ON THE CLASSIFICATION OF OBSERVATION DATA INTO DISTINCT GROUPS

By R. v. Mises

Harvard University

Introduction. In scholastic examinations as well as in the examination of industrial products the following probability problem arises. The individuals of a certain population are successively subjected to trials each of which leads to a definite score x (one real number or a group of m real numbers). Each individual is supposed to belong to one of n classes. These classes are characterized by n probability densities $p_1(x)$, $p_2(x)$, $\cdots$ $p_n(x)$. One has to decide on the basis of the observed value x to which class the respective individual belongs and one wishes to make this decision with the smallest possible risk of failure.

For example, let us consider an examination where the three grades A, B, C are attributed on the basis of a simple score x (case $m = 1$, $n = 3$). It may be assumed that an individual of the class A has a mean expected value of x equal to $\vartheta_1 = 75$ and a normal distribution with the standard deviation $\sigma_1 = 4/\sqrt{2}$. The analogous values for the classes B and C may be $\vartheta_2 = 50$, $\sigma_2 = 8/\sqrt{2}$ and $\vartheta_3 = 25$, $\sigma_3 = 12/\sqrt{2}$. In this case, the solution developed in the present paper allows the conclusion that the best way of grading would be to attribute the grade A to scores x beyond 70.0, the grade C to scores below 40.0 and B to the rest. The corresponding error risk will be 3.9% or the success rate 0.961.

There exists, of course, one case where the solution is trivial. If the probability densities $p_r(x)$ are limited to n non-overlapping regions R_r (with $p_r = 0$ at points outside R_r) an obvious decision can be made without any risk of failure. An assumption of this kind underlies the usual procedure of grading. If, in the foregoing example, an individual of class A is supposed to have at any rate a score beyond 60 and a class C individual less than 40, it is obvious how the grades should be attributed without incurring any risk. It seems, however, that in many problems the assumption of normal distributions or some other kind of overlapping distributions is more appropriate. Then, the probability problem has to be solved.

The solution submitted in the present paper is derived from the simplest principles of calculus of probability without any arbitrary assumption or hypothesis. If n equals 2, the problem can also be considered as a problem of testing a simple statistical hypothesis with a two-valued parameter.[1] It has been shown in an earlier paper[2] that under this restriction success rates higher than 50% are obtainable.

[1] See A. Wald, *Annals of Math. Stat.*, Vol. 15 (1944), p. 145. Here, both $p_1(x)$ and $p_2(x)$ are supposed to be normal distributions with the same covariance matrix. The problem treated by Wald is different from the one considered in the present paper since in Wald's paper the parameters of the two multivariate normal distributions are assumed to be unknown.

[2] R. v. Mises, *Annals of Math. Stat.*, Vol. 14 (1943), p. 238.

1. Statement of the problem. For each of n classes of individuals a probability density $p_\nu(x)$, $\nu = 1, 2, \cdots n$, is given. We subdivide the m-dimensional x-space into n regions $R_1, R_2, \cdots R_n$ and assign the region R_ν to the νth class. The probability, for an individual of this class, to have its x-value falling in R_ν is

$$P_\nu = \int_{(R_\nu)} p_\nu(x)\, dX, \qquad \nu = 1, 2, \cdots n \tag{1}$$

where dX denotes the element of the x-space ($dX = dx$ in the case $m = 1$). In the N first trials of the indefinite sequence of trials, N_ν individuals that belong to the νth class will be tested. Out of these only those individuals whose x-value falls in R_ν will be ascribed to the νth class. Their number according to the definition of probability, equals $N_\nu(P_\nu + \epsilon_\nu)$ where ϵ_ν tends towards zero as N_ν goes to infinity. The total number of correct decisions during the N first trials is therefore

$$N_1(P_1 + \epsilon_1) + N_2(P_2 + \epsilon_2) \cdots N_n(P_n + \epsilon_n) \tag{2}$$

and the relative number is

$$\frac{N_1}{N}(P_1 + \epsilon_1) + \frac{N_2}{N}(P_2 + \epsilon_2) + \cdots \frac{N_n}{N}(P_n + \epsilon_n). \tag{2'}$$

If N increases indefinitely a part of the N_ν must become infinite. For these classes, ϵ_ν converges toward zero. For the other classes N_ν/N diminishes to zero. Thus, the relative number of right decisions converges towards

$$\frac{1}{N}(N_1P_1 + N_2P_2 + \cdots N_nP_n). \tag{3}$$

The N_ν are unknown. Every one of these unknowns can take each value from zero to N. If P_μ is the smallest P_ν, the most unfavorable case, where the expression (3) has its smallest value, will occur with $N_\mu = N$, all other N_ν being zero. This value is obviously P_μ. Thus it is seen that the frequency of correct assignments is at least equal to the smallest P_ν which may be written as $P_{\min}$. The greatest risk of making a false decision is $1 - P_{\min}$.

Now the problem to be solved in the present paper can be stated as follows: *For n given densities $p_\nu(x)$, find the subdivision of the x-space into n regions R_ν that gives to the smallest of the expressions P_ν defined in* (1) *its possibly greatest value.*

This problem has the type of a continuous variation problem with the integrals in question bounded within the limits zero to one. We may, therefore, assume that under reasonable restrictions for $p_\nu(x)$ a solution exists. Uniqueness of the solution cannot be expected in general. It seems very difficult to establish the conditions for unicity in other than the most simple cases. Existence of more than one solution would mean that each of them is an optimum with respect to infinitesimal modifications of the boundaries.

2. General solution. A simple problem of variation is considered as solved in principle when the nature of the extremals is known. In our case of a so-called minimax problem, where the minimum of n quantities is maximized, an additional relation between the n integrals is required. Both can easily be found in the actual case.

Let us first consider a partition of the x-space into n regions with not all P_ν being equal. The smallest P_ν will be called $P_{\min}$ and the smallest but one P^*. Among the k regions for which $P_\nu = P_{\min}$ there will be at least one, say, R_α that has a common border with a region R_β whose P-value is greater, so that $P_\beta \geqq P^*$. Now modify the boundary between R_α and R_β in such a way that the space covered by R_α is increased and that of R_β decreased. According to (1) the new values of P_α and P_β will be

$$P'_\alpha = P_\alpha + \Delta, \qquad P'_\beta = P_\beta - \Delta' \tag{4}$$

with both Δ and Δ' positive. The two quantities Δ and Δ' are not independent of one another, but they can be chosen both smaller than any given positive number ϵ. Therefore, the condition

$$P'_\alpha = P_\alpha + \Delta < P_\beta - \Delta' = P'_\beta \tag{5}$$

can be fulfilled. All other P_ν-values remain unchanged.

In the case $k = 1$, that is, if only one region R_ν had originally the minimum P-value, the modified system has a greater minimum P, which equals either $P_\alpha + \Delta$ or P^*. If $k > 1$ the new system has the same minimum P as the original one, but its k-value is diminished by one. If we repeat the same procedure $(k - 1)$ times we obtain a system of regions with one single P_ν having the minimum P-value and the next step leads to a partition of the x-space into n regions with a smallest P-value that is greater than the original $P_{\min}$. Thus it is seen that no partition with unequal P_ν-values can solve our problem.

Secondly, if $m > 1$, consider a system of n regions with $P = P_1 = P_2 = \cdots = P_n$. Take two points, x and y, on the border of any two neighboring regions R_ν and R_μ. An infinitesimal variation of the boundary would consist of adding to R_ν in the neighborhood of the point x a space element δS subtracting it from R_μ and, at the same time, adding to R_μ in the vicinity of y an element $\delta S'$ subtracting it from R_ν. Then, according to (1), the new values of P_ν and P_μ will be

$$\begin{aligned} P'_\nu &= P + p_\nu(x)\delta S - p_\nu(y)\delta S' \\ P'_\mu &= P - p_\mu(x)\delta S + p_\mu(y)\delta S'. \end{aligned} \tag{6}$$

Introducing $\Delta_\nu = P'_\nu - P$ and $\Delta_\mu = P'_\mu - P$, these equations solved for δS and $\delta S'$ give

$$\delta S = \frac{p_\nu(y)\Delta_\mu + p_\mu(y)\Delta_\nu}{D}, \qquad \delta S' = \frac{p_\nu(x)\Delta_\mu + p_\mu(x)\Delta_\nu}{D} \tag{7}$$

where

$$D = p_\nu(x)p_\mu(y) - p_\mu(x)p_\nu(y). \tag{7'}$$

If the determinant D is positive, we find two positive quantities δS and $\delta S'$ for any pair of positive Δ_μ and Δ_ν. If D is negative the same is true when x and y are interchanged. In both cases, that is, with $D \neq 0$, the original partition is replaced by a new system of regions in which only two regions, R_ν and R_μ, have increased P-values, while (if $n > 2$) still $P_{\min} = P$. If to this system the procedure as described in the foregoing is applied, a final partition with a greater minimum value of P can be derived. The conclusion is that no solution of our problem can include a boundary on which the determinant D is different from zero for any two points x and y. On the other hand, it is seen that $D = 0$ means that the ratio $p_\nu(x):p_\mu(x)$ has a constant value along the border. Thus the result is reached:

The partition of the x-space that solves our problem is characterized by two properties: (1) *for all n regions* R_ν *the value of* P_ν *is the same;* (2) *along the border between* R_ν *and* R_μ *the ratio* $p_\nu(x)/p_\mu(x)$ *is constant.*

In the one-dimensional case ($m = 1$) only the first of these two statements is relevant. In any case, the success rate, that is, the guaranteed ratio of correct decisions, equals the common value of all P_ν.

3. Illustrations. (a) One-dimensional case. Upon introducing the cumulative distribution functions

$$F_\nu(x) = \int_{-\infty}^{x} p_\nu(z)dz \tag{8}$$

the conditions $P_1 = P_2 = \cdots P_n$ take the form

$$F_1(x_1) = F_2(x_2) - F_2(x_1) = \cdots = F_{n-1}(x_{n-1}) - F_{n-1}(x_{n-2}) = 1 - F_n(x_{n-1}) \tag{9}$$

where $x_1, x_2, \cdots x_{n-1}$ determine the n intervals on the both-sides infinite x-axis. If all density functions have the same form except for an affine transformation, one has

$$F_\nu(x) = F[h_\nu(x - \vartheta_\nu)], \qquad \nu = 1, 2, \cdots n \tag{10}$$

Let us assume, for instance, that scores between 0 and 100 are attributed to three types of individuals. The first type may have an even chance to obtain a score between 0 and 50, the second between 40 and 80 and the third between 70 and 100. Here

$$F_\nu(x) = \tfrac{1}{2} + (x - \vartheta_\nu)p_\nu, \qquad |x - \vartheta_\nu| \leqq \frac{1}{2p_\nu} \tag{11}$$

with $\vartheta_\nu = 25, 60, 85$ and $p_\nu = \frac{1}{50}, \frac{1}{40}, \frac{1}{30}$. The conditions (9) supply

$$\frac{1}{2} + \frac{x_1 - 25}{50} = \frac{1}{40}(x_2 - x_1) = \frac{1}{2} - \frac{x_2 - 85}{30} \tag{12}$$

and this, solved for x_1, x_2 gives $x_1 = 41\frac{2}{3}$, $x_2 = 75$ while the three expressions (12) take the value 0.833. Therefore, in attributing all scores below $41\frac{2}{3}$ to the first class and all scores beyond 75 to the third one is safe to make under no circumstances more than $\frac{1}{6}$ incorrect decisions in the long run.

In the example quoted in the introduction one has

$$p_\nu(x) = \frac{1}{\sigma_\nu\sqrt{2\pi}}\, e^{(x-\vartheta_\nu)^2/2\sigma_\nu^2} \tag{13}$$

with $\vartheta_\nu = 75, 50, 25$ and $\sigma_\nu^2 = 8, 32, 72$. If $\Phi(x)$ denotes the integral

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2}\, dz$$

the conditions (9) become

$$1 + \Phi\left(\frac{x_1 - 25}{12}\right) = \Phi\left(\frac{x_2 - 50}{8}\right) - \Phi\left(\frac{x_1 - 50}{8}\right) = 1 - \Phi\left(\frac{x_2 - 75}{4}\right). \tag{14}$$

The first and last expression equated lead to $x_1 + 3x_2 = 250$. The complete solution can be found with the help of tables for Φ. It is $x_1 = 29.9920$, $x_2 = 70.0027$ with the common value twice 0.961 for the three expressions (14). Hence the result as quoted in the introduction.

Let us now take up the case of six normal distributions with equidistant mean values $\vartheta = \pm a, \pm 3a, \pm 5a$ and one and the same variance σ^2. Then, because of symmetry, two equations only have to be fulfilled:

$$1 + \Phi\left(\frac{x_1 + 5a}{\sigma\sqrt{2}}\right) = \Phi\left(\frac{x_2 + 3a}{\sigma\sqrt{2}}\right) - \Phi\left(\frac{x_1 + 3a}{\sigma\sqrt{2}}\right) = \Phi\left(\frac{a}{\sigma\sqrt{2}}\right) - \Phi\left(\frac{x_2 + a}{\sigma\sqrt{2}}\right)$$

For $\sigma^2/a^2 = 0.32$, the numerical solution gives

$$x_1 = -4.160a, \qquad x_2 = -2.062a.$$

The success rate, i.e. half the common value of the above expressions is 0.931. The six intervals extend from $-\infty$ to x_1, from x_1 to x_2, from x_2 to 0, from 0 to $-x_2$, from $-x_2$ to $-x_1$, and from $-x_1$ to ∞.

(b) *Case of more than one dimension.* Let us assume that two classes A and B have uniform distributions extending over volumes $V_1 = 1/p_1$ and $V_2 = 1/p_2$ respectively. If the two regions have a common part of volume V each surface within the common space fulfills the condition $p_1/p_2 =$ constant. Thus, the two regions R_1 and R_2 are not uniquely determined but subject to one condition only which determines the optimum success rate. If κV is cut out from V_1 and $(1 - \kappa)V$ from V_2, the relation must be fulfilled:

$$1 - p_1 V\kappa = 1 - p_2 V(1 - \kappa), \qquad i.e.\ \kappa = \frac{p_2 V}{p_1 + p_2}$$

and the success rate is

$$S = 1 - p_1 V\kappa = 1 - \frac{p_1 p_2 V}{p_1 + p_2} = 1 - p_2 V(1 - \kappa).$$

If three classes A, B, and C are considered with the densities $p_1 = 1/V_1$, $p_2 = 1/V_2$, $p_3 = 1/V_3$ and the first two regions have a space of volume V in common, the latter two a space of volume V', the conditions are

$$1 - p_1 V(1 - \kappa) = 1 - p_2(\kappa V + \lambda V') = 1 - p_3(1 - \lambda)V'$$

which supply

$$\kappa = 1 - \frac{p_2 + p_3}{p_1 p_2 + p_2 p_3 + p_3 p_1} \frac{V + V'}{V},$$

$$\lambda = 1 - \frac{p_1 p_2 p_3}{p_1 p_2 + p_2 p_3 + p_3 p_1} \frac{V + V'}{V'}$$

and the success rate has the value

$$S = 1 - (V + V') \frac{p_1 p_2 p_3}{p_1 p_2 + p_2 p_3 + p_3 p_1}.$$

If the p_ν are normal density functions, say

$$p_\nu(x, y) = \frac{\sqrt{D_\nu}}{\pi} e^{-\frac{1}{2}Q_\nu},$$

$$Q_\nu = \alpha_\nu (x - a_\nu)^2 + 2\beta_\nu (x - a_\nu)(y - b_\nu) + \gamma_\nu (y - b_\nu)^2$$

and D_ν the corresponding determinants, the curves separating the regions R_ν are the conics

$$Q_\nu - Q_\mu = \text{const.}$$

where the constants are determined by the conditions that all P_ν must be equal. If the α, β, γ have the same values for every ν, the borders consist of straight lines. In this case one can reduce the expressions for p_ν, by an affine transformation, to

$$p_\nu(x, y) = \frac{1}{\pi} e^{-(x-a_\nu)^2-(y-b_\nu)^2}.$$

In the transformed plane the borderline between the regions R_ν and R_μ is perpendicular to the straight line that connects the points $A_\nu(a_\nu, b_\nu)$ and $A_\mu(a_\mu, b_\mu)$. If all points A_ν lie on the same straight line (in particular, if $n = 2$) the whole problem is practically identical with the one-dimensional ($m = 1$). In the case $n = 3$, in general, the three regions are confined by three lines perpendicular to A_1A_2, A_2A_3, A_3A_1 passing through a point C whose coordinates are determined by the equations $P_1 = P_2 = P_3$. If r_ν denotes the distance $A_\nu C$ and φ_ν, ϑ_ν are the angles, $A_\nu C$ forms with the adjacent sides of the triangle $A_1A_2A_3$ one has to use the function

$$F(r, \varphi_1) = \frac{1}{2\sqrt{\pi}} \int_0^\infty \phi(r - z \tan \varphi) e^{-z^2} dz.$$

Then the two conditions for C read

$$F(r_1, \varphi_1) + F(r_1, \vartheta_1) = F(r_2, \varphi_2) + F(r_2, \vartheta_2) = F(r_3, \varphi_3) + F(r_3, \vartheta_3)$$

and the success rate equals 0.5 plus the common value of these three expressions.