A CHARACTERIZATION OF MULTISAMPLE DISTRIBUTION-FREE STATISTICS[1]

By C. B. Bell[2]

San Diego State College and University of Illinois

1. Introduction and summary. It has been observed [1], [3] that in the one-sample case the distribution-free statistics in common usage are both SDF (strongly distribution-free) and of the form $\psi[F(X_1), \cdots, F(X_n)]$, where $X_1, \cdots, X_n$ is a random sample and F is the hypothesized cpf (cumulative probability function); and it has been proved [1], [3] that the two properties above are equivalent in the one-sample case. In the multi-sample cases, one observes that except for the statistics of the Pitman conditional tests, which are not SDF, a major portion of distribution-free statistics (e.g. Kolmogorov-Smirnov, Cramér-von Mises, Wald-Wolfowitz, Mosteller-Tukey, Epstein-Rosenbaum, empty cell and the rank-sum types) in common usage have both the SDF and rank properties. Z. W. Birnbaum (in a personal communication in March, 1963) asks whether the SDF property implies the rank property in the two-sample case. An affirmative answer to this question and its converse would be of use both in analyzing and constructing multisample distribution-free statistics.

In this paper, it is shown that in the multisample case, the rank property implies the SDF property; and that, except for zero-probability sets, the two properties are equivalent if the k-sample statistic T satisfies Scheffé's [5] NB (null boundary) condition. The former result follows from the definitions of the rank and SDF properties. In proving the latter result one first shows that a completeness property of the class of strictly increasing continuous cpfs implies that each SDF, k-sample statistic T is AI (almost invariant) in the appropriate sense; and, then, that the NB condition and AI property imply invariance and, hence, the rank property almost everywhere.

2. Terminology and preliminaries. In the sequel as is usual in the k-sample case, $k, n_1, n_2, \cdots, n_k$ are arbitrary fixed positive integers with $k \geqq 2$; $N = n_1 + \cdots + n_k$; and $\{X(i,j): 1 \leqq j \leqq n_i; 1 \leqq i \leqq k\}$ are k independent univariate random samples with cpfs (cumulative probability functions) $F_1, \cdots, F_k$, respectively.

In order to exhibit the pertinent features of the various properties considered, one introduces the following.

(I) Ω', the class of strictly increasing continuous univariate cpfs, which is denoted by Ω_2^* in [5].

(II) $\Omega(n)$, the class of n-fold power cpfs, $F^{(n)}$, of elements of a class Ω of cpfs.

Received 17 September 1963; revised 17 December 1963.

[1] This research was supported by the National Science Foundation through grant NSFG-25220 and NSFG-21507.

[2] Now at San Diego State College.

(III) $\tilde{\Omega} = \Omega(n_1) \times \cdots \times \Omega(n_k) = \{F_1^{(n_1)} \cdot F_2^{(n_2)} \cdot \cdots \cdot F_k^{(n_k)}: F_j \varepsilon \Omega, 1 \leqq j \leqq k\}$.

(IV) $\mathfrak{N}(\Omega(N))$ and $\mathfrak{N}(\tilde{\Omega})$ the classes of sets of R_N which have zero probability wrt each cpf of $\Omega(N)$ and $\tilde{\Omega}$, respectively.

(V) $\mathcal{G}'$, the group of strictly increasing transformations of R_1 onto R_1.

(VI) $\mathcal{G}'(N) = \{g_N : g_N[y(1), \cdots, y(N)] = [g(y(1)), \cdots, g(y(N))]$ and $g \varepsilon \mathcal{G}'\}$, the N-fold power group of $\mathcal{G}'$.

(VII) $\mathcal{S}_N$, the symmetric group of $N!$ permutations of N elements. Further, for each element s of $\mathcal{S}_N$ and each point $\bar{y} = [y(1), \cdots, y(N)]$ in R_N, one defines

(VIII) $\bar{s}(\bar{y}) = [y(s(1)), \cdots, y(s(N))]$;

(IX) $B(s, N) = \{\bar{y}: y(s(1)) < \cdots < y(s(N))\}$; and

(X) $\tilde{D} = R_N - \bigcup B(s, N)$, where the union is taken over s in $\mathcal{S}_N$. $\tilde{D}$ is seen to be the set of $\bar{y}$'s having at least one coordinate tie. For the classes defined above one notes immediately

(A) T is a rank statistic if T is invariant under $\mathcal{G}'(N)$;

(B) $\mathcal{G}'$ generates Ω' in the sense that for each cpf

$$H \text{ of } \Omega', \Omega' = \{Hg: g \varepsilon \mathcal{G}'\};$$

(C) for each s in $\mathcal{S}_N$, $B(s, N)$ is similar wrt $\Omega'(N)$, has probability $[N!]^{-1}$; is invariant under $\mathcal{G}'(N)$, and is a set on which each rank statistic is constant; as the referee has essentially proved

(D) $\mathfrak{N}(\Omega'(N)) = \mathfrak{N}(\tilde{\Omega}')$; and

(E) $\tilde{D} \varepsilon \mathfrak{N}(\Omega'(N))$.

Before proceeding, it is necessary to define precisely the NB condition, the SDF and AI properties; and to introduce the symmetry natural for k-sample statistics; and the completeness property required.

A k-sample statistic T is SDF wrt Ω', if for each Borel set A, $P\{T \varepsilon A \mid F_1^{(n_1)}, \cdots, F_k^{(n_k)}\}$ depends only on the functions $F_1 \circ F_2^{-1}, \cdots, F_1 \circ F_k^{-1}$; is AI wrt $\mathcal{G}'(N)$ and $\tilde{\Omega}'$, if $\{T \neq T \circ f_N\} \varepsilon \mathfrak{N}(\tilde{\Omega}')$ for each f_N in $\mathcal{G}'(N)$; and is SWS (sample-wise symmetric) if it is invariant under all permutations within each of the k sets of sample values.

The desired completeness property is, then, as follows. A class Ω of univariate cpfs is said to be SWSC or sample-wise symmetrically complete if (for arbitrary $k, n_1, \cdots, n_k$), (a) T is a SWS k-sample statistic and (b) $\int T\, dH = 0$ for all H in $\tilde{\Omega}$, together imply $\{S \neq 0\} \varepsilon \mathfrak{N}(\tilde{\Omega})$. It can be shown that

(F) Ω' is SWSC.

In the sequel it will sometimes be more feasible to work with the inverse image sets $T^{-1}(A)$ than with the statistic T. For this reason one defines a set W to be SWS; SDF wrt Ω', or AI wrt $\mathcal{G}'(N)$ *and* $\tilde{\Omega}'$ if $W = T^{-1}(A)$, where A is a Borel set, and T is, respectively, SWS, SDF wrt Ω' or AI wrt $\mathcal{G}'(n)$ and $\tilde{\Omega}'$. Following Scheffé [5], W *has structure* $\mathcal{S}_b(0 \leqq b \leqq N!)$ if there exists a set E in $\mathfrak{N}(\Omega'))$ such that for each $\bar{y}$ in $R_N - \tilde{D} - E$, W contains exactly b of the $N!$ points in $\{\bar{s}(\bar{y}) : s \varepsilon \mathcal{S}_N\}$, and W satisfies the NB condition if the boundary of W is an element of $\mathfrak{N}(\Omega'(N))$.

From these definitions it is clear that

(G) for $0 \leqq b \leqq N!$ each union of b elements of $\{B(s, N): s \varepsilon \mathcal{S}_N\}$ is a set having structure S_b, and satisfying the NB condition.

Hence, if T is a rank statistic, $T^{-1}(A)$ satisfies the NB condition for each Borel set A.

With these preliminaries one can now show that the SDF and SWSC properties imply the AI property; that the NB condition and the AI property imply the rank property almost everywhere, and finally that the rank property implies the SDF property.

3. Almost invariance, null boundaries and rank statistics. First one proves

LEMMA 1. *If k-sample statistic T is* SDF wrt Ω' *and is* SWS, *then T is* AI wrt $\mathcal{G}'(N)$ *and* $\tilde{\Omega}'$.

PROOF. For an arbitrary Borel set A, f in $\mathcal{G}'$, $\{G_i\} \subset \Omega'$, let $W = T^{-1}(A)$ and $\bar{P} = P\{W \mid G_1^{(n_1)}, \cdots, G_k^{(n_k)}\}$. Using the SDF property, and then the substitution $y(i,j) = f(x(i,j))$ for all i and j, one derives

$$\bar{P} = P\{W \mid (G_1f)^{(n_1)}, \cdots, (G_kf)^{(n_k)}\} = P\{f_N(W) \mid G_1^{(n_1)}, \cdots, G_k^{(n_k)}\}.$$

Hence, $\int [I(W) - I(f_N(W)) \prod_i \prod_j dG_i(x(i,j)) = 0$ for all $\{G_i\} \subset \Omega'$ and all f in $\mathcal{G}'$, where $I(\cdot)$ is the indicator function. Therefore, for each f_N in $\mathcal{G}'(N)$, and each $W = T^{-1}(A)$, one has that $\{I(W) \neq I(f_N(W))\}$ and $\{T \neq T \circ f_N\}$ are elements of $\mathfrak{N}(\tilde{\Omega}')$; and that T is AI wrt $\mathcal{G}'(N)$ and $\tilde{\Omega}'$.

Now from result (A) of Section 2 it is seen that at this point one needs conditions under which an AI statistic is equivalent to an invariant or rank statistic. The desired result would follow from Lehmann's theorem ([4], p. 225) if one could construct a Haar-type measure on the group $\mathcal{G}'(N)$. However, the approach to be employed here makes use of a theorem of Scheffé ([5]) and places on T the NB restriction, the complete significance of which is not clear to the author. The pertinent theorem of Scheffé ([5]) is essentially

LEMMA 2.

(i) *Each set W of structure S_b is similar* wrt $\Omega'(N)$ *and has probability* $b[N!]^{-1}$; *and further,*

(ii) *if W satisfies the* NB *condition, then W is similar* wrt $\Omega'(N)$ *iff there exists* $0 \leqq b \leqq N!$ *such that W has structure S_b.*

From Scheffé's result one now establishes the principal lemma.

LEMMA 3. *If W is* SWS; SDF wrt Ω'; *and satisfies the* NB *condition, then for each permutation s in $\mathcal{S}_N$, $W \cdot B(s, N) \equiv B(s, N)$ or the empty set $\varnothing$* wrt *each* cpf *in* $\Omega'(N)$.

PROOF. From Lemmas 1 and 2 one sees that if W satisfies the hypothesis, W is AI wrt $\mathcal{G}'(N)$ and $\tilde{\Omega}'$; is, hence, similar wrt $\Omega'(N)$; and has structure S_b for some $0 \leqq b \leqq N!$.

For arbitrary s in $\mathcal{S}_N$, $B(s,N)$ has boundary in $\mathfrak{N}(\Omega'(N))$. Hence, $W \cdot B(s,N)$ also has null boundary. Further, since the class of AI sets is closed under intersections, one has that $W \cdot B(s, N)$ is similar wrt $\Omega'(N)$; and, consequently, must have structure S_b for some $0 \leqq b \leqq N!$.

Since $B(s, N)$ has structure S_1, $W \cdot B(s, N)$ can have structure S_b only for $b = 0$ or 1. If $b = 0$, then $W \cdot B(s, N) \equiv \emptyset$; if $b = 1$, then $W \cdot B(s, N) \equiv B(s, N)$.

One now proves the main theorem.

THEOREM 4. *Let T be a k-sample statistic with $k \geqq 2$.*

(i) *If T is a rank statistic, T is* SDF wrt $\Omega' = \Omega_2^*$.

(ii) *If T is* SWS, *and* SDF wrt Ω', *then T is almost invariant* wrt $\mathcal{G}'(N)$ *and* $\tilde{\Omega}'$.

(iii) *If, in addition to the conditions of* (ii), $T^{-1}(B)$ *satisfies the* NB *condition for each Borel set B, then T is equivalent to a rank statistic.*

PROOF.

(i) If T is a rank statistic, then for each F in Ω',

$$T[X(1, 1), \cdots, X(k, n_k)] = T[F(X(1, 1)), \cdots, F(X(k, n_k))]$$

Therefore, on making the substitution, $V(i, j) = F_i(x(i, j))$ for $1 \leqq j \leqq n_i$, $1 \leqq i \leqq k$, one finds that

$$\begin{aligned} &P\{T[X(1, 1), \cdots, X(k, n_k)] \,\varepsilon\, A \mid F_1^{(n_1)}, \cdots, F_k^{(n_k)}\} \\ &\quad = P\{T[V(1, 1), \cdots, V(1, n_1); F_1 \circ F_2^{-1}(V(2, 1)), \cdots, F_1 \circ F_2^{-1}(V(2, n_2)); \\ &\qquad\qquad \cdots F_1 \circ F_k^{-1}(V(k, n_k))] \,\varepsilon\, A \mid V_0^{(n_1)}, \cdots, V_0^{(n_k)}\} \text{ for all } \{F_i\} \subset \Omega', \end{aligned}$$

where V_0 is the cpf uniform on the interval $(0, 1)$.

T, then, satisfies the definition of an SDF statistic.

(ii) This is Lemma 1.

(iii) T is equivalent to a rank statistic iff for each Borel set A, $T^{-1}(A) \cdot B(s, N)$ is equivalent to $B(s, N)$ or $\emptyset$, since a rank statistic is constant on each $B(s, N)$. The result then follows by applying Lemma 3 to each of the sets $T^{-1}(A)$.

For a variety of reasons, it is sometimes important to consider classes of absolutely continuous cpfs. Since Scheffé's basic result (Lemma 2) can be extended to the classes Ω_3^*, the class of absolutely continuous cpfs of $\Omega_2^* = \Omega'$; and Ω_4^*, the class of cpfs of Ω_3^* which have continuous densities; one proves readily

COROLLARY 5. *The results of Theorem 4 remain valid if one replaces $\Omega_2^* = \Omega'$ by Ω_3^* or by Ω_4^*, except in the expression defining the* NB *condition.*

REFERENCES

[1] BELL, C. B. (1960). On the structure of distribution-free statistics. *Ann. Math. Statist.* **31** 703–709.

[2] BELL, C. B., BLACKWELL, D. and BREIMAN, L. (1960). On the completeness of order statistics. *Ann. Math. Statist.* **31** 794–797.

[3] BIRNBAUM, Z. W. and RUBIN, H. (1954). On distribution-free statistics. *Ann. Math. Statist.* **25** 593–598.

[4] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[5] SCHEFFÉ, H. (1943). On a measure problem arising in the theory of non-parametric tests. *Ann. Math. Statist.* **14** 227–233.