LIMIT THEOREMS FOR FUNCTIONS OF SHORTEST TWO-SAMPLE SPACINGS AND A RELATED TEST[1]

By Saul Blumenthal

New York University

1. Introduction and notation. Limit theorems for certain functions of two-sample "sample spacings" are given, and then applied to obtain the large sample properties of a procedure for testing whether two distribution functions ($F(x)$ and $G(x)$) are the same. The present limit results extend earlier work of Blum and Weiss [1], and the proposed test is analogous to one used by Weiss [6].

Denote observations from one population by $X_1, X_2, \cdots, X_m$ and from the other population by $Y_1, Y_2, \cdots, Y_n$, with labels chosen so that $m = \theta n$ with $\theta \geqq 1$. The X's are independent with common distribution function $F(x)$, and the Y's are independent with common distribution function $G(x)$. Let p_0 $(0 < p_0 < 1)$ be given (choice of a value for p_0 will be discussed in Section 3).

The ordered X-values will be denoted $X_1' \leqq \cdots \leqq X_m'$, and the ordered Y's by $Y_1' \leqq \cdots \leqq Y_n'$. Let Y_0' denote $-\infty$ and Y_{n+1}' denote $+\infty$. By S_i we denote the number of $X_1, \cdots, X_m$ which are contained in the interval $[Y_{i-1}', Y_i')$ $(i = 1, \cdots, n+1)$. The S_i are the numbers of X's "separating" adjacent ordered Y's and are sometimes referred to as "sample spacings." S_i will be seen to be a measure of the "probability content" of the interval $[Y_{i-1}', Y_i')$.

For an arbitrary k and collection of indices $(i_1, \cdots, i_k)$ we write

$$I_n = \bigcup_{j=1}^{k} [Y_{i_j-1}', Y_{i_j}') \tag{1.1}$$

and we denote the "content" of I_n as

$$H_n = \sum_{j=1}^{k} (S_{i_j} + 1)/(n + m + 1). \tag{1.2}$$

We shall study $I_n(p_0)$ where the indices i_j are chosen so that intervals $[Y_{i-1}', Y_i')$ with small corresponding S_i values are included in $I_n(p_0)$, and enough intervals are included so that $H_n(p_0)$ is as close to p_0 as possible without exceeding p_0. Thus if any interval with an S_i value of r is included in $I_n(p_0)$, all intervals with S_i values of less than r will be included. Generally many intervals will have a given S_i value, and if inclusion of all intervals with $S_i = r_0$ (say) would make $H_n(p_0) > p_0$, then an arbitrary subset of those intervals can be chosen subject to

$$p_0 - [(r_0 + 1)/(n + m + 1)] < H_n(p_0) \leqq p_0. \tag{1.3}$$

To formalize the definition of $I_n(p_0)$, we define K_n as the largest integer such

Received 18 June 1965; revised 11 August 1966.

[1] Part of this work is contained (slightly modified) in a thesis submitted to Cornell University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Research supported in part by Air Force Contract No. AF49(683)-230, Cornell University, and by National Science Foundation Grant GP4933, Rutgers—The State University.