EXACT AND LIMITING PROBABILITY DISTRIBUTIONS OF SOME SMIRNOV TYPE STATISTICS

Miklós Csörgő *

(received January 2, 1964)

1. Summary. Let $F(x)$ be the continuous distribution function of a random variable X and $F_n(x)$ be the empirical distribution function determined by a random sample $X_1, \ldots, X_n$ taken on X. Using the method of Birnbaum and Tingey [1] we are going to derive the exact distributions of the random variables

$$\sup_{F(x) \leq b} (F(x) - F_n(x)), \quad \sup_{a \leq F(x)} (F_n(x) - F(x)), \quad \sup_{F_n(x) \leq b} (F(x) - F_n(x))$$

and $\sup_{a \leq F_n(x)} (F_n(x) - F(x))$, where $0 < a < 1$, $0 < b < 1$ and where the indicated sup' s are taken over all x' s such that $-\infty < x < x_b$ and $x_a \leq x < +\infty$ with $F(x_b) = b$, $F(x_a) = a$ in the first two cases and over all x' s so that $F_n(x) \leq b$ and $a \leq F_n(x)$ in the last two cases. We are also going to discuss briefly the asymptotic behaviour of these random variables and the consistency of the relevant statistical tests.

2. Introduction. Let $Y = F(x)$. Then Y is a uniformly distributed random variable on $(0, 1)$ and we have $Y_1, \ldots, Y_n$ as a random sample on Y resulting from this transformation. Let $Y_{(1)} < \ldots < Y_{(n)}$ be the corresponding order statistics which determines the empirical distribution function

* This research was partially supported by the Office of Naval Research 042-023, Contract Nonr 1858(05) to Princeton University.

$$(2.1) \qquad G_n(Y) = \begin{cases} 0 & \text{for } Y < Y_{(1)} \\ \frac{k}{n} & \text{for } Y_{(k)} \le Y < Y_{(k+1)} \\ 1 & \text{for } Y_{(n)} \le Y . \end{cases}$$

We are going to need the following result of Birnbaum and Tingey:

$$(2.2) \quad P\{ \sup_{-\infty < x < +\infty} (F(x) - F_n(x)) < \varepsilon \} = P\{ \sup_{0 < Y < 1} (Y - G_n(Y)) < \varepsilon \}$$

$$= n! \int_{Y_{(0)}=0}^{\varepsilon} \int_{Y_{(1)}}^{\frac{1}{n}+\varepsilon} \cdots \int_{Y_{(k)}}^{\frac{k}{n}+\varepsilon} \int_{Y_{(k+1)}}^{1} \cdots \int_{Y_{(n-2)}}^{1} \int_{Y_{(n-1)}}^{1}$$

$$dY_{(n)} dY_{(n-1)} \dots dY_{(k+2)} dY_{(k+1)} \dots dY_{(2)} dY_{(1)}$$

$$= 1 - \sum_{j=0}^{k} T_{j,n}(\varepsilon) ,$$

where $T_{j,n}(\varepsilon) = \binom{n}{j}(1-\varepsilon-\frac{j}{n})^{n-j}(e+\frac{j}{n})^{j-1}\varepsilon$, $k = [n(1-\varepsilon)]$ = greatest integer contained in $n(1-\varepsilon)$ and $0 < \varepsilon \le 1$.

3. Exact distributions of random variables of section 1. Using the notation of sections 1 and 2 we are going to prove:

THEOREM 1.

$$(3.1) \quad P\{ \sup_{F(x) \le b} (F(x) - F_n(x)) < \varepsilon \} = P\{ \sup_{0 < Y \le b} (Y - G_n(Y)) < \varepsilon \}$$

$$= 1 - \sum_{j=0}^{k} T_{j,n}(\varepsilon) = N_1(\varepsilon, b, n) ,$$

where $k = [n(b-\varepsilon)]$ with $0 < \varepsilon \le b$.

COROLLARY 1.

$$(3.2)\quad P\{\sup_{a \le F(x)} (F_n(x) - F(x)) < \varepsilon\} = P\{\sup_{a \le Y < 1} (G_n(Y) - Y) < \varepsilon\}$$

$$= N_1'(\varepsilon, a, n),$$

where $N_1'(.)$ is obtained by putting $b = 1-a$ in $N_1(.)$ of Theorem 1.

The statement of Corollary 1 follows immediately from Theorem 1 after putting $b = 1-a$ and replacing $1-F(x)$ by $F(x)$ and $1-F_n(x)$ by $F_n(x)$ in it. We also note here that Corollaries 2, 3 and 4 as stated below follow exactly the same way from their respective preceding theorems.

THEOREM 2.

$$(3.3)\quad P\{\sup_{F_n(x) \le b} (F(x) - F_n(x)) < \varepsilon\} = P\{\sup_{0 < G_n(y) \le b} (Y - G_n(Y)) < \varepsilon\}$$

$$= 1 - \sum_{j=0}^{k} T_{j,n}(\varepsilon) = N_2(\varepsilon, b, n),$$

where $k = \min\{[nb], [n(1-\varepsilon)]\}$, $0 < \varepsilon \le 1$.

COROLLARY 2.

$$(3.4)\quad P\{\sup_{a \le F_n(x)} (F_n(x) - F(x)) < \varepsilon\} = P\{\sup_{a \le G_n(Y) < 1} (G_n(Y) - Y) < \varepsilon\}$$

$$= N_2'(\varepsilon, a, n),$$

where $N_2'(.)$ is obtained by putting $b = 1-a$ in $N_2(.)$ of Theorem 2.

<u>Proof of Theorem 1.</u> It is clear that the distribution of the random variable $\sup_{F(x) \le b} (F(x) - F_n(x))$ is the same as that

https://doi.org/10.4153/CMB-1965-012-5 Published online by Cambridge University Press

of $\sup_{Y \leq b} (Y - G_n(Y))$, and saying that

$$\sup_{Y \leq b} (Y - G_n(Y)) < \varepsilon \tag{3.5}$$

is equivalent to saying: $Y < G_n(Y) + \varepsilon$ for all $Y \leq b$. From the definition of $G_n(Y)$ it follows that $Y < G_n(Y) + \varepsilon$ for all $Y \leq b$ occurs if and only if the ordered random sample

$$0 < Y_{(1)} < Y_{(2)} < \ldots < Y_{(n)} < 1 \tag{3.6}$$

falls into the region

$$\begin{aligned} &Y_{(j-1)} < Y_{(j)} < \frac{j-1}{n} + \varepsilon \quad \text{for } j = 1, 2, \ldots, k+1 , \\ &Y_{(j-1)} < Y_{(j)} < 1 \qquad\qquad \text{for } j = k+2, \ldots, n , \end{aligned} \tag{3.7}$$

where $Y_{(0)} \equiv 0$ and k is the greatest integer so that

$$\frac{k}{n} + \varepsilon \leq b , \tag{3.8}$$

(that is $k = [n(b-\varepsilon)]$ with $0 < \varepsilon \leq b$) .

The density function of (3.6) is given by

$$p(Y_{(1)}, \ldots, Y_{(n)}) = n!\, dY_{(1)} \ldots dY_{(n)} \tag{3.9}$$

and thus the probability that (3.6) falls into the region (3.7) is given by the last two lines of (2.2) with $k = [n(b-\varepsilon)]$ and $0 < \varepsilon \leq b$. This completes the proof of Theorem 1.

The proof of Theorem 2 is exactly the same as that of Theorem 1. To indicate its main lines we have there that $Y < G_n(Y) + \varepsilon$ for all Y such that $G_n(Y) \leq b$ occurs if and only if (3.6) falls into region (3.7) where k is now defined as the greatest integer such that $\frac{k}{n} + \varepsilon \leq b + \varepsilon \leq 1$, that is

https://doi.org/10.4153/CMB-1965-012-5 Published online by Cambridge University Press

$k = [nb] \leq [n(1-\varepsilon)]$ with $0 < \varepsilon \leq 1$. Thus $k = \min\{[nb], [n(1-\varepsilon)]\}$. From here we can proceed exactly the same way as we did above when proving Theorem 1.

4. Limiting distributions. If we put $\varepsilon = \frac{\lambda}{\sqrt{n}}$ in theorems 1 - 2 of section 3 then the following statements hold:

THEOREM 3.

$$\lim_{n\to\infty} N_1\left(\frac{\lambda}{\sqrt{n}}, b, n\right) = \phi_1(\lambda, b), \tag{4.1}$$

where

$$\phi_1(\lambda, b) = 1/\sqrt{2\pi} \int_{-\infty}^{\alpha} e^{-t^2/2}\, dt - (e^{-2\lambda^2}/\sqrt{2\pi}) \int_{-\infty}^{\beta} e^{-t^2/2}\, dt,$$

and $\alpha = \frac{\lambda}{\sqrt{b(1-b)}}$, $\beta = \frac{\lambda - 2\lambda(1-b)}{\sqrt{b(1-b)}}$. We note here that when $b = 1$, that is when $\alpha = \beta = +\infty$, then we have $\phi_1(\lambda, 1) = 1 - e^{-2\lambda^2}$, the original theorem of Smirnov [4].

COROLLARY 3.

$$\lim_{n\to\infty} N_1'\left(\frac{\lambda}{\sqrt{n}}, a, n\right) = \phi_2(\lambda, a), \tag{4.2}$$

where

$$\phi_2(\lambda, a) = 1/\sqrt{2\pi} \int_{-\infty}^{\delta} e^{-t^2/2}\, dt - (e^{-2\lambda^2}/\sqrt{2\pi}) \int_{-\infty}^{\gamma} e^{-t^2/2}\, dt$$

and $\delta = \frac{\lambda}{\sqrt{a(1-a)}}$, $\gamma = \frac{\lambda - 2\lambda a}{\sqrt{a(1-a)}}$. When $a = 0$, that is when $\delta = \gamma = +\infty$, then $\phi_2(\lambda, 0) = 1 - e^{-2\lambda^2}$, the above quoted Smirnov theorem again.

THEOREM 4.

$$\lim_{n\to\infty} N_2\left(\frac{\lambda}{\sqrt{n}}, b, n\right) = \phi_1(\lambda, b), \tag{4.3}$$

where $\phi_1(\lambda, b)$ is as it was defined in Theorem 3. Thus Theorems 1 and 2 are equivalent in the limit.

COROLLARY 4.

$$\lim_{n\to\infty} N_2'\left(\frac{\lambda}{\sqrt{n}}, a, n\right) = \phi_2(\lambda, a), \tag{4.4}$$

where $\phi_2(\lambda, a)$ is as it was defined in Corollary 3. Thus Corollaries 1 and 2 are equivalent in the limit.

Having got the explicit forms of Theorems 1 and 2, a natural way to derive theorems 3 and 4 would be through making use of Stirling's approximation for large factorials and some change-of-variable techniques. In fact we would have to prove only Theorem 3 this way, for we are going to show that Theorem 3 implies Theorem 4. Thus we will have to have an actual derivation for Theorem 3 only.

Theorem 3 itself could also be verified through manipulations with generating functions and their limiting forms, the Laplace transforms, the way Feller proved the Kolmogorov-Smirnov theorems in [2].

None of these ways of proof is simple and they are definitely not short. However, we can get Theorem 3 and its corollary as immediate by-products of a theorem of Manija, which we are going to quote here. Using the method of Feller's paper [2], he proved the following theorem:

THEOREM (Manija [3]):

$$\lim_{n\to\infty} P\left\{ \sup_{a \le F(x) \le b} (F(x) - F_n(x)) < \frac{\lambda}{\sqrt{n}} \right\} = \phi(a, b; \lambda),$$

where $0 < a < b < 1$ and

https://doi.org/10.4153/CMB-1965-012-5 Published online by Cambridge University Press

$$\phi(a,b;\lambda) = C \int_{-\infty}^{\delta}\int_{-\infty}^{\alpha} e^{-\frac{1}{2}\bar{\theta}(t_1,t_2)} dt_1 dt_2 - Ce^{-2\lambda^2} \int_{-\infty}^{\gamma}\int_{-\infty}^{\beta} e^{-\frac{1}{2}\theta(t_1,t_2)} dt_1 dt_2,$$

where $C = 1/2\pi\sqrt{1-R^2}$, $R = \sqrt{\frac{a(1-b)}{b(1-a)}}$, $\theta(t_1,t_2) = 1/(1-R^2)[t_1^2+2Rt_1t_2+t_2^2]$, $\bar{\theta}(t_1,t_2) = 1/(1-R^2)[t_1^2-2Rt_1t_2+t_2^2]$, and $\alpha, \beta, \delta, \gamma$ are as defined in Theorem 3 and Corollary 3.

If, in the above theorem, $a = 0$ we immediately get Theorem 3 and, when $b = 1$, Corollary 3 is gained. We remark here that we can actually equate a to zero and b to one in Manija's theorem, for Feller's method of proof does not require the restriction $0 < a < b < 1$ and is valid for $a = 0$ or $b = 1$.

It remains to show that Theorem 3 implies Theorem 4. To do this, let us consider the event $|Y - G_n(Y)| \leq \delta$, where $\delta > 0$ and is arbitrarily small. In case of Theorem 7 we have that $0 < G_n(Y) \leq b$ and thus it follows that $|Y-b| \leq \delta$ or $|Y-b| \geq \delta$. The second case can only result from $Y-b \leq -\delta$, and this together with $-\delta \leq Y - G_n(Y)$ implies that $G_n(Y) \leq Y + \delta \leq b$; thus

$$\text{(4.5)} \qquad \sup_{Y \leq b-\delta} (Y - G_n(Y)) \leq \sup_{G_n(Y) \leq b} (Y - G_n(Y)) .$$

Let A be the event that $\sup_{G_n(Y) \leq b} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}$, and let A' be the event that $\sup_{Y \leq b-\delta} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}$. Then, by (4.5), $A \subseteq A'$, and if we let B be the event $|Y - G_n(Y)| \leq \delta$, then $AB \subseteq A'B$. Thus

$$A = AB^C \cup AB \subseteq B^C \cup A'B \subseteq B^C \cup A' ,$$

where B^C denotes the complementary event of B. Therefore $P(A) \leq P(B^C) + P(A')$, that is

https://doi.org/10.4153/CMB-1965-012-5 Published online by Cambridge University Press

$$P\{\sup_{G_n(Y) \leq b} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}\}$$

(4.6)

$$\leq P\{|Y - G_n(Y)| > \delta\} + P\{\sup_{Y \leq b-\delta} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}\}.$$

It can be similarly shown that

$$P\{\sup_{Y \leq b+\delta} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}\}$$

(4.7)

$$\leq P\{|Y - G_n(Y)| > \delta\} + P\{\sup_{G_n(Y) \leq b} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}\}.$$

We also have

(4.8) $$\lim_{n \to \infty} P\{|Y - G_n(Y)| > \delta\} = 0$$

and Theorem 3 states that

(4.9) $$\lim_{n \to \infty} P\{\sup_{Y \leq b} (Y - G_n(Y)) < \frac{\lambda}{\sqrt{n}}\} = \phi_1(\lambda, b).$$

It follows then from (4.6), (4.7), (4.8) and (4.9) that

(4.10)
$$\lim_{n \to \infty} \sup N_2(\frac{\lambda}{\sqrt{n}}, b, n) \leq \phi_1(\lambda, b-\delta),$$
$$\lim_{n \to \infty} \inf N_2(\frac{\lambda}{\sqrt{n}}, b, n) \geq \phi_1(\lambda, b+\delta).$$

Since δ can be chosen arbitrarily small, and an integral is a continuous function of its upper limit, it follows that

$$\lim_{n \to \infty} N_2(\frac{\lambda}{\sqrt{n}}, b, n) = \phi_1(\lambda, b),$$

and this terminates the proof of Theorem 4.

Theorems 1 - 4 and their corollaries provide statistical tests or one-sided confidence contours for unknown continuous distribution functions when we would want to work with truncated theoretical or empirical distribution functions.

5. Consistency. Let us consider the null hypothesis $H_o: F(x) = F_o(x)$ which we would like to test against the alternative $H_1: F(x) = F_1(x)$, where $F_o(x)$ is a given continuous distribution function, $F_1(x)$ is continuous too and satisfies the relation

$$(5.1) \qquad \sup_{F_o(x) \leq b} (F_o(x) - F_1(x)) = d > 0 ,$$

and let $x_o \leq x_b$, where $F_o(x_b) = b$, be a value of x such that

$$(5.2) \qquad F_o(x_o) - F_1(x_o) = d .$$

We are going to use the test-statistic of Theorem 1 to test this statistical hypothesis. The critical region of this test is defined by

$$(5.3) \qquad P\{ \sup_{F_o(x) \leq b} (F_o(x) - F_n(x)) \geq \varepsilon_{n,\alpha} \} \leq \alpha ,$$

where $\varepsilon_{n,\alpha}$ is chosen as the smallest positive number such that (5.3) holds and can be found from $N_1(\varepsilon, b, n)$ of Theorem 1.

To show consistency of this test against the class of alternatives specified in (5.1), we take $\varepsilon_{n,\alpha} = \lambda_\alpha / \sqrt{n}$ where λ_α is such that

$$(5.4) \qquad \lim_{n \to \infty} P\{ \sup_{F_o(x) \leq b} (F_o(x) - F_n(x)) < \lambda_\alpha / \sqrt{n} \} = 1 - \alpha ,$$

and can be found from (4.3). Thus we have

$$(5.5)\quad \lim_{n\to\infty} P\{\sup_{F_o(x)\le b}(F_o(x)-F_n(x)) \ge \lambda_\alpha/\sqrt{n}\} = \alpha ,$$

and the test is called consistent if

$$(5.6)\quad \lim_{n\to\infty} P\{\sup_{F_o(x)\le b}(F_o(x)-F_n(x)) \ge \lambda_\alpha/\sqrt{n}\,|\,F_1(x)\} = 1 .$$

Using relation (5.2) we have

$$(5.7)\quad \begin{aligned} &P\{\sup_{F_o(x)\le b}(F_o(x)-F_n(x)) \ge \lambda_\alpha/\sqrt{n}\,|\,F_1(x)\} \\ &\ge P\{F_o(x_o)-F_n(x_o) \ge \lambda_\alpha/\sqrt{n}\,|\,F_1(x)\} \\ &= P\{F_n(x_o)-F_1(x_o) \le d - \lambda/\sqrt{n}\} \end{aligned}$$

and thus, taking limits on both sides of (5.7), we get (5.6), that is, consistency, as a straightforward consequence of the weak convergence of the individual sample quantiles to the corresponding true quantiles.

Consistency of a possible statistical test based on Corollary 1 can be shown similarly. We have shown that in the limit Theorem 2 and Corollary 2 with $\varepsilon = \lambda/\sqrt{n}$, are equivalent to Theorem 1 and Corollary 1 respectively, and so the statistical tests based on them are the same asymptotically as the ones treated above.

6. Acknowledgments. Theorems 1 and 2 and their corollaries owe their present simple forms to a suggestion made by the referee of this paper. I take this opportunity to thank him for this and other valuable remarks. I would also like to thank my colleagues Professors H. Furstenberg and K. M. Rao for valuable discussions on some parts of this paper.

REFERENCES

1. Z. W. Birnbaum and Fred H. Tingey (1951), One-sided confidence contours for probability distribution functions, Ann. Math. Statist. 22, pp. 592-596.

2. William Feller (1948), On the Kolmogorov-Smirnov limit theorems for empirical distributions, Ann. Math. Statist. 19, pp. 177-189.

3. G. M. Manija (1949), Obobschenije kriterija A. N. Kolmogorova dlja otcenki zakona racpredelenija po empiricheskim dannym, Dokl. Akad. Nauk. SSSR 69, pp. 495-497.

4. N. Smirnov (1939), Sur les écarts de la courbe de distribution empirique, Rec. Math. (Mat. Sbornik), 6, pp. 3-26.

Princeton University

https://doi.org/10.4153/CMB-1965-012-5 Published online by Cambridge University Press