NONPARAMETRIC ESTIMATION, III. STATISTICALLY EQUIVALENT BLOCKS AND MULTIVARIATE TOLERANCE REGIONS—THE DISCONTINUOUS CASE

By John W. Tukey

Princeton University

1. Summary. In Paper II of this series [2, 1947] it was shown that if n functions and a sample of n were used to divide the population space into $n + 1$ blocks in a particular way, and *if the joint cumulative of the functions were continuous*, then the $n + 1$ fractions of the population, corresponding to the $n + 1$ blocks, were distributed symmetrically and simply.

In Paper I of this series [1, 1945] it was shown that the one-dimensional theory of tolerance regions could be extended to the discontinuous case, if equalities were replaced by inequalities.

In this paper the results of Paper II will be extended to the discontinuous case with the same weakening of the conclusion. The devices involved are more complex, but the nature of the results is the same (See Section 5).

As a tool, it is shown that any n-variate distribution can be represented in terms of an n-variate distribution with a continuous joint cumulative (in fact, with uniform univariate marginals), where each variate of the given distribution is a different monotone function of the corresponding variate from the continuous distribution.

2. Introduction. The importance of extending the simple results of the continuous case to the more complex results of the discontinuous case may not be clear at first thought. Yet all the data with which the statistician actually works comes from *discontinuous distributions*. Often these distributions are very fine-grained—the distributions of the number of eggs laid by codfish and of the measured wavelengths of a spectral line (measured in 0.000001 Å) do not have *large* concentrated probabilities, but *all* their probability *is* concentrated at discrete points. Insofar as the considerations of the theoretical statistician apply to the data as received rather than to the "data" of a more or less imaginary model, these considerations apply to data with a discrete distribution. When his theories are erected on a basis of a probability density function, or even a continuous cumulative, there is a definite extrapolation from theory to practice. It is, ultimately, a responsibility of the mathematical statistician to study discrete models and find out the dangerous large effects and the pleasant small effects which go with such extrapolation. *We all deal with discrete data, and must sooner or later face this fact.*

In order to deal with the discontinuous case, we must face two problems: (we assume that the reader is familiar with Paper II [2])

(1) What to do about "ties"?

(2) Finite probabilities associated with cuts.

The first of these is peculiar to the multivariate situation and can be easily explained by an example. Consider the three points in the plane with coordinates (1, 9), (3, 9) and (2, 6). Let the first two functions be y and x, then the procedure of Section 4 of Paper II [2] is not unique—two possibilities arise:

Alternative A. (1, 9) is selected as having the largest y, and (3, 9) as having the largest x among the remaining (two) points, hence $S_1 = \{(x, y)|y > 9\}$, $S_2 = \{(x, y)|y < 9, x > 3\}$, $S_{2|4} = \{(x, y)|y < 9, x < 3\}$.

Alternative B. (3, 9) is selected as having the largest x, and (2, 6) as having the largest x among the remaining (two) points, hence $S_1 = \{(x, y)|y > 9\}$, $S_2' = \{(x, y)|y < 9, x > 2\}$, $S_{2|4}' = \{(x, y)|y < 9, x < 2\}$.

Notice that $S_2' \neq S_2$. The procedure is not unique. In the continuous case, ties happen with probability zero, hence their consequences could be neglected. This is now no longer the case.

This difficulty is solved by using more functions and the idea of lexicographical (like a dictionary!) ordering. In the simplest case, we add no new functions and proceed as follows: If there is a unique i for which $\varphi_1(w_i)$ is maximal, select it. Otherwise look among the w_i for which $\varphi_1(w_i)$ is maximal—look at the values of $\varphi_2(w_i)$. If there is a unique such i for which $\varphi_2(w_i)$ is maximal, select it. If not, go on to $\varphi_3(w_i) \cdots$. This procedure leads to a specific i unless $\varphi_h(w_j) = \varphi_h(w_k)$ for h and some $j \neq k$. But in this case it does not matter whether j or k is selected, the set of m-tuples $(\varphi_1(w_i), \varphi_2(w_i), \cdots, \varphi_m(w_i))$ remaining will be the same, although the indices i will not. But the indices play no role in the actual construction.

As an example, consider the following 20 four-letter words as a sample and let there be four functions—φ_i being the negative of the position in the alphabet of the i-th letter of the word. (Thus $a > b > c > \cdots > z$.)

Sample: meet, west, made, gone, come, back, said, that, maid, well, with, with, just, week, very, near, edge, this, last, have. (*The Law of the Three Just Men*, Edgar Wallace, pp. 159–160).

Selections: *back*, m*a*d*e*, n*e*a*r*, (gon*e*, com*e*, edg*e*, hav*e*. The fourth selection to be made at random among these four.) The inferences which can be made about the four-letter words in Edgar Wallace's writing vocabulary are left to the reader.

We have just given one rule for breaking ties, one which chooses Alternative B in our example. But we might prefer a rule which chooses Alternative A. To get more generality, we have only to take M functions, $M \geq m$, and let $\varphi_{p(1)}$, $[\varphi_{p(2)}, \cdots, \varphi_{p(m)}$, (where we may suppose $p(1) = 1$ without loss of generality) play the role just taken by $\varphi_1, \varphi_2, \cdots, \varphi_m$. Thus if the maximum of $\varphi_1(w)$ is not unique proceed to $\varphi_2(w)$, thence to $\varphi_3(w), \cdots$, thence to $\varphi_m(w)$. For the second block, start with $\varphi_{p(2)}$, then $\varphi_{p(2)+1}, \varphi_{p(2)+2}, \cdots, \varphi_m$. And so on. The choice

$$\varphi_1(x, y) = y,$$

$$\varphi_2(x, y) = -x,$$

$$\varphi_3(x, y) = xe^y,$$
$$\varphi_4(x, y) = x,$$
$$\varphi_5(x, y) = y^2,$$

with $p(1) = 1$ and $p(2) = 4$, leads to Alternative A above. (Note that φ_3 is a dummy in the sense that it is never used.) The problem of ties, which was a problem in uniqueness of construction, is thus dealt with.

Next we must deal with the cuts. When we made S_1, S_2 and $S_{2|4}$ in Alternative A, we omitted some points, namely

$$T_1 = \{(x, y) \mid y = 9\}, \quad \text{and} \quad T_2 = \{(x, y) \mid y < 9, x = 2\}.$$

In the continuous case this did not matter, since these sets had probability zero and could be avoided. Here they cannot, and we shall have to consider a family of blocks (in the wide sense) as consisting of the blocks S *and the cuts* T. The solution of the univariate case in Paper I [1] shows us that what we must expect is that:

$$\Pr \{ \text{coverage } S_i + T_{i-1} + T_i > t\} \geq \Pr \{ \text{coverage of one continuous-case block} > t\} \geq \Pr \{ \text{coverage } S_i > t\}.$$

That is, if we want a certain set of blocks to cover (together) *at least* a certain amount with a certain probability we must add the adjoining cuts; and if we want a certain set of blocks to cover *at most* a certain amount with a certain probability we may add only these cuts which do not adjoin blocks not in our set. By introducing the cuts explicitly, we solve the second problem.

In order to reduce the size of the cuts, our detailed definitions will differ in detail from those which we have used so far. In the example, where the functions leading to Alternative A are used; we place in S_1 not only the points with $y > 9$, but also those with $y = 9$ and $-x > -1$; we place in S_2 not only the points with $y < 9$ and $x > 3$ and the points with $y < 9$, $x = 3$, $y^2 > 49$, but also those with $y = 9$ and $-x < -1$. Proceeding in this way, we reduce T_1 to the point $x = 1$, $y = 9$ and T_2 to the point $x = 3$, $y = 9$. This reduction can only diminish the probability associated with the cuts, but we cannot be sure that it will reduce it to zero.

Only in the quasi-trivial case, where the probability that all functions shall tie together is zero, do we return to the simplicity of the continuous case. This case is quasi-trivial because it does not arise with discrete probabilities, and real observations always involve discrete probabilities.

Having discussed the results, we should now briefly touch on the methods. The proof of the main theorems depends on two facts:

(1) a representation theorem, (5.3), and

(2) a lemma, (6.1) which shows that m functions would be enough if (i) the distribution were fixed, and (ii) cases of probability zero were neglected. The

representation theorem has been outlined in the summary. It is analogous to, but a definite extension of the one used in Paper I [1]. It seems to be new in statement, though not in thought—it will surprise few probability theorists. The novel element is the monotonicity of the functions, which is utterly essential for our purposes.

The lemma allows us to reduce the general case to the case of no extra functions, where the reduction must be made differently for each underlying distribution. The reduced functions are then represented by the representation theorem and the results of Paper II [2] are taken over. The results are stated in a form independent of the underlying distribution and the particular representation, hence they apply in general.

The last paragraph stresses the principle common to Paper I [1] and this paper. It is natural to call it the "iceberg principle," and to sketch it as follows: "We have some information about the visible one-ninth of the iceberg, and we want to conclude something about this visible part. If we can imagine another eight-ninths, consistent with the part we know, and if using that we can prove something expressed solely in terms of the visible part, then this is the required proof. (The only essential is to be able to match *every* visible part.)" Both the reduced functions (which depend on the underlying distribution) and the uniform variables used to represent them are part of the invisible eight-ninths which "could be there."

3. Terminology and Notation. In general we use the terminology and notation of Paper II [2], and we shall continue to assume that all functions concerned in the argument are measurable.

Given two finite sequences of the same length, we write $(a_1, a_2, \cdots, a_m) > (b_1, b_2, \cdots, b_m)$ if *any of the following hold:*

$$a_1 > b_1,$$

$$a_1 = b_1, \text{ and } a_2 > b_2,$$

$$a_1 = b_1, a_2 = b_2, \text{ and } a_3 > b_3,$$

$$\cdots$$

$$a_i = b_i \text{ for } i < m, \text{ and } a_m > b_m.$$

This is the lexicographical order referred to above. (We interpret $(a_1, a_2, \cdots, a_m) < (b_1, b_2, \cdots, b_m)$ to mean $(b_1, b_2, \cdots, b_m) > (a_1, a_2, \cdots, a_m)$ and $=$ to mean identity.)

3.1 DEFINITION: *Given a sequence of real-valued functions* $\varphi_1, \varphi_2, \cdots, \varphi_M$ *and a sequence of starting indices* $p(1), p(2), \cdots, p(m)$, (which we shall often refer to, briefly, as an m-system of functions, $\varphi_1, \varphi_2, \cdots, \varphi_M$, without explicitly mentioning the starting indices), *the functions* $\Phi_1, \Phi_2, \Phi_3, \cdots, \Phi_m$ *are defined as follows:*

$$\Phi_k(w) = \{\varphi_{p(k)}(w), \varphi_{p(k)+1}(w), \cdots, \varphi_M(w)\}, \tag{3.2}$$

the values of Φ_k being sequences of $M - p(k) + 1$ numbers. (In these terms, the rule for tie-breaking already explained becomes "select an i for which $\Phi_k(w_i)$ is maximal (in the sense of lexicographical ordering)".)

4. The blocks and cuts determined by n points. 4. DEFINITION: *Given an* m*-system of functions $\varphi_1, \varphi_2, \cdots, \varphi_M$ and n points $w_1, w_2, \cdots, w_n$, $(m \leq n)$ the corresponding blocks and cuts are given by the following procedure:* (the Φ's are defined in 3.1) *First $i(1)$ is selected to maximize $\Phi_1(w_i)$, when*

$$S_1 = \{w \mid \Phi_1(w) > \Phi_1(w_{i(1)})\},$$

$$T_1 = \{w \mid \Phi_1(w) = \Phi_1(w_{i(1)})\}.$$

Next, $i(2)$ is selected $\neq i(1)$ and to maximize $\Phi_2(w_i)$ among such i, when

$$S_2 = \{w \mid \Phi_1(w) < \Phi_1(w_{i(1)}), \Phi_2(w) > \Phi_2(w_{i(2)})\},$$

$$T_2 = \{w \mid \Phi_1(w) < \Phi_1(w_{i(1)}), \Phi_2(w) = \Phi_2(w_{i(2)})\}.$$

$\cdots\cdots\cdots$(the construction is perfectly analogous to II-4.1)

$$S_{m|n+1} = \{w \mid \Phi_k(w) < \Phi_k(w_{i(k)}), k = 1, 2, \cdots, m\}.$$

4.2 DEFINITION: *If $m = n$, then $S_{n|n+1}$ is also denoted by S_{n+1}.*

If $m > n$, then only $\Phi_1, \Phi_2, \cdots, \Phi_n$ are used and $S_{n|n+1}$ is also denoted by S_{n+1}.

We denote by λ a subset (possibly none, possibly all) of the indices $1, 2, \cdots, m$ and $m|n+1$ or, in case $m \geq n$ of the indices $1, 2, \cdots, n+1$.

4.3 DEFINITION: *The block-group B_λ consists of the union of all S_i with i in λ and all T_i with both i and $i+1$ in λ ($m+1$ means $m \mid n+1$).*

The closed block-group $\bar{B}_\lambda$ consists of the union of all S_i with i in λ and all T_i with either i or $i+1$ in λ.

Given any set we define its *coverage* as the proportion of the population falling into it (here the underlying probability distribution appears for the first time in this section), and we use

4.4 DEFINITION: *The coverage of B_λ is denoted by $C(\lambda)$ and that of $\bar{B}_\lambda$ by $\bar{C}(\lambda)$.*

Thus, given a family of functions φ and n points w, the space of the w is divided into blocks and cuts, these are joined together into block-groups, and these block-groups have coverages. Thus, if the family of functions is fixed, the n points determine these coverages, and, if the points are chance points, the coverages are chance numbers.

5. Statement of results. Having discussed the construction, we can now state the results.

(5.1) THEOREM $A^*_{m|n+1}$. *Let $\varphi_1, \varphi_2, \cdots, \varphi_M$ be any* m*-system of functions and let $W_1, W_2, \cdots, W_n$, where $m \leq n$, be a sample from any distribution, let the blocks, cuts, block-groups and coverages be formed, as described above, using the*

same (unknown) *distribution for forming the coverages. Then, if* $\alpha_1, \alpha_2, \cdots, \alpha_p$ *are any set of* λ*'s (each* λ *is a set of indices!),*

$$\Pr\{C(\alpha_1) < a_1, C(\alpha_2) < a_2, \cdots, \bar{C}(\alpha_k) > a_k, \cdots, \bar{C}(\alpha_p) > a_p\}$$
$$\geq \Pr\{t(\alpha_1) < a_1, t(\alpha_2) < a_2, \cdots, t(\alpha_k) > a_k, \cdots, t(\alpha_p) > a_p\},$$

where $t(\lambda) = \Sigma t_i$ *for* i *in* λ, $t_{m|n+1} = t_{m+1} + \cdots + t_{n+1}$, *and* $t_1, t_2, \cdots, t_{n+1}$ *have a uniform distribution on the barycentric simplex.* (Compare Theorem $A_{m|n+1}$ of Paper II [2].)

In particular,

$$\Pr\{C(i) < a\} \geq I_a(1, n) \geq \Pr\{\bar{C}(i) < a\}, \quad i = 1, 2, \cdots, m,$$

where $I_a(1, n)$ *is the incomplete Beta-function.*

(5.2) THEOREM B^*_{n+1}. *Let* $\varphi_1, \varphi_2, \cdots, \varphi_M$ *be any* n*-system of functions and let* $W_1, W_2, \cdots, W_n$ *be a sample from any distribution. Then*

$$\Pr\{C(\alpha_1) < a_1, C(\alpha_2) < a_2, \cdots, \bar{C}(\alpha_k) > a_k, \cdots, \bar{C}(a_p) > a_p\}$$
$$\geq \Pr\{t(\alpha_1) < a_1, t(\alpha_2) < a_2, \cdots, t(\alpha_k) > a_k, \cdots, t(\alpha_p) > a_p\},$$

where $t(\lambda) = \Sigma t_i$ *for* i *in* λ *and* $t_1, t_2, \cdots, t_{n+1}$ *have a uniform distribution on the barycentric simplex. In particular,*

$$\Pr\{C(i) < a\} \geq I_a(1, n) \geq \Pr\{\bar{C}(i) < a\}, \quad i = 1, 2, \cdots, n + 1.$$

For convenience of reference, we also state the representation theorem as:

(5.3) THEOREM C. *Let* $X_1, X_2, \cdots, X_n$ *have any joint* n*-variate distribution. Then there exist (real) functions* $g_1, g_2, \cdots, g_n$ *and a joint distribution for* $U_1, U_2, \cdots, U_n$ *such that,*

(i) *the marginal distribution of each* U_i *is uniform on* [0, 1],

(ii) *each function* g *is non-decreasing,*

(iii) *the distribution of* $g_1(U_1), g_2(U_2), \cdots, g_n(U_n)$ *is identical with that of* $X_1, X_2, \cdots, X_n$.

6. The functions ψ. The aim of this section is to prove

(6.1) LEMMA. *Given any* m*-system of functions* $\varphi_1, \varphi_2, \cdots, \varphi_M$, *there exist real functions* $\psi_1, \psi_2, \cdots, \psi_M$ *such that,* if $W_1, W_2, \cdots, W_n$ *are a sample from the distribution concerned:*

(6.2) $\Pr\{\psi_i(W_j) = \psi_i(W_k)$, *but* $\psi_{i+h}(W_j) \neq \psi_{i+h}(W_k)$ *for some* $h > 0\} = 0$.

(6.3) $\Pr\{\Phi_i(W_j)$ *has a different relation to* $\Phi_i(W_k)$ *than that of* $\psi_i(W_j)$ *to* $\psi_i(W_k)\} = 0$, *where by relation is meant* $>$, $=$, *or* $<$.

The ψ_i will depend on the underlying probability distribution. Thus they are useful in the proof, but could not replace the Φ_i in the statement of the theorems.

(6.4) LEMMA. *Let* $\Phi(w)$ *have its values in a totally ordered set,* (i.e. *always either* $\Phi_1 < \Phi_2$, $\Phi_1 = \Phi_2$ *or* $\Phi_1 > \Phi_2$) *and let* W *have a distribution. Consider the function* ψ,

$$\psi(w) = \Pr\{\Phi(W) < \Phi(w)\}.$$

Let $W_1, W_2, \cdots, W_n$ be a sample from the same distribution, then, with probability one, the relation ($<$, $=$, or $>$) between $\Phi(W_j)$ and $\Phi(W_k)$ is the same as that between $\psi(W_j)$ and $\psi(W_k)$.

If $\Phi(w_j) < \Phi(w_k)$, then $\psi(w_j) \leq \psi(w_k)$, if $\psi(w_j) < \psi(w_k)$, then $\Phi(w_j) < \Phi(w_k)$. These follow directly from the definition. To prove the lemma, then, we must show that

(i) $\psi(w_j) = \psi(w_k)$ but $\Phi(w_j) < \Phi(w_k)$ occurs with probability zero.

We may clearly assume that the totally ordered set is complete, and that, in particular, it contains the symbols $-\infty$ and $+\infty$. Consider the real function of an abstract variable,

$$F(s) = \Pr\{\Phi(W) < s\}.$$

It is a monotone function, with $F(-\infty) = 0$ and $F(+\infty) = 1$. We can therefore, given $t > 0$, select elements $-\infty = s_0 < s_1 < s_2 < \cdots < s_k = +\infty$ such that

$$0 \leq F(s_{i+1}) - F(s_i + 0) < \epsilon.$$

If (*i*) occurs, then $\Phi(w_j)$ and $\Phi(w_k)$ belong either to the same open interval (s_1, s_{i+1}) or one belongs to an open interval and the other is its upper endpoint. The probability of either of these happening is at most

$$\frac{n(n-1)}{2}\{F(s_{i+1}) - F(s_i + 0)\}^2 + n\{F(s_{i+1}) - F(s_i + 0)\}\{F(s_{i+1} + 0) - F(s_{i+1})\}.$$

Summing this over all intervals yields an estimate of

$$\frac{n(n-1)}{2} \operatorname*{Max}_i \{F(s_{i+1}) - F(s_i + 0)\} = \frac{n(n-1)}{2}\epsilon.$$

Since this goes to zero, the lemma is established.

We turn now to the proof of (6.1). The system of functions $\varphi_1, \varphi_2, \cdots, \varphi_M$ define the $\Phi_1, \Phi_2, \cdots, \Phi_m$ according to Section 3. These define $\psi_1, \psi_2, \cdots, \psi_m$ according to lemma (6.4) just proved. Applying this m times proves (6.3). Recalling that $\Phi_i(w_j) = \Phi_i(w_k)$ implies $\Phi_{i+h}(w_j) = \Phi_{i+h}(w_k)$, we see that (6.3) implies (6.2).

7. The notation $F(x + \lambda \cdot 0)$. All practitioners of analysis are familiar with $F(x + 0)$ and $F(x - 0)$, defined by

$$F(x \pm 0) = \lim_{h \downarrow 0} F(x \pm h).$$

We now generalize this formal notation to

$$F(x + \lambda \cdot 0) = \frac{1 + \lambda}{2} F(x + 0) + \frac{1 - \lambda}{2} F(x - 0), \tag{7.1}$$

where we will, in our immediate applications, need only λ's between -1 and $+1$

(although the definition applies in general). Notice, for example, that

$$F(x - 0) \leq F(x + \lambda \cdot 0) \leq F(x + 0), \quad \text{for } -1 \leq \lambda \leq 1,$$

that if F is continuous at x,

$$F(x + \lambda \cdot 0) = F(x \pm 0) = F(x),$$

that the condition for F to be normalized is

$$F(x + 0 \cdot 0) = F(x).$$

A similar definition is made for functions of two variables, namely

$$F(x + \lambda \cdot 0, y + \mu \cdot 0) = \frac{1 + \mu}{2} F(x + \lambda \cdot 0, y + 0) + \frac{1 - \mu}{2} F(x + \lambda \cdot 0, y - 0)$$

$$= \frac{1 + \lambda}{2} F(x + 0, y + \mu \cdot 0) + \frac{1 - \lambda}{2} F(x - 0, y + \mu \cdot 0),$$

where the two right-hand sides are equal if, as is the case for cumulatives, all doubly one-sided limits exist.

If $F(x_1, x_2)$ is the joint cumulative of two variates, then, when all ordinates and abscissas involved are ordinates and abscissas of continuity,

$$Pr\{a \leq x \leq b, c \leq y < d\} = F(b, d) - F(b, c) - F(a, d) + F(a, c) \geq 0.$$

Passing to the limit in assorted ways, and taking linear combinations gives

$$F(b + \mu \cdot 0, d + \rho \cdot 0) - F(b + \mu \cdot 0, c + \nu \cdot 0) - F(a + \lambda \cdot 0, d + \rho \cdot 0) + F(a + \lambda \cdot 0, b + \nu \cdot 0) \geq 0, \tag{7.2}$$

for $-\infty \leq a, b, c, d \leq +\infty$ and $-1 \leq \lambda, \mu, \nu, \rho \leq 1$. This will be of use shortly.

8. The representation theorem. It was shown in Paper I [1] of this series, that the uniform distribution on [0,1] could serve as the prototype of any variate —that is, that given a distribution, there is a monotone function g, so that $g(U)$ has the given distribution, where U has the uniform distribution on [0, 1]. (In Paper I, U was denoted by X^*).

In the notation of the last section, there is a function $\lambda(u)$, with $|\lambda(u)| \leq 1$, so that

$$F(g(u) + \lambda(u) \cdot 0) = u, \tag{8.1}$$

for all u. (We may, and shall, require that $g(u) = -\infty$, for $u \leq 0$, and $g(u) = +\infty$ for $u \geq 1$). It is easy to see that $g(u)$ is unique except on a set of probability zero and that $\lambda(u)$ is unique (and in fact linear) on each open interval which contains no value of $F(x)$.

Each cumulative $F(x)$, then serves to define $g(u)$ and $\lambda(u)$ by the equation

(8.1). Two or more *independent* variates can be thrown back on a set of *independent* uniform variates by applying this process to their cumulatives separately.

Our present problem is to prove Theorem C (5.3), which applies to variates $X_1, X_2, \cdots, X_n$ which need not be independent. Let $F_i(x_i)$ be the (marginal cumulative of X_i, and use (8.1) to define $g_i(u_i)$ and $\lambda_i(u_i)$. Then define the joint distribution of $U_1, U_2, \cdots, U_n$ by

$$G(u_1, u_2, \cdots, u_n) = F(g_1(u_1) + \lambda_1(u_1)\cdot 0, \cdots, g_n(u_n) + \lambda_n(u_n)\cdot 0),$$

where $F(x_1, x_2, \cdots, x_n)$ is the joint cumulative of the $X_1, X_2, \cdots, X_n$.

We shall verify that this is the desired distribution in the case $n = 2$, leaving the general case to the reader. Consider $G(u_1, +\infty) = G(u_1, 1) = F(g_1(u_1) + \lambda, (u_1)\cdot 0, +\infty)$. This is a cumulative, and so is $G(+\infty, u_2)$. In fact, using (8.1) they are each the uniform cumulative

$$G(u) = \begin{cases} 0, & u \leq 0, \\ u, & 0 \leq u \leq 1, \\ 1, & 1 \leq u. \end{cases}$$

By (7.2) all second differences are positive, and hence $G(u_1, u_2)$ is a joint cumulative. Since its marginals are uniform, it is continuous.

Finally,

$$Pr\{g_1(U_1) < s_1, g_2(U_2) < s_2\} = G(F(s_1 - 0, +\infty), F(+\infty, s_2 - 0))$$
$$= F(s_1 - 0, s_2 - 0),$$

since $g_1(u_1) < s_1$ is equivalent to $u_1 < F(s_1 - 0, +\infty)$ and $g_2(u_2) < s_2$ is equivalent to $u_2 < F(+\infty, s_2 - 0)$. Thus $g_1(U_1)$ and $g_2(U_2)$, have the given bivariate distribution.

9. Proof of main theorems. We come now to the proof of Theorems $A^*_{m|n+1}$ and B^*_{n+1}, and we begin with $A^*_{m|n+1}$. According to Lemma (6.1), the various indices, $i(1), i(2), \ldots, i(m)$ selected to determine the blocks will be the same, excluding cases of probability zero, whether the Φ_i or the ψ_i are used. Consider the first block, which takes the forms:

$$S'_1 = \{W \mid \Phi_1(W) > \Phi_1(w_{i(1)})\}.$$
$$S''_1 = \{W \mid \psi_1(W) > \psi_1(w_{i(1)})\}.$$

Another application of Lemma (6.1) shows that these sets differ by a set of probability zero, and hence their coverages are identical. It will thus suffice to prove theorem $A^*_{m|n+1}$ for a fixed underlying distribution and the corresponding $\psi_1, \psi_2, \cdots, \psi_m$.

According to Theorem C (5.3), the m-variate distribution of the $\psi_i(W)$ can be represented in terms of uniformly distributed variates $U_1, \cdots, U_m$ and monotone functions $g_1(U_1), \cdots, g_m(U_m)$. Now $U_1, U_2, \cdots, U_m$ have a continuous joint

cumulative, so that theorem $A_{m|n+1}$ applies to a sample of n drawn from this m-variate population, with the coordinates themselves as the m functions. We shall denote the coordinates of the i-th element of this sample by $u_1(i), \cdots, u_m(i)$. Consider the first block,

$$S_1 = \{(U_1, \cdots, U_m) \mid U_1 > u_1(i(1))\}.$$

Its image, $g(S) = \{(g_1(U_1), \cdots, g_m(U_m)) \mid U_1 > u_1(i(1))\}$
contains

$$S_1^* = \{(g_1(U_1), \cdots, g_m(U_m)) \mid g(U_1) > g(u_1(i(1)))\},$$

and is contained in the union of S_1^* and T_1^*, where

$$T_1^* = \{(g_1(U_1), \cdots, g_m(U_m)) \mid g(U_1) = g(u_1(i(1)))\}.$$

Thus the conclusions of Theorem $A_{m|n+1}^*$ hold for $S_1^*, T_1^*, \cdots, S_m^*, T_m^*, S_{m|n+1}^*$.

Now while Theorem $A_{m|n+1}^*$ mentions the underlying W's implicitly, careful study shows that they are not really involved; only the joint distribution of the φ_i, which in our present case are the ψ_i, matters. Since this is the same for the $\psi_i(W)$ and the $g_i(U_s)$, Theorem $A_{m|n+1}^*$ must hold for the ψ_i and the theorem is proved.

Theorem B_{n+1}^* is again a special case of Theorem $A_{m|n+1}^*$.

REFERENCES

[1] H. Scheffé and J. W. Tukey, "Nonparametric Estimation I. Validation of order statistics," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187—192 (Also cited as Paper I).

[2] J. W. Tukey, "Nonparametric Estimation II. Statistically equivalent blocks and multivariate tolerance regions. The continuous case," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 529–539 (Also cited as Paper II).