LETTERS TO THE EDITOR

A GENERAL THEORY OF DISCRIMINATION WHEN THE ALTERNATIVE POPULATION DISTRIBUTIONS INVOLVE UNKNOWN PARAMETERS

The problem of discrimination, that is of assigning an observed individual to its proper group, admits a simple solution when the distributions of measurements in the alternative populations are completely specified. In practice, the alternative distributions are not completely known but may be estimated on the basis of samples drawn from each alternative population. It can be formally demonstrated that if the parameters are based on large samples, then the above solution in which the estimates are substituted for parameters is asymptotically the most efficient. The small sample case needs some investigation. The problem may be stated as follows:

Two samples of sizes n_1 and n_2 are available from two populations $P_1(x/\theta_1)$ and $P_2(x/\theta_2)$, where x stands for all the measurements and θ for all the parameters. An individual with known measurements y has to be assigned as a member of one of the two groups basing the decision on the observed values only, the parameters occurring in the alternative distributions being unknown. Wald[1] proposed to solve this problem in the case of p-variate normal populations by finding the distribution of the estimated discriminant function earlier suggested by Fisher for classificatory purpose. This involves unknown parameters and cannot, therefore, be applied in practice. In this note an attempt is being made to lay down a decision rule independent of the unknown parameters.

If the measurements are p in number, we have a total of $(n_1 + n_2 + 1)p$ observations which can be represented by a point in a Euclidean space. The decision rule requires the division of the space into two regions R_1 and R_2 such that when the point of observations falls in R_1 the individual is assigned to the first group and otherwise to the second. Whatever may be the set of regions, it should have the property that errors of classification when the alternative populations are different must be smaller than those when the populations are the same. This criterion leads to the restriction that the size of each region should be the *same* whenever the two probability densities P_1 and P_2 are identical irrespective of what the actual values of the common parameters are. This part of the problem is immediately solved if similar regions exist with respect to all the unknown parameters when the probability densities are the same.

We have now to fix the size of the regions $\alpha_1(\theta, \theta)$ and $\alpha_2(\theta, \theta)$ of R_1 and R_2 when P_1 and P_2 are identical. When the population distributions are identical, the decision may be equivalent to that of tossing an unbiased coin so that it is reasonable to take $\alpha_1 = \alpha_2 = ½$. We could also fix them by assigning the ratio of errors of classification at a specified set of parameters in the two different distributions. The special case of fixing $\alpha_2(\theta, \theta)$ at the 5% level leads to a test of the null hypothesis that the individual belongs to the first group at level $\leqslant 5\%$, the alternative being the second group. Similarly, the other null hypothesis can be tested.

The problem is now to determine such similar divisions R_1, R_2 covering the entire space which have fixed values when the two distributions are identical and for which the errors of classification is a minimum. In many cases, similar regions are constructed by choosing portions from surfaces of constant values of sufficient statistics. The problem of minimising the errors reduces to suitably dividing the region common to the surfaces of sufficient statistics as in the general problem of testing composite hypothesis.

Again, in all cases, no uniformly best division is possible on the surfaces of sufficient statistics. We may then determine regions for which the errors of classification is least, at least locally, *i.e.*, for small departures from the equality of populations.

By using these principles, it has been possible to solve the discrimination problem as well as testing of hypothesis in a number of cases including multivariate normal distributions by just utilizing the observed values only and making no assumption about the unknown parameters.

The theory is general and can be applied even when the alternative distributions are more than two. The detailed mathematical treatment will be published elsewhere. Related problems of selection, greater mean, etc., can also be easily solved by this method.

Statistical Laboratory, C. RADHAKRISHNA RAO.
Presidency College, Calcutta-12,
February 24, 1953.

1. Wald, A., *Ann. Math. Stat.*, 1944, **15**, 145.

OCCURRENCE OF GYPSUM IN PALLADAM TALUK, COIMBATORE DISTRICT

THE whole area near Anuppapatti village, Palladam taluk, is a vast stretch of black cotton soil extending to thousands of acres and the topography is very undulating with heavy formations of gullies due to soil erosion. Gypsum occurs as modules of varying sizes both in the profile and as an outcrop as well in a basin covering about 200 acres. The profile deposits are present at a depth of 2′-8′ depending on the topography. Blocks of granites, gneisses and *kankar* of different sizes are found on the surface.

The clay content of the soil ranges between 48 to 55 per cent., the lower depths being heavier than the surface and subsoil. The soils are alkaline in reaction (pH 8·2 to 8·6) and contain high amounts of water-soluble salts, mostly the sulphates of sodium, calcium and magnesium; the carbonates and bicarbonates of sodium being completely absent. The water-soluble salt content also shows an increase with depth, the lower depths having higher concentrations than the surface and subsoil. A petrological analysis of the samples reveal the presence, in addition to gypsum, of the following minerals, *viz.*, quartz, ferro-magnesium minerals, etc. Gypsum is found associated with *kaolin* and *kankar*, the latter being present throughout profile either in a nodular form or as streaks.

The soils of this area are similar to the black cotton soils with gypsum occurring in the Rayalaseema (Ceded Districts) in all respects, *viz.*, parent material, mode of formation, salt concentration, texture and other physical characteristics.

In view of the absence in the soil of salts toxic to plants or inhibitory to their growth in the top zones, the high lime status and the dry intensive cultivation that is being followed, the area would continue to support good vegetation provided suitable precautions are taken against soil erosion and consequent depletion of soil fertility.

The formation of a zone of gypsum concentration is only secondary which may be due to the absence of heavy leaching and to the high retentive and absorptive capacity of the clay complex present in the soil. The gypsum nodules contain the hydrated sulphate of lime in a crystalline form, crystallised in the monoclinic system in flat rhomboid tables and the crude material as obtained fresh from the field contains 65% $CaSO_4$. A rough estimate puts