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.