SOME THEOREMS ON THE RATIO OF EMPIRICAL DISTRIBUTION TO THE THEORETICAL DISTRIBUTION

S. C. TANG

1. Introduction. Let $X_1, X_2, \cdots, X_n$ be mutually independent random variables with the common cumulative distribution function $F(x)$. Let $X_1^*, X_2^*, \cdots, X_n^*$ be the same set of variables rearranged in increasing order of magnitude. In statistical language $X_1, X_2, \cdots, X_n$ form a sample of n drawn from the distribution with distribution function $F(x)$. The empirical distribution of the sample $X_1, \cdots, X_n$ is the step function $F_n(x)$ defined by

$$F_n(x) = \begin{cases} 0 & \text{for } x \leqq X_1^* \\ \dfrac{k}{n} & \text{for } X_k^* < x \leqq X_{k+1}^* \\ 1 & \text{for } x > X_n^* . \end{cases} \tag{1}$$

A. Kolmogorov developed a well-known limit distribution law for the difference between the empirical distribution and the corresponding theoretical distribution, assuming $F(x)$ continuous:

$$\lim_{n\to\infty} P\left\{\sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| < z\right\} = \begin{cases} 0 , & \text{for } z \leqq 0 , \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2} , & \text{for } z > 0 . \end{cases}$$

Equally interesting is Smirnov's theorem:

$$\lim_{n\to\infty} P\left\{\sqrt{n} \sup_{-\infty<x<\infty} [F_n(x) - F(x)] < z\right\} = \begin{cases} 0 , & \text{for } z \leqq 0 , \\ 1 - e^{-2z^2} , & \text{for } z > 0 . \end{cases}$$

In this paper we shall study the ratio of the empirical distribution to the theoretical distribution, and evaluate the distribution functions of the upper and lower bounds of the ratio. We shall prove the following four theorems.

THEOREM 1. *If F is everywhere continuous then*

$$P\left\{\sup_{0\leq F(x)\leq 1} \frac{F_n(x)}{F(x)} < z\right\} = s(z) = \begin{cases} 0 & \text{for } z \leqq 1 \\ 1 - \dfrac{1}{z} & \text{for } z > 1 . \end{cases}$$

Received July 28, 1961. I am deeply indebted to the referee for his corrections of my English writing. Without his aid, the present paper would not be read clearly, although the responsibility of errors rests with me.

THEOREM 2. *Let c be a positive number no larger than n and suppose that $F(x)$ is continuous in the range $0 \leqq F(x) \leqq c/n$. Then*

$$P\left\{\sup_{0<F(x)\leqq\frac{c}{n}} \frac{F_n(x)}{F(x)} < z\right\} = \begin{cases} \sum_{r=0}^{[cz]} \binom{n}{r} \left(1-\frac{c}{n}\right)^{n-r} \left(\frac{c}{n}\right)^r \\ \quad - \sum_{k=1}^{[cz]} \binom{n}{k} \frac{k^{k-1}(nz-k)^{n-k}}{(nz)^n} \\ \quad \cdot \sum_{r=k}^{[cz]} \binom{n-k}{r-k} \left(\frac{cz-k}{nz-k}\right)^{r-k} \left(1-\frac{cz-k}{nz-k}\right)^{n-r} \\ \qquad\qquad\qquad\qquad \text{for } 0 < z \leqq \frac{n}{c} \\ 0 \qquad\qquad\qquad\qquad \text{for } z \leqq 0 \\ 1 - \frac{1}{z} \qquad\qquad\qquad\qquad \text{for } z > \frac{n}{c}\,. \end{cases}$$

THEOREM 3. *Under the assumption of Theorem 1,*

$$P\left\{\inf_{0<F_n(x)\leqq 1} \frac{F_n(x)}{F(x)} \leqq z\right\} = I_n(z) = \begin{cases} 0 \qquad\qquad\qquad\qquad \text{for } z < \frac{1}{n} \\ \left(1-\frac{1}{nz}\right)^n \\ \quad + \sum_{k=1}^{nz} \binom{n}{k} \frac{(k-1)^{k-1}(nz-k)^{n-k}}{(nz)^n} \\ \qquad\qquad\qquad\qquad \text{for } \frac{1}{n} \leqq z \leqq 1 \\ 1 \qquad\qquad\qquad\qquad \text{for } z > 1\,. \end{cases}$$

THEOREM 4. *Under the assumptions of Theorem 2, if $1 \leqq c \leqq n$ then*

$$P\left\{\inf_{0<F_n(x)\leqq\frac{c}{n}} \frac{F_n(x)}{F(x)} \leqq z\right\} = \begin{cases} 0 \qquad\qquad\qquad\qquad \text{for } z < \frac{1}{n} \\ \left(1-\frac{1}{nz}\right)^n + \sum_{k=1}^{[nz]} \binom{n}{k} \frac{(k-1)^{k-1}(nz-k)^{n-k}}{(nz)^n} \\ \qquad\qquad\qquad\qquad \text{for } \frac{1}{n} \leqq z \leqq \frac{c}{n} \\ \left(1-\frac{1}{nz}\right)^n + \sum_{k=1}^{[c]} \binom{n}{k} \frac{(k-1)^{k-1}(nz-k)^{n-k}}{(nz)^n} \\ \qquad\qquad\qquad\qquad \text{for } z > \frac{c}{n}\,. \end{cases}$$

2. An elementary lemma. We shall use the following lemma.

LEMMA. *Let k be a positive integer and α an arbitrary real number. Then*

$$\sum_{r=1}^{k} \frac{r^{r-1}}{r!} \frac{(\alpha - r)^{k-r}}{(k-r)!} = \frac{\alpha^{k-1}}{(k-1)!}, \tag{2}$$

$$\frac{(\alpha-1)^k}{k!} + \sum_{r=1}^{k} \frac{(r-1)^{r-1}}{r!} \frac{(\alpha - r)^{k-r}}{(k-r)!} = \frac{\alpha^k}{k!}. \tag{3}$$

If $\alpha = k$ (a special case) then

$$\sum_{r=1}^{k} \frac{r^{r-1}}{r!} \frac{(k-r)^{k-r}}{(k-r)!} = \frac{k^{k-1}}{(k-1)!} = \frac{k^k}{k!}, \tag{4}$$

$$\frac{(k-1)^k}{k!} + \sum_{r=1}^{k} \frac{(r-1)^{r-1}}{r!} \frac{(k-r)^{k-r}}{(k-r)!} = \frac{k^k}{k!}. \tag{5}$$

Proof. Let $f(x)$ be a polynomial whose degree is less than k. Then

$$\sum_{r=0}^{k} \binom{k}{r} (-1)^r f(r) = (-1)^n [\Delta_n f(x)]_{x=0} = 0$$

i.e.

$$\sum_{r=1}^{k} \binom{k}{r} (-1)^r f(r) = -f(0).$$

Now we can directly obtain (2) and (3):

$$\begin{aligned}
\sum_{r=1}^{k} \frac{r^{r-1}}{r!} \frac{(\alpha-\gamma)^{k-r}}{(k-r)!} &= \sum_{r=1}^{k} \frac{r^{r-1}}{r!} \sum_{s=0}^{k-r} \frac{\alpha^s (-r)^{k-r-s}}{s!\,(k-r-s)!} \\
&= \sum_{s=0}^{k-1} \frac{(-1)^{k-s}\alpha^s}{s!} \sum_{r=1}^{k-s} \frac{(-1)^r r^{k-s-1}}{r!\,(k-r-s)!} \\
&= \frac{(-1)^{k-(k-1)}\alpha^{k-1}}{(k-1)!}(-1) = \frac{\alpha^{k-1}}{(k-1)!}.
\end{aligned}$$

$$\begin{aligned}
&\frac{(\alpha-1)^k}{k!} + \sum_{r=1}^{k} \frac{(r-1)^{r-1}}{r!} \frac{(\alpha-r)^{k-r}}{(k-r)!} \\
&= \frac{(\alpha-1)^k}{k!} + \sum_{r=1}^{k} \frac{(r-1)^{r-1}}{r!} \sum_{s=0}^{k-r} \frac{(\alpha-1)^s(-r+1)^{k-r-s}}{s!\,(k-r-s)!} \\
&= \frac{(\alpha-1)^k}{k!} + \sum_{s=0}^{k-1} \frac{(-1)^{k-s}(\alpha-1)^s}{s!} \sum_{r=1}^{k-s} \frac{(-1)^r (r-1)^{k-s-1}}{r!\,(k-r-s)!} \\
&= \frac{(\alpha-1)^k}{k!} + \sum_{s=0}^{k-1} \frac{(-1)^{k-s}(\alpha-1)^s(-1)^{k-s}}{s!\,(k-s)!} \\
&= \sum_{s=0}^{k} \frac{(\alpha-1)^s}{s!\,(k-s)!} = \frac{\alpha^k}{k!}.
\end{aligned}$$

3. Proof of Theorem 1. First we consider the case $z \geqq 1$. Let y_k be the largest root of the equation

$$F(x) = \frac{k}{nz} \qquad (1 \leqq k \leqq n) .$$

Since F is continuous, y_k is well defined. Now we evaluate probability of the inequality

$$(6) \qquad \sup_{0 \leqq F(x) \leqq 1} \frac{F_n(x)}{F(x)} \geqq z$$

that is, the probability that there exists an x such that

$$(7) \qquad \frac{F_n(x)}{F(x)} \geqq z .$$

If (7) is true, then, since $F(x)$ is nondecreasing and $\lim_{x\to\infty} (F_n(x)/F(x)) = 1$, there exists an x_0 such that

$$\frac{F_n(x_0)}{F(x_0)} = z .$$

By the definition of F_n, $F_n(x_0) = k/n$ for some k, so that $F(x_0) = k/nz$. Therefore we can take x_0 as one of y_k. In other words, for one value of y_k we have

$$F_n(y_k) = \frac{k}{n}$$

and the inequality

$$(8) \qquad X_k^* < y_k \leqq X_{k+1}^*$$

is true. Now let A_k be the event that this inequality holds.

Clearly (6) occurs if, and only if, at least one among the events

$$A_1, A_2, A_3, \cdots, A_n$$

occurs. Generally the A_k are not mutually exclusive events so that the additive law is of no use. We may, with the help of an associated set of mutually exclusive events, deal with this situation. Put

$$u_k = \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{k-1} A_k \qquad (k = 1, 2, \cdots, n) .$$

where $\bar{A}$ denotes the complementary event of A. We have

$$(9) \qquad P\left\{\sup_{0<F(x)\leqq 1} \frac{F_n(x)}{F(x)} \geqq z\right\} = P\left\{\sum_{k=1}^{n} A_k\right\} = P\left\{\sum_{k=1}^{n} u_k\right\} = \sum_{k=1}^{n} P(u_k) .$$

If we employ the following conditional probability formula, we can evaluate $P(u_k)$.

$$(10) \qquad P(A_k) = \sum_{r=1}^{k} P(u_r) P\{A_k \mid u_r\} = \sum_{r=1}^{k} P(u_r) P\{A_k \mid A_r\} .$$

Now A_k occurs if k of the n observations fall to the left of y_k, and $n - k$ to the right; hence

$$P(A_k) = \binom{n}{k}\left(\frac{k}{nz}\right)^k\left(1 - \frac{k}{nz}\right)^{n-k}.$$

And $P(A_k \mid A_r)$ is the probability that of $(n - r)$ observations, known to lie to the right of y_r, $(k - r)$ lie to the left of y_k and $n - k$ to the right. The probability of occurrence of an event, whose value is greater than or equal to y_r and is on the interval $y_r \leqq x \leqq y_k$ is

$$\frac{k/nz - r/nz}{1 - r/nz} = \frac{k - r}{nz - r}.$$

Therefore

$$P(A_k \mid A_r) = \binom{n - r}{k - r}\left(\frac{k - r}{nz - r}\right)^{k-r}\left(1 - \frac{k - r}{nz - r}\right)^{n-k}.$$

If we employ the notation

$$p_k(z) = \frac{(nz - n + k)^k}{k!},$$

we get

$$P(A_k) = \frac{p_k(1)p_{n-k}(z)}{p_n(z)}, \qquad P(A_k \mid A_r) = \frac{p_{k-r}(1)p_{n-k}(z)}{p_{n-r}(z)}.$$

Equation (10) can be reduced to

$$\frac{p_k(1)p_{n-k}(z)}{p_n(z)} = \sum_{r=1}^{k} P(u_r)\,\frac{p_{k-r}(1)p_{n-k}(z)}{p_{n-r}(z)},$$

$$p_k(1) = \sum_{r=1}^{k} P(u_r)\,\frac{p_n(z)}{p_{n-r}(z)}\,p_{k-r}(1).$$

Hence $P_k(1) = k^k/k!$ and $P_{k-r}(1) = (k - r)^{k-r}/(k - r)!$. By (3) of our lemma we know that

$$P(u_r)\frac{p_n(z)}{p_{n-r}(z)} = \frac{r^{r-1}}{r!}, \quad P(u_r) = \frac{r^{r-1}}{r!}\frac{p_{n-r}(z)}{p_n(z)} = \frac{n!}{(nz)^n}\frac{r^{r-1}}{r!}\frac{(nz - r)^{n-r}}{(n - r)!}.$$

And (9) and the (2) of our lemma tell us that

$$P\left\{\sup_{0<F(x)\leqq 1}\frac{F_n(x)}{F(x)} < z\right\} = 1 - P\left\{\sup_{0\leqq F(x)\leqq 1}\frac{F_n(x)}{F(x)} \geqq z\right\} = 1 - \sum_{k=1}^{n} P(u_k)$$

$$= 1 - \frac{n!}{(nz)^n}\sum_{k=1}^{n}\frac{k^{k-1}}{k!}\frac{(nz - k)^{n-k}}{(n - k)!}$$

$$= 1 - \frac{n!}{(nz)^n}\frac{(nz)^{n-1}}{(n - 1)!} = 1 - \frac{1}{z}, \qquad (z \geqq 1).$$

This completes the proof of Theorem 1.

Proof of Theorem 2. Since the ratio is nonnegative, the probability of the inequality is zero if $z \leqq 0$. Under the condition $z > n/c$, we know that the event $\{\sup_{0<F(x)\leqq c/n} F_n(x)/F(x) \geqq z\}$ is still equal to $\sum_{k=1}^{n} A_k$ by the result of Theorem 1. Suppose $0 < z \leqq n/c$. Then

$$P\left\{\sup_{0<F(x)\leqq \frac{c}{n}} \frac{F_n(x)}{F(x)} \geqq z\right\} = P\left\{\sum_{k=1}^{[cz]} A_k + A_{cz}^*\right\} = \sum_{k=1}^{[cz]} P(u_k) + P(u_{cz}^*) ,$$

where

$$A_{cz}^* = \{X_{[cz]+1}^* < y_{cz}\} , \qquad y_{cz} = \sup\left\{x: F(x) = \frac{c}{n}\right\} ,$$
$$u_{cz}^* = \bar{A}_1\bar{A}_2 \cdots \bar{A}_{[cz]}A_{cz}^* .$$

Since $P(U_k)$ has been computed, we need only evaluate $P(U_{cz}^*)$. We have

$$P(A_{cz}^*) = \sum_{k=1}^{[cz]} P(u_k)P\{A_{cz}^* \mid A_k\} + P(u_{cz}^*)P\{A_{cz}^* \mid A_{cz}^*\} ,$$
$$P(u_{cz}^*) = P(A_{cz}^*) - \sum_{k=1}^{[cz]} P\{A_{cz}^* \mid A_k\} .$$

From this we obtain

$$\begin{aligned} P\left\{\sup_{0\leqq F(x)\leqq \frac{c}{n}} \frac{F_n(x)}{F(x)} < z\right\} &= 1 - \sum_{k=1}^{[cz]} P(u_k) - P(u_{cz}^*) \\ &= 1 - P(A_{cz}^*) - \sum_{k=1}^{[cz]} P(u_k)[1 - P\{A_{cz}^* \mid A_k\}] \\ &= P(\bar{A}_{cz}^*) - \sum_{k=1}^{[cz]} P(u_k)P\{\bar{A}_{cz}^* \mid A_k\} \\ &= \sum_{r=0}^{[cz]} \binom{n}{r}\left(\frac{c}{n}\right)^r\left(1 - \frac{c}{n}\right)^{n-r} - \sum_{k=1}^{[cz]} \binom{n}{k} \frac{k^{k-1}(nz-k)^{n-k}}{(nz)^n} \\ &\quad \cdot \sum_{r=k}^{[cz]} \binom{n-k}{r-k}\left(\frac{cz-k}{nz-k}\right)^{r-k}\left(1 - \frac{cz-k}{nz-k}\right)^{n-r} . \end{aligned}$$

Since

$$\begin{aligned} P(\bar{A}_{cz}^*) = P\{y_{cz} \leqq X_{[cz]+1}^*\} &= \sum_{r=0}^{[cz]} P\{X_r^* < y_{cz} \leqq X_{r+1}^*\} \\ &= \sum_{r=0}^{[cz]} \binom{n}{r}\left(\frac{c}{n}\right)^r\left(1 - \frac{c}{n}\right)^{n-r} \end{aligned}$$

we have

$$P(\bar{A}_{cz}^* \mid A_k) = \sum_{r=k}^{[cz]} \binom{n-k}{r-k}\left(\frac{cz-k}{nz-k}\right)^{r-k}\left(1 - \frac{cz-k}{nz-k}\right)^{n-r}$$

which completes the proof. The distribution of Theorem 1 is continuous; the distribution of Theorem 2 is not continuous on the interval $0 \leqq z \leqq n/c$.

Theorem 3 is a special case of Theorem 4 under the condition $c = n$. In fact, by (4) of our lemma, setting $z = 1$, we deduce Theorem 3 as follows:

$$P\left\{\inf_{0<F_n(x)\leqq 1}\frac{F_n(x)}{F(x)} \leqq 1\right\} = \left(1-\frac{1}{n}\right)^n + \sum_{k=1}^{n}\binom{n}{k}\frac{(k-1)^{k-1}(n-k)^{n-k}}{n^n}$$

$$= \frac{n!}{n^n}\left[\frac{(n-1)^n}{n!} + \sum_{k=1}^{n}\frac{(k-1)^{k-1}}{k!}\frac{(n-k)^{n-k}}{(n-k)!}\right]$$

$$= \frac{n!}{n^n}\frac{n^n}{n!} = 1 .$$

Thus we need only prove Theorem 4. The distribution function of Theorem 3 has a discontinuity corresponding to $z = 1$; the distribution function of Theorem 4 is continuous on the interval $1 \leqq c \leqq n$.

5. Proof of Theorem 4. On the internal $F_n(x) > 0$ the maximum of $F(x)$ is 1; the minimum of $F_n(x)$ is $1/n$. From these results it follows that the lower bound of the ratio is no smaller than $1/n$. Therefore

$$P\left\{\inf_{0<F_n(x)\leqq\frac{c}{n}}\frac{F_n(x)}{F(x)} \leqq z\right\} = 0$$

if $z < 1/n$. Let z_k be the least root of the equation

$$F(x) = \frac{k}{nz} \qquad \left(z \geqq \frac{1}{n},\quad 1 \leqq k \leqq [nz]\right) .$$

Define events

$$B_0\colon\ X_1 \geqq z_1 ,$$

$$B_k\colon\ X_k^* < z_k \leqq X_{k+1}^* .$$

If $1/n \leqq z \leqq c/n$, the event

$$\inf_{0<F_n(x)\leqq\frac{c}{n}}\frac{F_n(x)}{F(x)} \leqq z$$

is the union of the events

$$B_0, B_1, B_2, \cdots, B_{[nz]} .$$

For the purpose of evaluating the probability of $\sum_k B_k$ we put

$$V_0 = B_0 , \qquad V_k = \bar{B}_0\bar{B}^1\cdots\bar{B}_{k-1}B_k \qquad (k = 1, 2, \cdots, [nz]) .$$

We have

$$P(B_k) = \sum_{r=0}^{k} P(V_r)P(B_k \mid B_r), \qquad (k = 0, 1, \cdots, [nz]). \tag{11}$$

Here

$$P(V_0) = P(B_0) = \left(1 - \frac{1}{nz}\right)^n,$$

$$P\{B_k \mid B_0\} = \binom{n}{k}\left(\frac{k-1}{nz-1}\right)^k\left(\frac{nz-k}{nz-1}\right)^{n-k}.$$

Now we transform (11) into the following form:

$$\frac{p_k(1)p_{n-k}(z)}{p_n(z)} = \frac{(k-1)^k p_{n-k}(z)}{k!\, p_n(z)} + \sum_{r=1}^{k} P(V_r)\frac{p_{k-r}(1)p_{n-k}(z)}{p_{n-r}(z)},$$

$$p_k(1) = \frac{(k-1)^k}{k!} + \sum_{r=1}^{k} P(V_r)\frac{p_n(z)}{p_{n-r}(z)}p_{k-r}(1).$$

By (4) of our lemma we obtain

$$P(V_r) = \frac{(r-1)^{r-1}}{r!}\frac{p_{n-r}(z)}{p_n(z)} = \binom{n}{r}\frac{(r-1)^{r-1}(nz-r)^{n-r}}{(nz)^n}, \quad (1 \leqq r \leqq [nz]).$$

It follows that if $1/n \leqq z \leqq c/n$,

$$P\left\{\inf_{0<F_n(x)\leq\frac{c}{n}}\frac{F_n(x)}{F(x)} \leqq z\right\} = P\left\{\sum_{k=0}^{[nz]} B_k\right\} = \sum_{k=0}^{[nz]} P(V_k)$$

$$= \left(1 - \frac{1}{nz}\right)^n + \sum_{k=1}^{[nz]}\binom{n}{k}\frac{(k-1)^{k-1}(nz-k)^{n-k}}{(nz)^n}.$$

If $z > c/n$,

$$P\left\{\inf_{0<F_n(x)\leq\frac{c}{n}}\frac{F_n(x)}{F(x)} \leqq z\right\} = P\left\{\sum_{k=0}^{[c]} B_k\right\} = \sum_{k=0}^{[c]} P(V_k)$$

$$= \left(1 - \frac{1}{nz}\right)^n + \sum_{k=1}^{[c]}\binom{n}{k}\frac{(k-1)^{k-1}(nz-k)^{n-k}}{(nz)^n}.$$

Theorem 4 is thus proved.

6. Conclusions. Theorems 1 and 3 can be applied to test whether statistical data correspond with the theoretical distribution or not. In the process of testing, z_s and z_i are separately representative of the ratio's upper and lower bounds. If $S(z_s)$ is small and $I_n(z_i)$ large, we conclude that the empirical data in two extreme tails correspond with the theoretical distribution; conversely, if $S(z_s)$ is very large and $I_n(z_i)$ very small, the statistical data do not correspond with the theoretical distribution. Our test is more sensitive in the two tails, but less sensitive in the central part, than Smirnov's test.