ON A TEST WHETHER TWO SAMPLES ARE FROM THE SAME POPULATION[1]

By A. Wald[2] and J. Wolfowitz

1. The Problem.[3] Let X and Y be two independent stochastic variables about whose cumulative distribution functions nothing is known except that they are continuous. Let $x_1, x_2, \cdots, x_m$ be a set of m independent observations on X and let $y_1, \cdots, y_n$ be a set of n independent observations on Y. It is desired to test the hypothesis (the null hypothesis) that the distribution functions of X and Y are identical.

An important step in statistical theory was made when "Student" proposed his ratio of mean to standard deviation for a similar purpose. In the problem treated by "Student" the distribution functions were assumed to be of known (normal) form and completely specified by two parameters. It is clear that in the problem to be considered here the distributions cannot be specified by any finite number of parameters.

It might nevertheless be argued that by virtue of the limit theorems of probability theory, "Student's" ratio might be used in our problem for large samples. Such a procedure is open to very serious objections. The population distributions may be of such form (e.g., Cauchy distribution) that the limit theorems do not apply. Furthermore, the distributions of X and Y may be radically different and yet have the same first two moments; clearly "Student's" ratio will not distinguish between two such distributions.

The Pearson contingency coefficient is a useful test specifically designed for the problem we are discussing here, but one which also possesses some disadvantages. The location of the class intervals is to a considerable extent arbitrary. In order to use the χ^2 distribution, the numbers in each class interval must not be small; often this can be done only by having large class intervals, thus entailing a loss of information.

2. Preliminary remarks. Denote by $P\{X < x\}$ the probability of the relation in braces. Let $f(x)$ and $g(x)$ be the distribution functions of X and Y respectively; e.g., $P\{X < x\} = f(x)$. Throughout this paper we shall assume that $f(x)$ and $g(x)$ are continuous.

Let the set of $m + n$ elements $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ be arranged in

---

[1] Presented to the Institute of Mathematical Statistics at Philadelphia, December 27, 1939.

[2] Research under a grant-in-aid from the Carnegie Corporation of New York.

[3] The authors are indebted to Prof. S. S. Wilks for proposing this problem to them.