ON DISTRIBUTION-FREE STATISTICS[1]

By Z. W. Birnbaum and H. Rubin

University of Washington and Stanford University

1. Introduction. 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 F. It has been observed [1] that, to the authors' knowledge, all distribution-free statistics considered in the past 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 $\{U: 0 \leqq U_i \leqq 1, i = 1, 2, \cdots, n\}$. It is the purpose of this paper to study the relationship between the class of statistics which can be written in this particular form and the class of distribution-free statistics.

2. Distribution-free statistics and statistics of structure (*d*). Let Ω and Ω' be two families of cumulative probability functions. A real quantity $W = S(X_1, X_2, \cdots, X_n, G)$ will be called a *statistic in* Ω *with regard to* Ω' if, for any $G \varepsilon \Omega$, $F \varepsilon \Omega'$, and $X_1, X_2, \cdots, X_n$ in the n-dimensional sample-space for a random variable X which has the cumulative probability function F,

(i) $S(X_1, X_2, \cdots, X_n, G)$ is defined almost everywhere in the sample-space $X_1, X_2, \cdots, X_n$ (i.e. with the possible exception of a set of probability zero), and

(ii) $W = S(X_1, X_2, \cdots, X_n, G)$ has a probability distribution; this probability distribution will be denoted by $\mathcal{P}(W; F) = \mathcal{P}[S(X_1, X_2, \cdots, X_n, G); F]$.

For example, Kolmogorov's statistic

$$D_n = \sup_{-\infty < x < \infty} | F_n(x) - G(x) |, \tag{2.1}$$

where F_n is the empirical cumulative distribution function determined by the sample $X_1, X_2, \cdots, X_n$, satisfies (i) and (ii) when $\Omega = \Omega' = \Omega_1$, the class of all nondegenerate cumulative probability functions[2], hence D_n is a statistic in Ω_1 with regard to Ω_1.

If for a statistic $S(X_1, X_2, \cdots, X_n, G)$ in Ω with regard to Ω' there exists a function Φ defined on the n-dimensional unit cube and symmetric in its arguments, such that for any $G \varepsilon \Omega$, $F \varepsilon \Omega'$ we have

$$S(X_1, X_2, \cdots, X_n, G) = \Phi[G(X_1), G(X_2), \cdots, G(X_n)]$$

Received 8/6/53.

[1] Work done under the sponsorship of the Office of Naval Research.

[2] The notations for various classes of cumulative probability functions are those introduced by Scheffé [2].

almost everywhere[3] in the sample space $X_1, X_2, \cdots, X_n$ for the random variable X which has the cumulative probability function F, then we shall say that $S(X_1, X_2, \cdots, X_n, G)$ is *a statistic of structure* (d).

Kolmogorov's statistic (2.1) is an example of a statistic of structure (d), since it can be written as

$$D_n = \max_{i=1,\cdots,n} \left\{ \max \left[G(X_i') - \frac{i-1}{n}, \frac{i}{n} - G(X_i') \right] \right\},$$

where $X_1', X_2', \cdots, X_n'$ are the numbers $X_1, X_2, \cdots, X_n$, ordered increasingly.

If $\Omega = \Omega'$ and the statistic $S(X_1, X_2, \cdots, X_n, G)$ has the property that the probability distribution $\mathcal{P}[S(X_1, X_2, \cdots, X_n, G); G]$ is independent of G for $G \varepsilon \Omega$, we shall say that $S(X_1, X_2, \cdots, X_n, G)$ is a *distribution-free statistic in* Ω.

Let us now assume $\Omega = \Omega' = \Omega_2$, the class of all continuous cumulative probability functions. Denoting by R the rectangular distribution in (0, 1) and by $U_1, \cdots, U_n$ a sample of size n of a random variable with distribution R we have

$$\mathcal{P}\{\Phi[G(X_1), \cdots, G(X_n)]; G\} = \mathcal{P}\{\Phi(U_1, \cdots, U_n); R\}.$$

It follows that *if a statistic in* Ω_2 *with regard to* Ω_2 *has structure* (d) *then it is distribution-free in* Ω_2.

All distribution-free statistics considered in literature happen to have structure (d), with $\Omega = \Omega' = \Omega_2$. Nevertheless the conjecture that every distribution-free statistic, symmetric in $X_1, X_2, \cdots, X_n$, with $\Omega = \Omega' = \Omega_2$, must have structure (d) is not true. This can be seen from the following counter-example.

Let ω_1 and ω_2 be nonempty, mutually exclusive subsets of Ω_2 such that $\omega_1 \cup \omega_2 = \Omega_2$. Denoting by F_n again the empirical cumulative distribution function determined by a sample of size n, we define

$$S = \begin{cases} \sup\limits_{-\infty<x<\infty} [F(x) - F_n(x)] = S_1, & \text{if } F \varepsilon \omega_1 \\ \sup\limits_{-\infty<x<\infty} [F_n(x) - F(x)] = S_2, & \text{if } F \varepsilon \omega_2. \end{cases}$$

Since S_1 and S_2 are distribution-free statistics with the same probability distribution, S is a distribution-free statistic. It is, however, clearly not a statistic of structure (d).

3. Strongly distribution-free statistics.

Let Ω^* be the family of all continuous cumulative probability functions such that if $G \varepsilon \Omega^*$ then G is strictly increasing at all x for which $0 < G(x) < 1$. Clearly if $G \varepsilon \Omega^*$ then the inverse function $G^{(-1)}$ is defined on the open unit interval.

[3] The exceptional set of probability zero may depend on G.

We now consider a statistic $S(X_1, X_2, \cdots, X_n, G)$ in Ω^* with regard to some family Ω' of cumulative probability functions. This statistic shall be called *strongly distribution-free in* Ω^* *with regard to* Ω' if the probability distribution $\mathcal{P}[S(X_1, X_2, \cdots, X_n, G); F]$ depends only on the function $\tau = FG^{(-1)}$ for all $G \varepsilon \Omega^*$, $F \varepsilon \Omega'$.

It is easily seen that, for $\Omega' = \Omega^*$, *a strongly distribution-free statistic is distribution-free*. For if $\mathcal{P}[S(X_1, X_2, \cdots, X_n, G); F]$ depends only on $FG^{(-1)}$ for all $F, G \varepsilon \Omega^*$, then in particular $\mathcal{P}[S(X_1, X_2, \cdots, X_n, G); G]$ depends only on $GG^{(-1)} = I$, hence is independent of G. One also verifies immediately that *if a statistic in* Ω^* *with regard to* Ω^* *has structure* (*d*) *then it is strongly distribution-free*, since then $\mathcal{P}\{\Phi[G(X_1), G(X_2), \cdots, G(X_n)]; F\} = \mathcal{P}\{\Phi[U_1, U_2, \cdots, U_n]; FG^{(-1)}\}$.

Since all practically important distribution-free statistics are symmetric in $X_1, X_2, \cdots, X_n$ and strongly distribution-free, as well as of structure (*d*), one again may conjecture that under some fairly general assumptions these two properties are equivalent. This conjecture is found to be correct for $\Omega = \Omega' = \Omega^*$. We have already seen that if a statistic has structure (*d*) it is strongly distribution-free; it remains only to prove the converse statement:

THEOREM. *If a statistic* $W = S(X_1, X_2, \cdots, X_n, G)$ *in* Ω^* *with regard to* Ω^* *is symmetric in* $X_1, X_2, \cdots, X_n$ *and strongly distribution-free, then it has structure* (*d*).

The proof of this theorem makes use of a lemma which will be presented in the next section.

4. LEMMA. *Let* H *be a strictly increasing continuous function on the closed unit-interval, such that* $H(0) = 0$, $H(1) = 1$; μ_H *the measure defined by* H *on the unit-interval* I_1; $\mu_H^{(n)}$ *the corresponding product-measure on the n-dimensional unit-cube* I_n. *Then, for any set* $M \subset I_n$ *with* $\mu_H^{(n)}(M) > 0$ *and any* $\epsilon > 0$, *there exist sets* $Q_1, Q_2, \cdots, Q_n$ *in* I_1 *such that*

(i) $Q_1, Q_2, \cdots, Q_n$ *are disjoint*, μ_H*-measurable, with* $\mu_H(Q_i) > 0$ *for* $i = 1, 2, \cdots, n$;

(ii) *for* $Q_0 = Compl. \bigcup_{i=1}^{n} Q_i$ *we have* $\mu_H(Q_0) > 0$;

(iii) *if* Q_i *is placed on the* y_i*-axis*, $i = 1, 2, \cdots, n$, *then the product-set* $Q = Q_1 \times Q_2 \times \cdots \times Q_n$ *in* I_n *has the property*

$$\mu_H^{(n)}(Q \cap M)/\mu_H^{(n)}(Q) > 1 - \epsilon.$$

PROOF. It may be assumed without loss of generality that $H(y) = y$, so that μ_H and $\mu_H^{(n)}$ are Lebesgue measures. Let $C_{\eta, y_1, \cdots, y_n}$ denote the cube $| Y_i - y_i | < \eta$ in the $(Y_1, Y_2, \cdots, Y_n)$ space, with center $(y_1, y_2, \cdots, y_n)$ and volume $\mu_H^{(n)}(C_{\eta, y_1, \cdots, y_n}) = (2\eta)^n$.

It is well known that

$$\lim_{\eta \to 0} (2\eta)^{-n} \mu_H^{(n)}(M \cap C_{\eta, y_1, \cdots, y_n}) = 1 \tag{4.1}$$

for almost all points in M (see e.g. [3] p. 129). The subset of those points of M for which no two coordinates are equal and none is 0 or 1 has the same measure as M. Let M_1 be the set of all points of M for which (4.1) holds and which have no two coordinates equal and no coordinate 0 or 1. Then $\mu_H^{(n)}(M_1) = \mu_H^{(n)}(M) > 0$. Let $y_1^0, \cdots, y_n^0$ be a point in M_1, and let

$$\lambda = \min \{\min_{(i)} y_i^0, \min_{(i)} (1 - y_i^0), \min_{i \neq j} | y_i^0 - y_j^0 |\}.$$

Clearly $0 < \lambda < \frac{1}{2}$, and for $0 < \eta < \lambda/2$ the intervals

$$Q_i\colon (y_i^0 - \eta, y_i^0 + \eta), \qquad i = 1, 2, \cdots, n, \tag{4.2}$$

are all in I_1 and satisfy (i) and (ii). If Q_i is placed on the Y_i-axis then the product-set $Q = Q_1 \times Q_2 \times \cdots \times Q_n$ is the cube $C_{\eta, y_1^0, \cdots, y_n^0}$. According to (4.1) there exists an $\eta_0 > 0$ such that

$$(2\eta)^{-n} \mu_H^{(n)}(M \cap C_{\eta, y_1^0, \cdots, y_n^0}) > 1 - \epsilon$$

for $\eta < \eta_0$. Choosing $\eta < \min(\eta_0, \lambda/2)$ and constructing the intervals (4.2) one obtains the Q_i required by the lemma.

5. Proof of theorem. When the random variable X has the cumulative probability function F, the random variable $Y = G(X)$ has the cumulative probability function $H = FG^{(-1)}$. Setting $Y_i = G(X_i)$ we, therefore, have

$$W = S(X_1, \cdots, X_n, G) = S[G^{(-1)}(Y_1), \cdots, G^{(-1)}(Y_n), G]$$

and

$$\begin{aligned} \mathcal{P}[S(X_1, \cdots, X_n, G); F] &= \mathcal{P}\{S[G^{(-1)}(Y_1), \cdots, G^{(-1)}(Y_n), G]; FG^{(-1)}\} \\ &= \mathcal{P}\{S[G^{(-1)}(Y_1), \cdots, G^{(-1)}(Y_n), G]; H\}. \end{aligned}$$

By assumption, this last probability distribution depends only on the cumulative probability function H, and not on G. From this and the symmetry assumption we wish to conclude that $S[G^{(-1)}(Y_1), \cdots, G^{(-1)}(Y_n), G]$ can be written in the form of a function $\Phi(Y_1, \cdots, Y_n)$, independent of G except on a set of H-measure zero.

To prove this, we assume that for some $G_1, G_2 \varepsilon \Omega^*$ we have $S[G_1^{(-1)}(Y_1), \cdots, G_1^{(-1)}(Y_n), G_1] \neq S[G_2^{(-1)}(Y_1), \cdots, G_2^{(-1)}(Y_n), G_2]$ on a set of positive H-measure. Without loss of generality we may assume

$$\begin{aligned} \infty > k > S[G_1^{(-1)}(Y_1), \cdots, G_1^{(-1)}(Y_n), G_1] & \\ - S[G_2^{(-1)}(Y_1), \cdots, G_2^{(-1)}(Y_n), G_2] &> \eta > 0 \end{aligned} \tag{5.1}$$

on a set M in the unit cube I_n, where M is symmetric and has positive measure. For any H, continuous and strictly increasing in I_1, and any $\epsilon > 0$, we construct sets $Q_1, Q_2, \cdots, Q_n$ according to the lemma in Section 4 and have

$$\mu_H^{(n)}(Q \cap M)/\mu_H^{(n)}(Q) > 1 - \epsilon. \tag{5.2}$$

For any

$$\alpha_i > 0, \quad i = 0, 1, \cdots, n, \quad \text{and} \quad \alpha_0 + \sum_{i=1}^{n} \alpha_i = 1 \tag{5.3}$$

we define the set function

$$K_{\alpha_1, \cdots, \alpha_n}(T) = \sum_{j=0}^{n} \alpha_j \frac{\mu_H(T \cap Q_j)}{\mu_H(Q_j)}$$

for any measurable $T \subset I_1$. This clearly is a probability measure in I_1. Taking for T the interval $(0, y)$ we obtain a strictly increasing continuous cumulative probability function which will be denoted by $K_{\alpha_1, \cdots, \alpha_n}$.

Without loss of generality, S may be assumed bounded, since otherwise we could consider $S/(1 + |S|)$. This assures the existence of the mathematical expectation of S. Since $S[G_1^{(-1)}(Y_1), \cdots, G_1^{(-1)}(Y_n), G_1]$ and $S[G_2^{(-1)}(Y_1), \cdots, G_2^{(-1)}(Y_n), G_2]$ have the same probability distribution if $Y_1, Y_2, \cdots, Y_n$ are a sample of a random variable Y with the cumulative probability function $K_{\alpha_1, \cdots, \alpha_n}$, their mathematical expectations are equal

$$\begin{aligned} E\{S[G_1^{(-1)}(Y_1), \cdots, G_1^{(-1)}(Y_n), G_1] \\ - S[G_2^{(-1)}(Y_1), \cdots, G_2^{(-1)}(Y_n), G_2]; K_{\alpha_1, \cdots, \alpha_n}\} = 0. \end{aligned} \tag{5.4}$$

Using the abbreviations

$$S[G_i^{(-1)}(Y_1), \cdots, G_i^{(-1)}(Y_n), G_i] = S_i(Y_1, \cdots, Y_n), \qquad i = 1, 2,$$

we write the left-hand side of (5.4) explicitly

$$\begin{aligned} &\int_{Y_1=0}^{1} \cdots \int_{Y_n=0}^{1} [S_1(Y_1, \cdots, Y_n) - S_2(Y_1, \cdots, Y_n)] \prod_{i=1}^{n} dK_{\alpha_1, \cdots, \alpha_n}(Y_i) \\ &= \sum_{j_1=0}^{n} \cdots \sum_{j_n=0}^{n} \int_{Y_1 \varepsilon Q_{j_1}} \cdots \int_{Y_n \varepsilon Q_{j_n}} [S_1(Y_1, \cdots, Y_n) - S_2(Y_1, \cdots, Y_n)] \\ &\quad \prod_{i=1}^{n} dK_{\alpha_1, \cdots, \alpha_n}(Y_i) \\ &= \sum_{j_1=0}^{n} \cdots \sum_{j_n=0}^{n} \frac{\alpha_{j_1} \cdots \alpha_{j_n}}{\mu_H(Q_{j_1}) \cdots \mu_H(Q_{j_n})} \\ &\quad \cdot \int_{Q_{j_1}} \cdots \int_{Q_{j_n}} [S_1(Y_1, \cdots, Y_n) - S_2(Y_1, \cdots, Y_n] \, dH(Y_n) \cdots dH(Y_1). \end{aligned} \tag{5.5}$$

Since $S_1(Y_1, \cdots, Y_n)$, $S_2(Y_1, \cdots, Y_n)$ and M are symmetric in $Y_1, \cdots, Y_n$, all the terms of the sum which correspond to different permutations of the same n subscripts $j_1, \cdots, j_n$ (out of the $n + 1$ possible values $0, 1, \cdots, n$) are equal.

Collecting these equal terms, we obtain a polynomial in $\alpha_0, \alpha_1, \cdots, \alpha_n$, which according to (5.4) vanishes identically under the restrictions (5.3). It follows that each of the integrals in the last term of (5.5) must vanish, and in particular

$$\int_{Q_1}\int_{Q_2}\cdots\int_{Q_n}[S_1(Y_1, Y_2, \cdots, Y_n) - S_2(Y_1, Y_2, \cdots Y_n)]\, dY_n \cdots dY_2\, dY_1 = 0;$$

which, for ϵ sufficiently small, contradicts (5.1) and (5.2).

REFERENCES

[1] Z. W. Birnbaum, "Distribution-free tests of fit for continuous distribution functions," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 1–8.
[2] H. Scheffé, "On a measure problem arising in the theory of nonparametric tests," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 227–233.
[3] S. Saks, "Theory of the integral," *Monografie Matematyczne*, Vol. 7, Warszawa-Lwow, 1937.