ON THE TWO SAMPLE PROBLEM: A HEURISTIC METHOD FOR CONSTRUCTING TESTS[1]

By V. P. Godambe

Science College, Nagpur, India

1. Introduction. The two-sample problem arises as follows. We are given two independent samples from populations A and B respectively and are required to investigate whether the population A could be considered as identical with B. In the usual terminology of hypothesis testing: Given two independent samples $x_1, \cdots, x_m$ and $x_{m+1}, \cdots, x_{m+n}$ from populations with unknown cumulative distribution functions F and G respectively, the problem is to test the composite hypothesis

$$H_0: F = G$$

against the alternatives

$$H_1: F \neq G,$$

F and G being completely or partially unspecified.

In the following lines we shall discuss a method (subsequently called the V-method), for testing H_0 against H_1, when F and G are partially specified (the exact meaning of this will be clear later). A test for the situation where F and G are completely unspecified is also put forward.

2. Notation. Suppose $F(x)$ and $G(x)$ to be two cumulative distribution functions on the real axis, $-\infty < x < \infty$, such that their frequency functions exist everywhere. Let $x_1, \cdots, x_m$ and $x_{m+1}, \cdots, x_{m+n}$ denote independent samples from F and G respectively. Now the combined sample from F and G can be represented as a point

$$\mathbf{x} = (x_1, \cdots, x_m, x_{m+1}, \cdots, x_{m+n}) \tag{2.1}$$

in the $m + n$ dimensional Euclidean space $\mathfrak{X}$ of all such points. It follows from the existence of the frequency functions that the probability measure of the set of points $\mathbf{x}$ in $\mathfrak{X}$ defined by $x_i = x_j$ for $i \neq j$ is zero. Next we define on $\mathfrak{X}$ a vector-valued function $\boldsymbol{\gamma}$,

$$\boldsymbol{\gamma}(\mathbf{x}) = (\gamma_1(\mathbf{x}), \cdots, \gamma_i(\mathbf{x}), \cdots, \gamma_{m+n}(\mathbf{x})), \tag{2.2}$$

where $\gamma_i(\mathbf{x})$ is the total number of the components of $\mathbf{x}$ less than or equal to x_i. Thus $\gamma_i(\mathbf{x})$ is the rank of x_i in the combined sample $\mathbf{x} = (x_1, \cdots, x_{m+n})$. Further we arrange the last n components of $\mathbf{x}$, that is $x_{m+1}, \cdots, x_{m+n}$, accord-

Received July 24, 1957; revised June 9, 1959.

[1] This work was submitted as a part of a Ph.D. thesis of the London University. The paper was supported (in part) by funds provided under contract AF-18 (600) 456 with the School of Aviation Medicine, USAF, Randolph Air Force Base, Texas.