ON THE STRUCTURE OF DISTRIBUTION-FREE STATISTICS[1]

C. B. BELL[2]

Stanford University and *San Diego State College*

Introduction and Summary. Let $X_1, X_2, \cdots, X_n$ be a sample of a one-dimensional random variable X which has the continuous cumulative probability function (cpf) F. It has been observed that the distribution-free statistics commonly appearing in the literature can be written in the form $\Phi[F(X_1), F(X_2), \cdots, F(X_n)]$ where Φ is a measurable symmetric function defined on the unit cube. Such statistics are said to have structure (d).

Birnbaum and Rubin [12] have proved that for the family Ω^*, of strictly monotone continuous cpf's, statistics of structure (d) possess a property stronger than that of being distribution-free.

The purpose of this paper is to study the extension of the Birnbaum-Rubin (B-R) result to other classes of cpf's and to present a different approach to these results. It is found that a one-sided extension of the B-R result is valid for all properly closed, symmetrically complete classes of cpf's. Then, from the existing literature on completeness, one can conclude that the extension is valid for several other classes of statistical interest.

The relation between statistics of structure (d) and strongly distribution-free statistics (Section 1) is of importance for two reasons. First of all, if one is designing distribution-free tests, the results here and in [12] guarantee that if one chooses a statistic of structure (d), one has a strongly distribution-free statistic for several large classes of cpf's.

On the other hand if one has a strongly distribution-free statistic, the results guarantee that it is of structure (d). Hence, its cpf can be written as the volume of a polyhedral region in the n dimensional unit cube. Under such circumstances the work of Smirnov [20], Feller [13], Anderson and Darling [4], and Birnbaum [9] indicate that it should be possible to evaluate the cpf explicity; reduce it to a system of recursion formulae; tabulate it with the aid of high-speed computers or at least evaluate its limiting distribution.

This article is divided into four sections. In Section 1 distribution-free statistics of various types are introduced. Section 2 contains some preliminary results concerning cpf's. The main theorem is proved in Section 3; and Section 4 contains a survey of the known pertinent completeness results as well as a corollary of the main theorem.

1. Distribution-free Statistics. Consistent with the notation of Scheffé [18] and B-R [12] let

Ω_0 = the class of all cpf's;

Received December 26, 1957; revised July 27, 1959.

[1] This research was supported by National Science Foundation Grant NSFG-3662.

[2] Present address: San Diego State College, San Diego 15, California.

Ω_1 = the class of all non-degenerate cpf's;
Ω_2 = the class of all continuous cpf's;
Ω^* = the class of all strictly monotone continuous cpf's;
Ω_3 = the class of all absolutely continuous (with respect to Lebesgue measure) cpf's;
Ω_4 = the class of all cpf's with continuous derivatives;
Ω_u = the class of all cpf's which are uniform within intervals [11], [12]; and
Ω_e = the class of all cpf's with densities of the form

$$C(\theta_1, \cdots, \theta_n) \exp\{-x^{2n} - \theta_1 x - \theta_2 x^2 - \cdots - \theta_n x^n\},$$

[16]. Analogously, for the unit interval I, one defines
$\Omega_0(I)$ = the class of all cpf's on I;
$\Omega_1(I)$ = the class of all non-degenerate cpf's on I;
$\Omega_2(I)$ = the class of all continuous cpf's on I; etc.

If Ω and Ω' are two arbitrary families of cpf's, a real-valued function

$$S_G = S_G(X_1, X_2, \cdots, X_n)$$

will be called a statistic in Ω with regard to (w.r.t.) Ω', if for every $G \in \Omega$, and $F \in \Omega'$; and $X_1, X_2, \cdots, X_n$ in the n-dimensional sample space for a random variable X which has cpf F,

(a) $S_G(X^{(n)}) = S_G(X_1, X_2, \cdots, X_n)$ is defined everywhere in the sample space, and

(b) $S_G = S_G(X^{(n)})$ has a probability distribution; this probability distribution will be denoted by $\mathcal{P}_F^{(n)} S_G^{-1}$.

For example, consider von Mises' statistic

$$w_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - G(x)]^2 \, dG(x) = (1/12n) + \sum_{i=1}^{n} [G(X_i) - (2i - 1)/n]^2;$$

Kolmogoroff's statistic

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - G(x)| = \max_{i=1,\cdots,n} [G(X_i') - (i - 1)/n, (i/n) - G(X_i')];$$

Anderson and Darling's

$$K_n = \sup_{-\infty < x < \infty} \sqrt{n} |F_n(x) - G(x)| (\Psi[G(x)])^{\frac{1}{2}}$$

and

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - G(x)]^2 \Psi[G(x)] \, dG(x)$$

where $F_n(x)$ is the empirical cpf determined by the sample $X_1, \cdots, X_n$; and $X_1', X_2', \cdots, X_n'$ are the ordered sample values. All satisfy (a) and (b) when $\Omega = \Omega' = \Omega_2$. Hence w_n^2, D_n, K_n, and W_n^2 are all statistics in Ω_2 w.r.t. Ω_2.

If for a statistic $S_G(X^{(n)})$ in Ω w.r.t. Ω' there exists a (measurable) function Φ defined on the n-dimensional unit cube and symmetric in its arguments, such

that for any $G \in \Omega$, $F \in \Omega'$, we have $S_G(x^{(n)}) \equiv \Phi[G(x_1), \cdots, G(x_n)][\mathcal{P}_F]$, i.e. almost everywhere in the sample space $X^{(n)}$ for the random variable X which has cpf F, then $S_G(X^{(n)})$ is called a *statistic of structure* (d).

If $\Omega = \Omega'$ and $S_G(X^{(n)})$ has the property that $\mathcal{P}_G^{(n)} S_G^{-1}$, the probability distribution of S_G when X has cpf G, is independent of G for all $G \in \Omega$, then $S_G(X^{(n)})$ is a *distribution-free statistic* in Ω.

If $S_G(X^{(n)})$ is a statistic in $\Omega \subset \Omega^*$ w.r.t. some Ω', then $S_G(X^{(n)})$ is called a *strongly distribution-free statistic* in Ω w.r.t. Ω' if $\mathcal{P}_F^{(n)} S_G^{-1}$ depends only on the function $\tau = FG^{-1}$ for all $G \in \Omega$ and $F \in \Omega'$.

In view of the preceding definitions, it can be readily established that

(A) if a statistic in Ω_2 w.r.t. Ω_2 has structure (d) then it is distribution-free in Ω_2 ;

(B) if a statistic in Ω^* w.r.t. Ω^* is strongly distribution-free, then it is distribution-free in Ω^*; and

(C) if a statistic in Ω^* w.r.t. Ω^* has structure (d), then it is strongly distribution-free.

Further, it is seen that each of the statistics (von Mises, etc.) in the example above is, for properly chosen classes of cpf's, of structure (d); strongly distribution-free and symmetric; and distribution-free. Such also is the case for D_n^+ and D_n^- of Wald and Wolfowitz [21], and Birnbaum, [10]; the spacing statistics of Kimball [17] and Sherman [19]; and most of the other distribution-free statistics in the literature.

Birnbaum and Rubin [12] have shown that there exists a distribution-free statistic which is not strongly distribution-free; but the other two properties always seem to occur together in a statistic. For that reason it is of interest to find the conditions under which the property of having structure (d) is equivalent to being symmetric and strongly distribution-free.

It is known [12] that these two properties are equivalent for statistics in Ω^* w.r.t. Ω^*. In Section 3 it will be shown that the two properties are equivalent for statistics in Ω^* w.r.t. Ω', where Ω' satisfies certain closure and completeness properties.

Before proceeding with the proof of this theorem, it is worthwhile to recall some definitions and results concerning cpf's. This is done below in Section 2.

2. Probability Functions. In view of the nature of the problem, the work will deal primarily with probability spaces on the real line and on the unit interval. For that reason the following classes and sets should be defined.

Let R, $R^{(n)}$, I, $I^{(n)}$, $\mathfrak{B}$, $\mathfrak{B}^{(n)}$, $\mathfrak{B}_I$, and $\mathfrak{B}_I^{(n)}$, be respectively, the real line; euclidean n-space; the open unit interval; the n-dimensional open unit cube; and the respective classes of borel subsets of R, $R^{(n)}$, I, $I^{(n)}$.

A cpf, $F(x)$, on R is a non-decreasing, upper semi-continuous function defined on R and such that $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$. A cpf, $H(u)$, on I is a non-decreasing, upper semi-continuous function defined on I and such that $\lim_{u \to 1} H(u) = 1$ and $\lim_{u \to 0} H(u) = 0$.

It is well known ([2], p. 96) that each cpf on R induces and is induced by a probability distribution on $\mathfrak{B}$; similarly each cpf on I induces and is induced by a probability distribution on $\mathfrak{B}_I$. Let $\mathcal{P}_F$ denote the probability distribution induced by the cpf $F(x)$; and let $\mathcal{P}_F^{(n)}$ denote *power probability distribution* on the class $\mathfrak{B}^{(n)}$ generated by F, i.e. the probability distribution induced by n independent random variables each distributed with cpf F.

If $G, G_1 \varepsilon \Omega^*$, then G, G^{-1}, and G_1G^{-1} are all $1-1$ strictly monotone, continuous mappings; and, hence, preserve many of the properties of cpf's and their probability distributions. In fact,

(i) if $F \varepsilon \Omega_0[\Omega_1, \Omega_2, \Omega^*]$ and $G, G_1 \varepsilon \Omega^*$, then

(a) $FG^{-1} \varepsilon \Omega_0(I)[\Omega_1(I), \Omega_2(I), \Omega^*(I)]$ and

(b) $FG^{-1}G_1 \varepsilon \Omega_0[\Omega_1, \Omega_2, \Omega^*]$.

Since the closure property (b) is important in the sequel, it is worthwhile to give the following formal definition.

Ω' is said to be *closed under* Ω if $FG^{-1}G_1 \varepsilon \Omega'$, whenever $F \varepsilon \Omega'$ and $G, G_1 \varepsilon \Omega$. Therefore, one concludes from (i) that Ω_0, Ω_1, Ω_2 and Ω^* are each closed under Ω^*.

Further, it is seen that under such mappings numerical values are preserved in the following sense.

(ii) If $F \varepsilon \Omega_0$ and $G, G_1 \varepsilon \Omega^*$, then $(\alpha)\mathcal{P}_{FG^{-1}}^{(n)}(B) = \mathcal{P}_F^{(n)}G^{-1}(B)$ for all $B \varepsilon \mathfrak{B}_I^{(n)}$; and $(\beta)\mathcal{P}_{FG^{-1}G_1}^{(n)}[G_1^{-1}(B)] = \mathcal{P}_{FG^{-1}}^{(n)}(B)$ for all $B \varepsilon \mathfrak{B}_I^{(n)}$, where

$$[G(x^{(n)})] = [G(x_1), \cdots, G(x_n)]$$

and $G^{-1}(u, \cdots, u_n) = [G^{-1}(u_1), \cdots, G^{-1}(u_n)]$.

With these preliminary results one can proceed to establish the main theorem.

3. The Main Theorem. As mentioned in the introduction the object here is to demonstrate that for suitable classes of cpf's a statistic is symmetric and distribution-free if and only if it is of structure (d).

If a statistic, S_G, in Ω w.r.t. Ω' is of structure (d), there exists a measurable function Φ defined on $I^{(n)}$ and symmetric in its arguments, such that for any $G \varepsilon \Omega$ and $F \varepsilon \Omega'$, $S_G(x^{(n)}) \equiv \Phi[G(x^{(n)})][\mathcal{P}_F]$.

If A is an arbitrary element of $\mathfrak{B}^{(n)}$, then $S_G^{-1}(A) \equiv G^{-1}o\Phi^{-1}(A)$. In view of (ii), then, $\mathcal{P}_F S_G^{-1}(A) \equiv \mathcal{P}_F G^{-1}o\Phi^{-1}(A) \equiv \mathcal{P}_{FG^{-1}}\Phi^{-1}(A)$ providing FG^{-1} is well defined. Clearly, this will be so whenever $G \varepsilon \Omega^*$. Further, S_G is symmetric whenever Φ is. Therefore, one can conclude the following.

LEMMA 1: *If a statistic, S_G, in $\Omega \subset \Omega^*$ w.r.t. Ω' is of structure* (d), *then S_G is symmetric and strongly distribution-free.*

On the other hand if S_G, a statistic in $\Omega \subset \Omega^*$ w.r.t. Ω', is symmetric and strongly distribution-free, let $\Phi_1 \equiv S_{G_1} o G_1^{-1}$, where G_1 is an arbitrary fixed element of $\Omega \subset \Omega^*$.

It is clear that Φ_1 is symmetric. Therefore, in order to complete the proof one must demonstrate that $S_G(x^{(n)}) = \Phi_1[G(x^{(n)})][\mathcal{P}_F]$ for all $F \varepsilon \Omega'$ and all

$$G \varepsilon \Omega \subset \Omega^*.$$

Again let A be an arbitrary fixed element of $\mathfrak{B}^{(n)}$. Then,

$$\mathcal{P}_F\{\Phi_1[G(x^{(n)})] \varepsilon A\} = \mathcal{P}_F G^{-1} o \Phi_1^{-1}(A) = \mathcal{P}_{FG^{-1}}\Phi_1^{-1}(A) = \mathcal{P}_{FG^{-1}G_1} o\, S_{G_1}^{-1}(A)$$
$$= \mathcal{P}_{FG^{-1}G_1} S_{G_1}^{-1}(A) \qquad \text{for all } F \varepsilon \Omega' \text{ and all } G \varepsilon \Omega \subset \Omega^*.$$

Now, if $FG^{-1}G_1 \varepsilon \Omega'$, i.e. if Ω' is closed under $\Omega \subset \Omega^*$, then the fact that S_G is strongly distribution-free guarantees that $\mathcal{P}_{FG^{-1}G_1} S_{G_1}^{-1}(A) = \mathcal{P}_F S_G^{-1}(A)$ since $(FG^{-1}G_1)G_1^{-1} = (F)G^{-1}$. Under these circumstances one sees that

$$\mathcal{P}_F\{\Phi_1[G(x^{(n)})] \varepsilon A\} = \mathcal{P}_F\{S_G(x^{(n)}) \varepsilon A\} \qquad \text{for all } F \varepsilon \Omega', G \varepsilon \Omega \subset \Omega^*.$$

These results lead one to the following question. What conditions must the class Ω' satisfy in order that S_G and $\Phi o G$, which have identical distributions for each $F \varepsilon \Omega'$, be essentially equal? In answering this question, the following definition will be employed.

A class, Ω, of cpf's is said to be *symmetrically complete* if every unbiased, symmetric estimator of zero, with respect to the class of power probability distributions of Ω, is essentially zero, i.e., the conditions (1) f is symmetric; and (2) $\int_{R^{(n)}} f d\mathcal{P}_F^{(n)} = 0$ for all $F \varepsilon \Omega$, imply that $f = 0[\mathcal{P}_F^{(n)}]$ for all $F \varepsilon \Omega$.

In terms of this definition, the answer to the question is as follows.

LEMMA 2: *If S and Φ are symmetric measurable functions such that*

$$\mathcal{P}_F^{(n)}\{S \varepsilon A\} = \mathcal{P}_F^{(n)}\{\Phi \varepsilon A\}$$

for all $A \varepsilon \mathfrak{B}$ and all $F \varepsilon \Omega'$; and if Ω' is a symmetrically complete class, then

$$S \equiv \Phi[\mathcal{P}_F^{(n)}]$$

for all $F \varepsilon \Omega'$.

PROOF: Let $g(B, x^{(n)})$ be the indicator function of B, i.e.

$$g(B, x^{(n)}) = \begin{cases} 1 \text{ for } x^{(n)} \varepsilon B, \\ 0 \text{ otherwise;} \end{cases}$$

then for each $A \varepsilon \mathfrak{B}$ and each $F \varepsilon \Omega'$,

$$\int_{R^{(n)}} [g(S^{-1}(A), x^{(n)}) - g(\Phi^{-1}(A), x^{(n)})]\, d\mathcal{P}_F^{(n)} = \mathcal{P}_F^{(n)}\{S^{-1}(A)\}$$
$$- \mathcal{P}_F^{(n)}\{\Phi^{-1}(A)\} = 0.$$

Since S and Φ are symmetric, $g(S^{-1}(B), x^{(n)})$ and $g(\Phi^{-1}(B), x^{(n)})$ are symmetric, and so is their difference. Because of the completeness property of Ω',

$$g(S^{-1}(B), x^{(n)}) - g(\Phi^{-1}(B), x^{(n)}) = 0$$

and $g(S^{-1}(B), x^{(n)}) = g(\Phi^{-1}(B), x^{(n)})[\mathcal{P}_F^{(n)}]$ for all $F \varepsilon \Omega'$. Consequently,

$$\mathcal{P}_F^{(n)}(S^{-1}(A) \Delta \Phi^{-1}(A)) = 0$$

for all $F \varepsilon \Omega'$ and all $A \varepsilon B$.

[Note: $E \Delta F = (E \cup F) - (E \cap F)$.]

But

$$\mathcal{P}_F^{(n)}(S \neq \Phi) = \mathcal{P}_F^{(n)}(S > \Phi) + \mathcal{P}_F^{(n)}(\Phi > S) \leqq$$

$$\sum_{m=-\infty}^{\infty} \sum_{k=1}^{\infty} [\mathcal{P}_F^{(n)}(S > (m/k); \Phi < (m/k)) + \mathcal{P}_F^{(n)}(S < (m/k); \Phi > (m/k))]$$

$$\leqq \sum_{m=-\infty}^{\infty} \sum_{k=1}^{\infty} [\mathcal{P}_F^{(n)}((S \leqq (m/k)) \Delta (\Phi \leqq (m/k)))] = 0 \quad \text{for all } F \in \Omega'.$$

Therefore $S = \Phi[\mathcal{P}_F^{(n)}]$ for all $F \in \Omega$. The main theorem now follows immediately.

THE MAIN THEOREM. *If S_G is a statistic in Ω w.r.t. Ω', then the property of being symmetric and strongly distribution-free is equivalent to having structure* (d), *whenever the following three conditions are fulfilled.*

(α) $\Omega \subset \Omega^*$;

(β) Ω' *is closed under* Ω; *and*

(γ) Ω' *is a symmetrically complete class.*

The next question is: Which classes of statistical interest satisfy the hypotheses of the main theorem?

4. Closed and complete classes. As was previously mentioned one can conclude from (i) that Ω_0, Ω_1, Ω_2 and Ω^* are closed under all subsets of Ω^*. Also, it can be proved that Ω_3, Ω_4, Ω_u and Ω_e do not satisfy that closure property. However, one can verify that Ω_3 is closed under $\Omega_3 \cap \Omega^*$; and that Ω_4 is closed under $\Omega_4 \cap \Omega^*$.

The work of Halmos [16]; Fraser ([14], [15], [1], pp. 23–31); Lehmann ([3], p. 132), and Bell-Blackwell-Breiman [8] establish the fact that Ω_0, Ω_1, Ω_2, Ω_3, Ω_4, Ω_u and Ω_e are symmetrically complete. (It should be mentioned here that a class of cpf's is symmetrically complete if and only if the order statistic is a complete statistic with respect to the class of power probability distributions of the given class of cpf's.)

Therefore, Ω_0, Ω_1, Ω_2, Ω_3 and Ω_4 satisfy both the completeness and closure hypotheses of the main theorem. Consequently, the following corollary to the main theorem is valid.

COROLLARY: *If S_G is a statistic in Ω w.r.t. Ω', then the property of being symmetric and strongly distribution-free is equivalent to having structure* (d) *for each of the following cases.*

(1) $\Omega \subset \Omega^*$ and $\Omega' = \Omega_0$;
(2) $\Omega \subset \Omega^*$ and $\Omega' = \Omega_1$;
(3) $\Omega \subset \Omega^*$ and $\Omega' = \Omega_2$;
(4) $\Omega \subset \Omega^*$ and $\Omega' = \Omega^*$;
(5) $\Omega = \Omega_3 \cap \Omega^*$ and $\Omega' = \Omega_3$; and
(6) $\Omega = \Omega_4 \cap \Omega^*$ and $\Omega' = \Omega_4$.

REFERENCES

[1] D. A. S. FRASER, *Non-parametric Methods in Statistics*, New York, John Wiley and Sons, 1957.

[2] M. LOÈVE, *Probability Theory*, New York, D. van Nostrand, 1955.
[3] E. LEHMANN, *Testing Statistical Hypotheses*, New York, John Wiley and Sons, 1959.
[4] T. W. ANDERSON AND D. A. DARLING, "Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 183–212.
[5] C. B. BELL, "Application of distribution-free statistics to some problems in missile design and production," Douglas (Aircraft Co.), Santa Monica Report SM 18396 (1954).
[6] C. B. BELL, "On the structure of algebras and homomorphisms," *Proc. Amer. Math. Soc.*, Vol. 7 (1956), pp. 483–492.
[7] C. B. BELL, "On the structure of stochastic independence," *Ill. J. Math.*, Vol. 2 (1958), pp. 415–424.
[8] C. B. BELL, D. BLACKWELL AND L. BREIMAN, "A note on the completeness of order statistics," *Ann. Math. Stat.*, Vol. 31 (1960), pp. 794–797.
[9] Z. W. BIRNBAUM, "Numerical tabulation of the distribution of Kolmogoroff's statistic for finite sample size," *J. Amer. Stat. Assoc.*, Vol. 47 (1952), pp. 425–441.
[10] Z. W. BIRNBAUM AND F. H. TINGEY, "One-sided confidence contours for probability distribution functions," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 592–596.
[11] Z. W. BIRNBAUM, "Distribution-free tests for continuous distribution functions," *Ann. Math. Stat*, Vol. 24 (1953), pp. 1–8.
[12] Z. W. BIRNBAUM AND H. RUBIN, "On distribution-free statistics," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 593–598.
[13] W. FELLER, "On the Kolmogoroff-Smirnoff limit theorems for empirical distributions," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 177–189.
[14] D. A. S. FRASER, "Completeness of order statistics," *Can. J. Math.*, Vol. 6 (1953), pp. 42–45.
[15] D. A. S. FRASER, "Non-parametric theory: scale and location parameters," *Can. J. Math.*, Vol. 6 (1953), pp. 46–68.
[16] P. R. HALMOS, "The theory of unbiased estimation," *Ann. Math. Stat.*, Vol. 17 (1946), pp. 34–43.
[17] B. F. KIMBALL, "Some basic theorems for developing tests of fit for the case of the non-parametric probability distribution function, I," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 540–548.
[18] H. SCHEFFÉ, "On a measure problem arising in the theory of non-parametric tests," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 227–233.
[19] B. SHERMAN, "A random variable related to the spacing of sample values," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 339–361.
[20] N. SMIRNOV, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 279–281.
[21] A. WALD AND J. WOLFOWITZ, "Confidence limits for continuous distribution functions," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 105–118.