A CLASSIFICATION PROBLEM INVOLVING MULTINOMIALS[1]

By Oscar Wesler

University of Michigan

1. Introduction. The problem of the k-faced die. There are several distinct classification problems involving multinomials, each known as the problem of the k-faced die. For example, the problem may be to decide whether the die is loaded, and indeed there are several versions of this. One, say, in which the loading is specified; another, in which the unknown loading is estimated from one sample, and then a decision is made as to whether a second sample came from the loaded die or an honest one [1]. Or the problem may be to determine which of the k faces carries a known extra load [2]. Although distinct, all these problems are somewhat related and have certain features more or less in common. The version we shall treat in this paper is still another and a general one of its kind. It will be convenient for our purposes to state it formally as a fixed sample size two-decision statistical problem considered within the framework of composite hypotheses. We shall use the notation and terminology of [3].

Let the space Ω of nature's pure strategies consist of two subspaces Ω_1 and Ω_2, Ω_1 consisting of the $k!$ states got by permuting a known probability distribution $p = (p_1, p_2, \cdots, p_k)$ over the faces $1, 2, \cdots, k$ of a k-faced die, Ω_2 consisting similarly of the $k!$ states arising from a known distribution $q = (q_1, q_2, \cdots, q_k)$. We assume the p_i and q_i strictly positive, and shall further assume, without loss of generality, the vectors p and q written so that $p_1 \geqq p_2 \geqq \cdots \geqq p_k$, and $q_1 \geqq q_2 \geqq \cdots \geqq q_k$. The statistician wishes to make one decision if $\omega \,\varepsilon\, \Omega_1$ (the null hypothesis), and another decision if $\omega \,\varepsilon\, \Omega_2$ (the alternative hypothesis), the decision to be made on the basis of a sample of N observations $x = (x_1, \cdots, x_N)$, or rather on the basis of the sufficient statistic $r = (r_1, \cdots, r_k)$ representing the number of times each of the k faces appears. Let ϕ be a randomized statistical decision procedure such that if r is observed, the null hypothesis is accepted with probability $\phi(r)$ and the alternative hypothesis is accepted with probability $1 - \phi(r)$. Let $\bar{\alpha}$ and $\bar{\beta}$, the probabilities of the two kinds of errors, be the two functions given by

$$\bar{\alpha}(\omega \mid \phi) = 1 - \sum_r \phi(r) P(r \mid \omega), \qquad \omega \,\varepsilon\, \Omega_1$$

$$\bar{\beta}(\omega \mid \phi) = \sum_r \phi(r) P(r \mid \omega), \qquad \omega \,\varepsilon\, \Omega_2$$

Received May 12, 1958.

[1] Part of this research was carried out under contract with the Office of Naval Research while the author was at Stanford University, and completed at the University of Michigan with the partial support of the United States Army Signal Air Defense Engineering Agency, Contract No. DA-36-039 SC-64627.

and let

$$\alpha(\phi) = \max_{\omega \varepsilon \Omega_1} \bar{\alpha}(\omega \mid \phi),$$

$$\beta(\phi) = \max_{\omega \varepsilon \Omega_2} \bar{\beta}(\omega \mid \phi).$$

Then the problem we wish to consider is: Of all procedures ϕ satisfying the condition $\alpha(\phi) = \alpha_0$, to find that procedure ϕ^* which minimizes $\beta(\phi)$.

In this paper a game-theoretic minimax method is used to find an entire class of these desired procedures ϕ^*, optimal in this extended Neyman-Pearson sense. A simplification of the result is then given for the asymptotic case of large N. Finally, for illustrative purposes, a look is taken at the special binomial case of $k = 2$, a case deserving of special attention in its own right. We shall begin with a brief description of the minimax approach, referring the reader to Section 7.7 of [3] for detailed proofs.

2. The minimax method. This method of solving the problem can be understood in terms of a simple geometric picture. Consider the α, β-set in Euclidean 2-space given by $S = \{(\alpha(\phi), \beta(\phi)) : \phi \varepsilon \Phi\}$, where Φ is the class of all randomized strategies ϕ. S is a point set in the unit square of the first quadrant and contains the diagonal on which $\alpha + \beta = 1$. Although S is not necessarily convex (unlike the case of a simple hypothesis against a simple alternative) it can be shown that the subset T of S lying on or below the above-mentioned diagonal *is* convex, and that moreover, in the present case, the points on the lower boundary of T, a strictly decreasing convex curve joining the point (0, 1) to the point (1, 0), belong to T. Clearly, it is these points on the lower boundary of T which provide the solution to the statistical problem posed above: Given α_0, we are interested in that procedure ϕ^* which determines the point $(\alpha(\phi^*), \beta(\phi^*))$ on the lower boundary of T whose first coordinate is equal to α_0. Our problem, then, is to characterize the class of test procedures ϕ^* which sweep out the entire lower boundary of T.

Now let G_u be the statistical game that arises when the losses are taken to be 1 for an error of the first kind, a constant $u > 0$ for an error of the second kind, and zero otherwise. It is readily seen from our picture that a minimax solution ϕ_u^* for G_u determines that point on the lower boundary of T for which $\alpha(\phi_u^*) = u\beta(\phi_u^*)$. *It follows at once, therefore, that, by varying u between 0 and ∞, the class of corresponding minimax solutions ϕ_u^* will generate the desired lower boundary of T.*

3. The general solution. The minimax method requires that we solve the statistical game G_u for arbitrary $u > 0$. The problem is clearly invariant under the symmetric group of permutations on k letters. We shall therefore set about finding the class of (unique) invariant minimax procedures which will sweep out the lower boundary of T.[2] For fixed g, with $0 < g < 1$, let λ_g be the a priori probability distribution over Ω which selects Ω_1 with probability g, Ω_2 with prob-

[2] For the specific problem under consideration, it is very easily shown that only invariant procedures need be considered in minimizing $\beta(\phi)$. In more general situations, see [5].

ability $1 - g$, and which is uniform within both Ω_1 and Ω_2, i.e. $\lambda_g(\omega) = g/k!$ for $\omega \varepsilon \Omega_1$ and $\lambda_g(\omega) = (1 - g) / k!$ for $\omega \varepsilon \Omega_2$. Define the distribution function $p(r)$ over the set of all possible outcomes $r = (r_1, \cdots, r_k)$ of the N observations by the formula,

$$p(r) = \frac{1}{k!} \sum_{(i_1, \cdots, i_k)} \frac{N!}{r_1 ! \cdots r_k !} p_{i_1}^{r_1} p_{i_2}^{r_2} \cdots p_{i_k}^{r_k},$$

where the summation is taken over all permutations $(i_1, \cdots, i_k)$ of the indices $1, \cdots, k$. In other words, $p(r)$ is the average of all the multinomial distributions over the r's arising from the probability vector $p = (p_1, \cdots, p_k)$. $p(r)$ is clearly symmetric in r, that is, $p(r) = p(r_1, r_2, \cdots, r_k) = p(r_{i_1}, r_{i_2}, \cdots, r_{i_k})$. Similarly, define the symmetric distribution $q(r)$ as the average of the multinomial distributions determined by the probability vector $q = (q_1, \cdots q_k)$. The Bayes procedure against λ_g is readily computed in the usual way to be:

$$\phi_{c,t}(r) = \begin{cases} 1 & \text{if } \dfrac{q(r)}{p(r)} < \dfrac{g}{(1 - g)u} = c(g, u) = c \\ 0 & \text{if } \dfrac{q(r)}{p(r)} > c \\ t & \text{if } \dfrac{q(r)}{p(r)} = c \end{cases} \tag{1}$$

where t is any fixed number satisfying $0 \leqq t \leqq 1$. We note that the procedure $\phi_{c,t}$ is a symmetric function of r and depends on g and u only through c. Further, for any choice of t and c, we see that any pair of numbers g_0, u_0 leading to c will yield the same procedure $\phi_{c,t}$, which will be Bayes against λ_{g_0} in the game G_{u_0}. Moreover, c varies between 0 and ∞ as g and u vary respectively between 0 and 1, and 0 and ∞. Varying c between 0 and ∞, and t between 0 and 1, the symmetric partitions of the set of all possible r's into acceptance and rejection regions defined by the inequalities in (1) are clearly such that the acceptance regions vary monotonically from the empty set to the whole space, so that the corresponding α's vary from 1 to 0. By an obvious continuity argument, therefore, given any α_0, with $0 \leqq \alpha_0 \leqq 1$, we can by suitably choosing t find a c such that $\alpha(\phi_{c,t}) = \alpha_0$. Now, consider any one such procedure $\phi_{c,t}$. From the symmetry of $\phi_{c,t}$ it follows that its associated risk function $\rho(\omega, \phi_{c,t})$ is constant over Ω_1 and constant over Ω_2, equal to $\alpha(\phi_{c,t})$ and $u\beta(\phi_{c,t})$ respectively. Selecting u_0 so that $\alpha(\phi_{c,t}) = u_0\beta(\phi_{c,t})$ and then g_0 so that $g_0/(1 - g_0)u_0 = c$, we see that the procedure $\phi_{c,t}$ is Bayes against λ_{g_0} and has a constant risk function over Ω, hence $\phi_{c,t}$ is a minimax procedure for the game G_{u_0}, and determines the point $(\alpha(\phi_{c,t}), \beta(\phi_{c,t}))$ on the lower boundary of T. *By continuity, therefore, the class of invariant procedures $\phi_{c,t}$ with $0 \leqq c < \infty$, $0 \leqq t \leqq 1$, determines the entire lower boundary of T and provides the extended Neyman-Pearson solution to our classification problem.* Finally, since any minimax invariant procedure ϕ^* for a game G_{u_0} must be Bayes against the corresponding λ_{g_0}, and since the only such invariant Bayes procedures are of the form $\phi_{c,t}$, this proves the uniqueness of the class of invariant minimax procedures.

It might perhaps be useful for certain purposes to give the class of invariant partitions or procedures the following somewhat simpler geometric description. Let Ξ be the fundamental probability simplex in k-space, consisting of all vectors $\xi = (\xi_1, \cdots, \xi_k)$ with $\xi_i \geqq 0$, $\sum_{i=1}^k \xi_i = 1$, and let $\delta_i = r_i/N$, so that $\delta = (\delta_1, \cdots, \delta_k) \varepsilon \Xi$. By the symmetry of the problem and its solution, we shall assume without loss of generality that $r_1 \geqq \cdots \geqq r_k$, so that the observed probability vector δ as well as the two given vectors p and q all belong to the subsimplex Ξ' of Ξ defined by $\xi_1 \geqq \xi_2 \geqq \cdots \geqq \xi_k$. Then, given α_0, the desired optimal procedure $\phi_{c,t}$ may be characterized in the first place by the simple partition of Ξ' into two regions of δ's (the acceptance region and rejection region, with possible mixing on the boundary) given by the defining inequalities specified in (1); the symmetric images of this partition under all $k!$ permutations then give the corresponding symmetric partition of Ξ which constitutes the whole procedure $\phi_{c,t}$.

4. The case of large N. The kaleidoscopic procedures. For large N, an approximation suitable for most purposes may be given which materially simplifies the class of optimal partitions.

Confining our attention to Ξ', the inequality $q(r)/p(r) < c$ in (1) may be written, on factoring out the largest term from both numerator and denominator, as

$$\left[\left(\frac{q_1}{p_1}\right)^{\delta_1}\left(\frac{q_2}{p_2}\right)^{\delta_2}\cdots\left(\frac{q_k}{p_k}\right)^{\delta_k}\right]^N \frac{1+\Sigma_1}{1+\Sigma_2} < c$$

where, for large N, hence only approximately for large $r_1, r_2, \cdots, r_k$, both Σ_1 and Σ_2 may be taken as zero, or as certain positive integers less than $k!$, depending on the number of equalities existing among the given components q_i of q and among the p_i of p. In any event, taking logarithms, we may then regard the optimal partitions of Ξ' as approximately determined by the one parameter family of hyperplanes

$$\sum_{i=1}^{k} \delta_i \log \frac{q_i}{p_i} = b \qquad (b \text{ an arbitrary constant}). \tag{2}$$

That is, for arbitrary b and t, $0 \leqq t \leqq 1$, the optimal procedure $\phi_{b,t}$ amounts to placing the points δ of Ξ' on the side of the hyperplane where

$$\sum_{i=1}^{k} \delta_i \log \frac{q_i}{p_i} < b$$

into the acceptance region, the points on the other side into the rejection region, those on the hyperplane mixed in the proportion t, $1 - t$, and then taking the symmetric images of this partition to get the symmetric partition of Ξ as a whole. In view of our geometric description, we shall call these approximating $\phi_{b,t}$ the *kaleidoscopic procedures*.

To see at once the reason for this rather striking language, we invite the reader to consider the case of $k = 3$, and to draw a few simple figures, with the points δ or $\xi = (\xi_1, \xi_2, \xi_3)$ of Ξ represented in barycentric coordinates by a point in an equilateral triangle with unit altitude, the distance from the point to the three sides being the values ξ_1, ξ_2 and ξ_3. By selecting different pairs of points for p and q so as to get separating lines with different slopes, and by varying b so as to translate these lines in a parallel fashion, the reason for the term *kaleidoscopic* will appear before his eyes.

In general, by a suitable choice of b and of the mixing factor t, the symmetric or kaleidoscopic test can be made to give us any desired α_0 with approximately the minimum β.

5. The case of $k = 2$: The binomial case. We specialize the previous discussion to $k = 2$ as deserving of special attention, and also to provide a simple illustration of the discussion and results. We refer the reader to problem 7.7.3 of [3] and to pages 75–76 of [4] for a particular example.

We are now in the binomial case and shall speak of coins rather than dice. Two probability distributions, $p = (p_1, p_2)$ and $q = (q_1, q_2)$, are given. Under Ω_1 the coin falls heads with probability either p_1 or $p_2 = 1 - p_1$, under Ω_2 heads with probability either q_1 or $q_2 = 1 - q_1$. Assume things written so that $1 > p_1 > q_1 \geqq 1/2$. Let $r = (r_1, N - r_1)$ represent, respectively, the number of heads and tails appearing among the N tosses of the coin. Then $p(r)$, given by

$$p(r) = \binom{N}{r_1} \frac{p_1^{r_1} p_2^{N-r_1} + p_2^{r_1} p_1^{N-r_1}}{2},$$

and $q(r)$ given similarly, are symmetric bimodal distributions over the number r_1 of heads, $r_1 = 0, 1, \cdots, N$. The likelihood ratio $q(r)/p(r)$ which determines the optimal procedures in (1) may be written

$$\frac{q(r)}{p(r)} = \frac{q_1^{r_1} q_2^{N-r_1} + q_2^{r_1} q_1^{N-r_1}}{p_1^{r_1} p_2^{N-r_1} + p_2^{r_1} p_1^{N-r_1}} = \left(\frac{q_1 q_2}{p_1 p_2}\right)^{N-r_1} \left(\frac{q_1^{2r_1-N} + q_2^{2r_1-N}}{p_1^{2r_1-N} + p_2^{2r_1-N}}\right). \tag{3}$$

Without bothering to convert r to δ, the kaleidoscopic procedures given by (2) are immediately seen to be two-sided symmetric tests of the following form: Take action 1 (decide for p) if $r_1 < j$ or $r_1 > N - j$, action 2 (decide for q) if $j < r_1 < N - j$, and mix actions 1 and 2 in the proportion t, $1 - t$ if $r_1 = j$ or $N - j$, where j is allowed to run over the integers $0, 1, \cdots, [N/2]$ and t is any number satisfying $0 \leqq t \leqq 1$. Denote this two-sided symmetric test by (j, t). Then the kaleidoscopic procedures are given by the set of all pairs (j, t) with $j = 0, 1, \cdots, [N/2]$, $0 \leqq t \leqq 1$. Note that if $p_1 < q_1$ the procedures are of course the same, except that the actions taken are reversed.

The interesting feature here, which we shall prove, is that the optimal procedures of (1) and the (j, t) of (2) are the same, in exact agreement *for every* N, so that there is no question here of approximations or of loss of power in the approximation. To see this, let us agree because of symmetry to consider the likelihood ratio (3) for only those outcomes $r = (r_1, N - r_1)$ for which $2r_1 \geqq N$.

Then the assertion amounts to showing that the inequality $q(r)/p(r) < c$ for any one such r implies the same inequality for all r with a larger r_1. Suppose, therefore, $q(r)/p(r) < c$. Since $q_1q_2 > p_1p_2$, the first factor in the factored form of (3) decreases with increasing r_1. If we can show that the second factor decreases as well with increasing r_1, the assertion will be proved. We require the following simple lemma.

LEMMA: *The function* $y = [x^{m+1} + (1 - x)^{m+1}]/[x^m + (1 - x)^m]$ *considered on the unit interval* $0 < x < 1$, *with* m *an arbitrary non-negative integer, is decreasing for* $0 < x \leqq 1/2$ *and (by symmetry) increasing for* $1/2 \leqq x < 1$.

PROOF: Taking derivatives, a straightforward calculation will show that y' has the same sign as $[x^{2m} - (1 - x)^{2m}] + mx^{m-1}(1 - x)^{m-1}[x^2 - (1 - x)^2]$. Clearly, $y' \leqq 0$ for $0 < x \leqq 1/2$, hence the lemma.

Applying the lemma for successive values of m and multiplying functions together, there follows at once the more general fact that the same decreasing property over $0 < x \leqq 1/2$ and increasing property over $1/2 \leqq x < 1$ is exhibited by the function

$$y = \frac{x^n + (1 - x)^n}{x^m + (1 - x)^m}$$

for arbitrary integers $n > m \geqq 0$. From our assumption of $1/2 \leqq q_1 < p_1 < 1$ there follows the inequality

$$\frac{q_1^n + q_2^n}{q_1^m + q_2^m} < \frac{p_1^n + p_2^n}{p_1^m + p_2^m}$$

or, equivalently, the inequality

$$\frac{q_1^n + q_2^n}{p_1^n + p_2^n} < \frac{q_1^m + q_2^m}{p_1^m + p_2^m}$$

for arbitrary integers $n > m \geqq 0$. But this says that the second factor of (3) decreases with increasing r_1. The asserted agreement between the optimal and kaleidoscopic procedures is thus proved.

It should be pointed out that in practice, given α_0, it becomes a simple matter in this case, using tables of the binomial distribution $P_N(r_1 \mid p_1)$, to calculate the desired test (j, t) from the condition $\alpha(j, t) = \alpha_0$. The minimized β, namely $\beta(j, t)$, is then read off directly from the tables of $P_N(r_1 \mid q_1)$.

REFERENCES

[1] HERMAN CHERNOFF, "A classification problem," *Technical Report 33*, Applied Mathematics and Statistics Laboratory, Stanford University, 1956.

[2] S. G. ALLEN, JR., "A procedure for determining the loaded face of a die," *Technical Report 13*, Applied Mathematics and Statistics Laboratory, Stanford University, 1952.

[3] DAVID BLACKWELL AND M. A. GIRSHICK, *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954, pp. 200–207 in particular.

[4] OSCAR WESLER, *Solutions to Problems in Blackwell and Girshick's Theory of Games and Statistical Decisions*, John Wiley and Sons, New York, 1954.

[5] OSCAR WESLER, "Invariance theory and a modified minimax principle," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 1–20.