A METHOD OF TESTING THE HYPOTHESIS THAT TWO SAMPLES ARE FROM THE SAME POPULATION

By Harold C. Mathisen

Princeton University

1. Introduction. There are many cases in testing whether two samples are from the same population in which no assumption about the distribution function of the population can be made except that it is continuous. A. Wald and J. Wolfowitz, [1], have developed a method of testing the hypothesis that two samples come from the same population based on certain kinds of runs of the elements from each sample in the combined ordered sample. W. J. Dixon, [2], has introduced a criterion for testing the same hypothesis based on the number of elements of the second sample falling between each successive pair of ordered values in the first sample.

The problem considered here is that of devising a simple method of testing the hypothesis that two samples come from the same population, based on medians and quartiles, given only that the distribution function of the population is continuous. The simplest method may be described briefly as follows. We observe the number of elements, m_1, in the second sample whose values are lower than the median of the first sample. Since the distribution of m_1 is independent of the population distribution, we are able to compute significance points from the distribution of m_1. These points may then be used for testing the hypothesis at a given significance level. This will be referred to as the case of two intervals.

This method may be easily extended to the case of any number of intervals. In this note we shall consider the extension to four intervals by using the median and the two quartiles of the first sample to establish four intervals into which the elements of the second sample may fall. Then, if the second sample is of size $4m$, it will be shown that, under the hypothesis that the two samples come from the same population, $\frac{1}{4}$ of the second sample, or m elements will be expected to fall in each interval. Let the number in the second sample which actually fall in each interval be m_1, m_2, m_3, and m_4 respectively. The test function here proposed is,

$$C = \frac{(m_1 - m)^2 + (m_2 - m)^2 + (m_3 - m)^2 + (m_4 - m)^2}{9m^2}, \tag{1}$$

where $9m^2$ is a constant, which forces C to lie on the interval 0 to 1. If the m_i, ($i = 1, 2, 3, 4$), have values quite different from their expected value m, it is apparent that C will be large. Therefore the greater the value of C the more doubtful is the hypothesis that the two samples come from the same population. Significance values of C will be computed for several sample sizes. The question of whether C is the "best" four-interval criterion for testing the hypothesis that two samples come from the same continuous distribution is an open one