a good tabulation of L_α for several values of α, n, and k is available [3].

2. Derivation of the optimum procedure. Since $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, s_1^2, s_2^2, \ldots, s_k^2)$ constitute a set of sufficient statistics for the unknown parameters $(m_1, m_2, \cdots, m_k, \sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2)$ there will be no loss in considering only procedures which depend on this set of statistics. We can also show that any allowable procedure will depend only on $(s_1^2, s_2^2, \cdots, s_k^2)$. Let $\delta(\bar{x}_1, \cdots, \bar{x}_k, s_1^2, \cdots, s_k^2)$ denote any allowable decision procedure. If δ depends on one or more of the $\bar{x}_i$, then for some pair of sets $(\bar{x}_1', \cdots, \bar{x}_k')$, and $(\bar{x}_1'', \cdots, \bar{x}_k'')$ we have $\delta(\bar{x}_1', \cdots, \bar{x}_k', s_1^2, \cdots, s_k^2) \neq \delta(\bar{x}_1'', \cdots, \bar{x}_k'', s_1^2, \cdots, s_k^2)$. Now we define $b_i = \bar{x}_i'' - \bar{x}_i'$, and we have the following:

$$\delta(\bar{x}_1', \cdots \bar{x}_k', s_1^2, \cdots, s_k^2) \neq \delta(\bar{x}_1' + b_1, \cdots, \bar{x}_k' + b_k, s_1^2, \cdots, s_k^2).$$

This, however, contradicts restriction (d) which states that any allowable procedure is invariant if a constant b_i is added to each observation in the ith population. Also, because of restriction (c), any allowable procedure will depend only on the $k - 1$ statistics $s_1^2/s_k^2, \cdots, s_{k-1}^2/s_k^2$. Let $u_\alpha = s_\alpha^2/s_k^2$ for $\alpha = 1, 2, \cdots, k - 1$ and $v_\alpha = \sigma_\alpha^2/\sigma_k^2$ for $\alpha = 1, 2, \cdots, k - 1$. The joint probability density of $u_1, u_2, \cdots, u_{k-1}$ will depend on the parameters $v_1, v_2, \cdots, v_{k-1}$. We will let $\bar{D}_0$ denote the decision that $v_1 = v_2 = \cdots = v_{k-1} = 1$ and $\bar{D}_i$ $(i = 1, 2, \cdots, k - 1)$ denote the decision that $v_1 = v_2 = \cdots = v_{i-1} = v_{i+1} = \cdots = v_{k-1} = 1$ and $v_i = \lambda^2$, and let $\bar{D}_k$ be the decision that $v_1 = v_2 = \cdots = v_{k-1} = 1/\lambda^2$. Since any allowable procedure for selecting one of the set $(D_0, D_1, \cdots, D_k)$ will be a function of $(u_1, u_2, \cdots, u_{k-1})$, it can be transformed into a procedure for selecting one of the set $(\bar{D}_0, \bar{D}_1, \cdots, \bar{D}_k)$ by making D_i correspond to $\bar{D}_i$ for $i = 0, 1, \cdots, k$. Because of (a) the probability that any transformed procedure will select $\bar{D}_0$ when $v_1 = v_2 = \cdots = v_{k-1} = 1$ must be $1 - \alpha$. Also, the probability that any procedure will select $\bar{D}_i$ when $\sigma^2 = \sigma_1^2 = \cdots = \sigma_{i-1}^2 = \sigma_{i+1}^2 = \cdots = \sigma_k^2$ and $\sigma_i^2 = \lambda^2\sigma^2$ must be equal to the probability that the transformed procedure will select $\bar{D}_i$ when $\bar{D}_i$ is true. This probability must be the same for all i.

The proof that the indicated solution is the optimum solution consists mainly of showing that there exists a set of nonzero a priori probabilities $p_0, p_1, \cdots, p_k$ which are functions of λ, so that when the decision procedure is transformed into a procedure for selecting one of the set $(\bar{D}_0, \bar{D}_1, \cdots, \bar{D}_k)$ it will maximize the probability of making the correct decision among the set $(\bar{D}_0, \bar{D}_1, \cdots, \bar{D}_k)$ when p_i is the a priori probability that $\bar{D}_i$ is true. This will be equivalent to showing that the procedure for selecting one of the set $(\bar{D}_0, \bar{D}_1, \cdots, \bar{D}_k)$ is a Bayes solution with respect to $p_0, p_1, \cdots, p_k$ when we introduce the loss function $W_{ij} = 1$ if $i \neq j$ and $W_{ij} = 0$ if $i = j$, where W_{ij} represents the loss in making decision $\bar{D}_i$ when $\bar{D}_j$ is true. Assuming that we have shown this, it follows that the indicated solution is the optimum solution. For suppose there exists another allowable procedure δ^* which for some λ has a greater probability of making the correct decision when some population has slipped to the right. Then δ^*, which must be a function of $u_1, u_2, \cdots, u_{k-1}$ when transformed into a procedure for selecting one of the set $(\bar{D}_0, \bar{D}_1, \cdots, \bar{D}_k)$, will have a greater probability than the indicated solution of making the correct decision when $\bar{D}_i$ is true $(i = 1, 2, \cdots, k)$ and will have, because of (a), the same probability when $\bar{D}_0$ is true. This contradicts the fact that the indicated solution is a Bayes solution relative to the nonzero probabilities $p_0(\lambda), p_1(\lambda), \cdots, p_k(\lambda)$ since its Bayes risk is larger than that of δ^*.

Since our procedure will depend on $u_1, u_2, \cdots, u_{k-1}$, we will need to find the joint probability density of these random variables. It is easy to verify that this is given by

$$g(u_1, u_2, \cdots, u_{k-1}) = \frac{\Gamma\left[\frac{k(n-1)}{2}\right]}{\left[\Gamma\left(\frac{n-1}{2}\right)\right]^k} \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{[v_1 v_2 \cdots v_{k-1}]^{(n-1)/2}\left[\sum_{\alpha=1}^{k-1} \frac{u_\alpha}{v_\alpha} + 1\right]^{k(n-1)/2}}.$$

Let $g_i = g(u_1, u_2, \cdots u_{k-1} \mid \bar{D}_i)$ be the joint probability density of $u_1, u_2, \cdots, u_{k-1}$ when $\bar{D}_i$ is true. Let $p_0, p_1, \cdots, p_k$ be a set of a priori probabilities, where p_i is the a priori probability that $\bar{D}_i$ is true. The decision procedure which maximizes the probability of making the correct decision is the Bayes solution with respect to $p_0, p_1, \cdots, p_k$ and this is known to be given by the rule: for each j, $(j = 0, 1, \cdots, k)$, select $\bar{D}_j$ for all points in the $u_1, u_2, \cdots, u_{k-1}$ space where $p_j g_j = \max(p_0 g_0, p_1 g_1, \cdots, p_k g_k)$ [4]. Consider the special a priori distribution $p_0 = (1 - kp)$, $p_1 = p_2 = \cdots = p_k = p$. We can then calculate for each j the region where $\bar{D}_j$ is selected.

As an example we will compute the region where $\bar{D}_1$ is selected. We must have $g_1 > g_j$ for $j = 2, 3, \cdots, k$, and $pg_1 > (1 - kp)g_0$.

Region where $g_1 > g_j$.

$$g_1 = C \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{(\lambda^2)^{(n-1)/2}\left[\sum_{\alpha=1}^{k-1} u_\alpha - u_1\left(1 - \frac{1}{\lambda^2}\right) + 1\right]^{k(n-1)/2}}$$

$$g_j = C \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{(\lambda^2)^{(n-1)/2}\left[\sum_{\alpha=1}^{k-1} u_\alpha - u_j\left(1 - \frac{1}{\lambda^2}\right) + 1\right]^{k(n-1)/2}}$$

where

$$C = \Gamma\left[\frac{k(n-1)}{2}\right] \Big/ \left[\Gamma\left(\frac{n-1}{2}\right)\right]^k.$$

The region where $g_1 > g_j$ is given by

$$\left[\sum_{\alpha=1}^{k-1} u_\alpha - u_j\left(1 - \frac{1}{\lambda^2}\right) + 1\right]^{k(n-1)/2} > \left[\sum_{\alpha=1}^{k-1} u_\alpha - u_1\left(1 - \frac{1}{\lambda^2}\right) + 1\right]^{k(n-1)/2}$$

or equivalently $u_1 > u_j$ for $j = 2, 3, \cdots, k-1$.

Region where $g_1 > g_k$.

$$g_k = C \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{\left(\frac{1}{\lambda^2}\right)^{(k-1)(n-1)/2}\left[\lambda^2 \sum_{\alpha=1}^{k-1} u_\alpha + 1\right]^{k(n-1)/2}} = C \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{(\lambda^2)^{(n-1)/2}\left[\sum_{\alpha=1}^{k-1} u_\alpha + \frac{1}{\lambda^2}\right]^{k(n-1)/2}}$$

Hence we must have

$$\frac{1}{\sum_{\alpha=1}^{k-1} u_\alpha - u_1\left(1 - \frac{1}{\lambda^2}\right) + 1} > \frac{1}{\sum_{\alpha=1}^{k-1} u_\alpha + \frac{1}{\lambda^2}},$$

which reduces to $(\lambda^2 - 1)(1 - u_1) < 0$. Since $\lambda^2 > 1$ we must have $u_1 > 1$.

Region where $pg_1 > (1 - kp)g_0$.

$$g_0 = C \frac{[u_1 u_2 \cdots u_{k-1}]^{(n-3)/2}}{\left[\sum_{\alpha=1}^{k-1} u_\alpha + 1\right]^{k(n-1)/2}}.$$

Thus, $pg_1 > (1 - kp)g_0$ is equivalent with

$$\frac{p}{(\lambda^2)^{(n-1)/2}\left[\sum_{\alpha=1}^{k-1} u_\alpha - u_1\left(1 - \frac{1}{\lambda^2}\right) + 1\right]^{k(n-1)/2}} > \frac{(1 - kp)}{\left[\sum_{\alpha=1}^{k-1} u_\alpha + 1\right]^{k(n-1)/2}}.$$

This may be expressed as

$$\frac{p}{(\lambda^2)^{(n-1)/2}} - (1 - kp)\left[1 - \left(1 - \frac{1}{\lambda^2}\right)\frac{u_1}{\sum_{\alpha=1}^{k-1} u_\alpha + 1}\right]^{k(n-1)/2} > 0,$$

where the left-hand side is a monotonically increasing function of $u_1/(\sum_{\alpha=1}^{k-1} u_\alpha + 1)$. Therefore this region must be of the form $u_1/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$ where $\mathfrak{L}^*$ may be written as

$$\mathfrak{L}^* = \frac{\lambda^2}{1 - \lambda^2}\left[1 - \left(\frac{p}{(1 - kp)(\lambda^2)^{(n-1)/2}}\right)^{2/k(n-1)}\right].$$

Hence we would select $\bar{D}_1$ if $u_1 > 1$, $u_1 > u_j$ for $j = 2, 3, \cdots, k - 1$, and $u_1/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$. Similarly we can show that $\bar{D}_i$ is selected for $i = 1, 2, \cdots, k - 1$, if $u_i > 1$, $u_i > u_j$ for $j = 1, 2, \cdots, i - 1, i + 1, \cdots, k - 1$, and if $u_i/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$. It remains to calculate the region where $\bar{D}_k$ is selected. This is obtained in a similar manner and the result is stated without derivation. Accept $\bar{D}_k$ when $u_j < 1$ $(j = 1, 2, \cdots, k - 1)$ and $1/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$, where $\mathfrak{L}^*$ is the same constant as above.

Hence the Bayes solution with respect to $(1 - kp), p, p, \cdots, p$, is the following: for $1 \leqq j \leqq k - 1$ select $\bar{D}_j$ if $u_j > 1$, and $u_j > \max(u_1, u_2, \cdots, u_{j-1}, u_{j+1}, \cdots, u_{k-1})$, and $u_j/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$. Select $\bar{D}_k$ if $u_j < 1$ for $j = 1, 2, \cdots, k - 1$ and $1/(\sum_{\alpha=1}^{k-1} u_\alpha + 1) > \mathfrak{L}^*$. Otherwise select $\bar{D}_0$.

The existence of a priori probabilities will now be shown for which the above procedure, with $\mathfrak{L}^*$ replaced by the fixed L_α determined by condition (a), is a Bayes solution. Let us define the function

$$F(p) = \frac{p}{(\lambda^2)^{(n-1)/2}} - (1 - kp)\left[1 - L_\alpha\left(1 - \frac{1}{\lambda^2}\right)\right]^{k(n-1)/2}.$$

This is a continuous function of p with $F(0) < 0$, and $F(1/k) > 0$. Hence, there exists a p^* with $0 < p^* < 1/k$ which is a function of λ^2 so that $F(p^*) = 0$. To get the Bayes solution relative to $(1 - kp^*, p^*, \cdots, p^*)$ we merely replace $\mathfrak{L}^*$ by L_α.

We now substitute $u_\alpha = s_\alpha^2/s_k^2$ and the Bayes solution relative to $(1 - kp^*, p^*, \cdots, p^*)$ reduces to the following when $\bar{D}_i$ is made to correspond to D_i:

$$\text{if } s_M^2 \Big/ \sum_{\alpha=1}^{k} s_\alpha^2 > L_\alpha \text{ select } D_M,$$

$$\text{if } s_M^2 \Big/ \sum_{\alpha=1}^{k} s_\alpha^2 \leqq L_\alpha \text{ select } D_0,$$

where $s_M^2 = \max(s_1^2, s_2^2, \cdots, s_k^2)$. Since this is an allowable procedure we have proved it is an optimum one.

REFERENCES

[1] EDWARD PAULSON, "An optimum solution to the K sample slippage problem for the normal distribution," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 610–616.

[2] W. G. Cochran, "The distribution of the largest of a set of estimated variances as a fraction of their total," *Ann. Eugenics*, Vol. 11 (1941), pp. 47–52.
[3] Eisenhart, Hastay, and Wallis (Editors), *Selected Techniques of Statistical Analysis*, McGraw-Hill Book Co., 1947, pp. 390–391.
[4] A. Wald, *Statistical Decision Functions*, John Wiley and Sons, 1950, pp. 123–130.

AN EXTENSION OF THE BUFFON NEEDLE PROBLEM

By Nathan Mantel

Biometrics Section, National Cancer Institute[1]
Bethesda, Maryland

1. Introduction. An empirical determination of the value of π can be made from the relationship[2]

$$E = 2L_1L_2/(\pi A), \tag{1}$$

where E is the expected number of intersections of a group of line segments of total length L_1 with a group of line segments of total length L_2, both groups being distributed over an area A. This relationship applies under the following conditions.

(i) The arrangement of the two groups of line segments on the area A must be independent of each other, but the individual line segments of a group may have a systematic arrangement relative to each other.

(ii) The arrangement of at least one of the two groups of line segments on the area A must be at random. The randomness must be such that the probability of a specified point on a line segment falling into a sub-area of A is proportional to its area and the segment may assume any angle relative to some base line with equal probability.

Two applications of this relationship to the estimation of π are considered below.

2. The Buffon needle problem using a parallel line system. Consider an area A on which is superimposed a series of equally spaced parallel lines (without loss of generality we shall take the common distance between them to be unity), on which a straight line of length $L \leqq 1$ is allowed to fall at random. At each fall the line must either intersect the series of parallel lines only once, or not at all. Thus the expected number of intersections, E, is the probability, P, of an intersection occurring at a fall. And since for this system the total length of the

Received 6/12/52, revised 12/22/52.

[1] National Institutes of Health, Public Health Service, Department of Health, Education and Welfare.

[2] This relationship is developed in passing by Cornfield and Chalkley, "A problem in geometric probability," *J. Wash. Acad. Sci.*, July, 1951.