A COMBINATORIAL TEST FOR THE PROBLEM OF TWO SAMPLES FROM CONTINUOUS DISTRIBUTIONS

S. S. WILKS
PRINCETON UNIVERSITY

1. Introduction

The general problem of testing the null hypothesis that two independent samples come from identical continuous distributions against the alternative hypothesis that they come from *any* pair of different continuous distributions, has been considered by Smirnov [11], Wald and Wolfowitz [13], and others. The two-sample problem for testing the null hypothesis against various *restricted classes* of pairs of alternatives has been considered by Dixon [3], Wilcoxon [14], Mann and Whitney [7], Lehmann [5], Mood [8], Savage [10], Sukhatme [12], and other authors.

The purpose of this paper is to consider a simple combinatorial test for the general two-sample problem based on what are called "cell frequency counts" which one sample generates with respect to the other. The test proposed is consistent for testing the null hypothesis against alternatives in the class of all pairs of different continuous distributions subject to mild assumptions. The test criterion suggested, defined by (4.8), has as its limiting distribution in large samples a chi-square distribution under the null hypothesis. The power of the test is considered in some detail for alternatives in which the two distributions are "nearly" equal.

To be more precise let $\mathcal{C}$ be the class of all pairs of continuous c.d.f.'s $(F(x), G(x))$ and let $\mathcal{C}_0$ be the subset of $\mathcal{C}$ for which $F(x) \equiv G(x)$.

Let $(x_{(1)}, \cdots, x_{(n)})$, with $x_{(1)} < \cdots < x_{(n)}$, be the order statistics of a sample O_n from $F(x)$ and let $I_1, \cdots, I_{n+1}$ be the intervals $(-\infty, x_{(1)}]$, $(x_{(1)}, x_{(2)}]$, $\cdots$, $(x_{(n-1)}, x_{(n)}]$, $(x_{(n)}, +\infty)$, respectively. In an independent sample O'_m from $G(x)$, let $r_1, \cdots, r_{n+1}$ be the numbers of elements in O'_m which fall into $I_1, \cdots, I_{n+1}$, respectively. Then $(r_1, \cdots, r_{n+1})$ is a discrete vector random variable where $r_1, \cdots, r_{n+1}$ are nonnegative integers satisfying

$$r_1 + \cdots + r_{n+1} = m. \tag{1.1}$$

Next we define a new vector random variable $(s_0, s_1, \cdots, s_m)$ where

$$s_i = \text{number of } (r_1, \cdots, r_{n+1}) \text{ which are equal to } i, \tag{1.2}$$

Research partially supported by the Office of Naval Research.