Archivum Mathematicum

Jaroslav Michálek

The distribution of the estimate of entropy and its applications

Archivum Mathematicum, Vol. 15 (1979), No. 3, 133--141

Persistent URL: http://dml.cz/dmlcz/107033

ARCH. MATH. 3, SCRIPTA FAC. SCI. NAT. UJEP BRUNENSIS
XV: 133—142, 1979

THE DISTRIBUTION OF THE ESTIMATE OF ENTROPY AND ITS APPLICATIONS

JAROSLAV MICHÁLEK, Brno

(Received August 14, 1978)

Let $X_1, X_2, \ldots, X_n$ be a random sample of a size n taken from the continuous random variable with a density function $f(x)$. Then as it is known, the generally accepted method for estimating the unknown density function $f(x)$ is that by means of a construction of a histogram. This method is based on the fact of statistical convergence of relative frequencies $\tilde{p}_i = \frac{m_i}{n}$ to the estimated probabilities p_i. The speed of convergence is characterized by the dependence of dispersion $D\tilde{p}_i$ on the size n of random sample. As it is known the order of $D\tilde{p}_i$ is n^{-1}.

Tarasenko in [4] suggested an other estimate of density function $f(x)$ based on an order random sample. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order random sample arised from the random sample $X_1, X_2, \ldots, X_n$. We shall restrict our attention to the case, when $f(x) = 0$ for $x \notin [a, b]$, $-\infty < a < b < \infty$. Then the estimate $\tilde{f}(x)$ of $f(x)$ given by Tarasenko is

$$\tilde{f}(x) = \frac{1}{n+1} \sum_{j=0}^{n} \frac{1}{\Delta x_j} \pi(x, \Delta x_j) \tag{1}$$

where

$$\Delta x_j = X_{(j+1)} - X_{(j)} \qquad \text{for } j = 0, 1, \ldots, n$$
$$X_{(0)} = a, \qquad X_{(n+1)} = b;$$

$$\begin{aligned}
\pi(x, \Delta x_j) &= 1 \qquad \text{for } x \in (X_{(j)}, X_{(j+1)}], \qquad j = 0, 1, \ldots, n-1 \\
\pi(x, \Delta x_j) &= 0 \qquad \text{for } x \notin (X_{(j)}, X_{(j+1)}], \qquad j = 0, 1, \ldots, n-1 \\
\pi(x, \Delta x_n) &= 1 \qquad \text{for } x \in (X_{(n)}, b) \\
\pi(x, \Delta x_n) &= 0 \qquad \text{for } x \notin (X_{(n)}, b)
\end{aligned}$$

The estimate $\tilde{f}(x)$ estimates the density function $f(x)$ between two neighbouring ordered observations as

$$\tilde{f}_j = 1/[(n+1)\,\Delta x_j].$$

Tarasenko in [4] shows that the order of dispersion $D\tilde{f}_j$ is n^{-2}. Consequently, the dispersion of the estimate $\tilde{f}_j$ decreases no slower (with increasing n) than that of $\tilde{p}_i$. This eventually means that the estimate (1) is at least no worse than a histogram.

The very interesting outcome of the representation (1) is that it gives the possibility of estimating an entropy $\tilde{H}$ of the measured variable directly from observations $X_1, X_2, X_3, \ldots, X_n$. If we calculate the entropy integral for estimate $\tilde{f}(x)$, we get

$$\text{(2)} \qquad \tilde{H} = -\int_{-\infty}^{\infty} \tilde{f}(x) \log \tilde{f}(x)\, dx = \log(n+1) + \frac{1}{n+1} \sum_{j=0}^{n} \log \Delta x_j,$$

This relation simply means that to obtain the statistical estimation of the differential entropy, one needs to measure only the distances between neighbouring ordered observations $X_1, X_2, \ldots, X_n$.

A special attention must be paid to the case, when the density function $f(x)$ is that of uniformly distribution over the interval $[0, 1]$, because the random variable with arbitrary distribution can be transformed to the random variable uniformly distributed over the interval $[0, 1]$. Therefore it is necessary for an other statistical use of the statistic $\tilde{H}$ to know the distribution of $\tilde{H}$ under the condition that the $f(x)$ is the density function of uniform distribution over the interval $[0, 1]$. Tarasenko approximated the distribution of the statistic $\tilde{H}$ under above given condition by means of a normal distribution $N(\mu, \sigma^2)$ with parameters: expected value $\mu = E\tilde{H}$ and dispersion $\sigma^2 = D\tilde{H}$. He has proposed this approximation on the basis of "a mathematical experiment" performed on a computer.

The distribution of the statistic $\tilde{H}$ can be described by its characteristic function $\varphi(t)$ given by the following theorem.

Theorem 1: Let $X_1, X_2, \ldots, X_n$ be random sample of the size n taken from uniformly distributed random variable over an interval $[0, 1]$; $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ order random sample is arised from random sample $X_1, X_2, \ldots, X_n$; $X_{(0)} = 0$, $X_{(n+1)} = 1$ and $\Delta x_j = X_{(j+1)} - X_{(j)}, j = 0, 1, \ldots, n$.

Then the statistic $\tilde{H}$ given by (2) has characteristic function

$$\text{(3)} \qquad \varphi(t) = n!(n+1)^{it} \frac{\Gamma^{n+1}\left(1 + \frac{it}{n+1}\right)}{\Gamma(n+1+it)},$$

for $t \in (-\infty, \infty)$.

Proof: The statistic $\tilde{H}$ can be written in the following way

$$\text{(4)} \qquad \tilde{H} = \log(n+1) + H_0/(n+1)$$

where

$$\text{(5)} \qquad H_0 = \sum_{j=0}^{n} \log \Delta x_j = \sum_{j=0}^{n} \log(X_{(j+1)} - X_{(j)}).$$

First we shall find the characteristic function $\varphi_0(t)$ of the statistic H_0.
We receive

(6) $$\varphi_0(t) = \mathbf{E}\,\mathrm{e}^{itH_0} = \mathbf{E}\,\mathrm{e}^{it\sum_{j=0}^{n}\log(X_{(j+1)}-X_{(j)})} =$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \mathrm{e}^{it\sum_{j=0}^{n}\log(x_{j+1}-x_j)} g(x_1, \dots, x_n)\,\mathrm{d}x_1, \dots, \mathrm{d}x_n,$$

where $x_0 = 0$, $x_{n+1} = 1$ and $g(x_1, \dots, x_n)$ is the density function of the order random sample $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. By [3] or [6], the density function $g(x_1, \dots, x_n)$ can be written in the form

(7) $$\begin{aligned} g(x_1, x_2, \dots, x_n) &= n!\, h(x_1)\, h(x_2) \dots h(x_n) && \text{for } -\infty < x_1 < x_2 < \dots < x_n < \infty \\ g(x_1, x_2, \dots, x_n) &= 0 && \text{for the others } x_1, x_2, \dots, x_n \end{aligned}$$

where

$$\begin{aligned} h(x) &= 1 \qquad \text{for } x \in [0, 1] \\ h(x) &= 0 \qquad \text{for } x \notin [0, 1] \end{aligned}$$

is a density function of uniform distribution over the interval $[0, 1]$.

Considering the expression (7) for $g(x_1, \dots, x_n)$, we obtain from (6)

(8) $$\varphi_0(t) = n! \underset{0 \leqq x_1 < \dots < x_n \leqq 1}{\int \dots \int} \mathrm{e}^{it\sum_{j=0}^{n}\log(x_{j+1}-x_j)}\,\mathrm{d}x_1 \dots \mathrm{d}x_n.$$

Introducing the substitution

$$\begin{aligned} t_j &= x_{j+1} - x_j \qquad \text{for } j = 1, 2, \dots, n-1 \\ t_n &= 1 - x_n \end{aligned}$$

we reduce (8) to

$$\varphi_0(t) = n! \underset{M}{\int \dots \int} \mathrm{e}^{it\sum_{j=0}^{n}\log t_j + \log(1 - \sum_{j=0}^{n} t_j)}\,\mathrm{d}t_1 \dots \mathrm{d}t_n,$$

where

$$M = \{(t_1, \dots, t_n) \mid 0 < t_j < 1 \text{ for } j = 1, 2, \dots, n \text{ and } 0 < \sum_{j=1}^{n} t_j < 1\}$$

Hence, after simple modifications, we receive

$$\varphi_0(t) = n! \underset{M}{\int \dots \int} (1 - \sum_{j=1}^{n} t_j)^{it} (\prod_{j=1}^{n} t_j^{it})\,\mathrm{d}t_1 \dots \mathrm{d}t_n.$$

Now, using the theorem on repeated integrals, we obtain

$$\varphi_0(t) = n! \int_0^1 t_1^{it}\Big(\int_0^{1-t_1} t_2^{it} \dots \Big(\int_0^{1-(t_1+\dots+t_{n-2})} t_{n-1}^{it}\Big(\int_0^{1-(t_1+\dots+t_{n-1})} t_n^{it} \times$$

$$\times (1 - \sum_{j=1}^{n} t_j)^{it}\,\mathrm{d}t_n\Big)\,\mathrm{d}t_{n-1}\Big) \dots \mathrm{d}t_2\Big)\,\mathrm{d}t_1.$$

Using a notation

$$s_k = 1 - \sum_{j=1}^{k-1} t_j, \qquad k = 2, 3, \ldots, n,$$

for which the recurrent formula

$$s_k = s_{k-1} - t_{k-1}, \qquad k = 2, 3, \ldots, n$$

is valid, we receive

$$\varphi_0(t) = n! \int_0^{s_1} t_1^{it}\Big(\int_0^{s_2} t_2^{it} \ldots \Big(\int_0^{s_{n-1}} t_{n-1}^{it}\Big(\int_0^{s_n} t_n^{it}(s_n - t_n)^{it}\,\mathrm{d}t_n\Big)\,\mathrm{d}t_{n-1}\Big)\ldots\,\mathrm{d}t_2\Big)\,\mathrm{d}t_1. \tag{9}$$

Substituting in the k-th integral of the expression (9) variable y_k for t_k/s_k for $k = n, n - 1, \ldots, 2, 1$ we obtain consecutively by integrating step by step

$$\varphi_0(t) = n! \int_0^{s_1} t_1^{it}\Big(\int_0^{s_2} t_2^{it} \ldots \Big(\int_0^{s_{n-2}} t_{n-2}^{it}\Big(\int_0^{s_{n-1}} s_n^{2it+1} t_{n-1}^{it}\,\mathrm{d}t_{n-1}\Big)\,\mathrm{d}t_{n-2}\Big)\ldots\,\mathrm{d}t_2\Big)\,\mathrm{d}t_1 \times$$

$$\times \int_0^1 y_n^{it}(1 - y_n)^{it}\,\mathrm{d}y_n = \ldots = n! \prod_{k=1}^{n} \int_0^1 y_k^{it}(1 - y_k)^{(n-k+1)it+n-k}\,\mathrm{d}y_k =$$

$$= n! \prod_{k=1}^{n} \beta(it + 1, (n - k + 1)(1 + it)) = n! \prod_{k=1}^{n} \beta(1 + it, k(1 + it)).$$

Hence, by means of the well-known relation

$$\beta(z_1, z_2) = \frac{\Gamma(z_1)\,\Gamma(z_2)}{\Gamma(z_1 + z_2)}$$

between beta and gamma functions, we receive the final expression for the characteristic function $\varphi_0(t)$ in the form

$$\varphi_0(t) = n!\,\Gamma^{n+1}(1 + it)/\Gamma((n + 1)(1 + it)) \tag{10}$$

for $t \in (-\infty, \infty)$.

The statistic $\tilde{H}$ is a linear function of the statistic H_0 which is given by (4). Using known properties of characteristic function and (10), we can write the characteristic function $\varphi(t)$ of $\tilde{H}$ in the following way

$$\varphi(t) = e^{it\log(n+1)}\varphi_0(t)/(n + 1)) = n!(n + 1)^{it}\Gamma^{n+1}\left(1 + \frac{it}{n+1}\right)\Big/\Gamma(n + 1 + it). \tag{11}$$

Thus the theorem is proved.

Corollary 1: The statistic $\tilde{H}$ given by (2) has under the condition mentioned in Theorem 1 the expected value

$$E\tilde{H} = \log(n + 1) - \sum_{j=1}^{n} j^{-1} \tag{12}$$

and dispersion

(13) $$D\tilde{H} = \sum_{j=1}^{n} j^{-2} - \frac{n}{n+1}\frac{\pi^2}{6}.$$

Proof: First, we shall find the expected value and dispersion of the statistic H_0 given by (5) which has characteristic function $\varphi_0(t)$ given by (10).

Let us put
$$\psi(t) = \log \varphi_0(t).$$

Then we receive from the properties of the characteristic functions (see [2]) the relations

(14) $$EH_0 = i^{-1} \left.\frac{d\psi(t)}{dt}\right|_{t=0}$$

and

(15) $$DH_0 = -\left.\frac{d^2\psi(t)}{dt}\right|_{t=0}.$$

Now, we shall calculate these derivatives. We put $z_1 = 1 + it$ and $z_2 = (n+1)(1+it)$. Using (10) we obtain

$$\psi(t) = \log n! + (n+1)\log \Gamma(z_1) - \log \Gamma(z_2)$$

and

(16) $$\frac{d\psi(t)}{dt} = i(n+1)\left[\frac{1}{\Gamma(z_1)}\frac{d\Gamma(z_1)}{dz_1} - \frac{1}{\Gamma(z_2)}\frac{d\Gamma(z_2)}{dz_2}\right].$$

Applying the Gauss relation (see [5] p. 247)

$$\frac{1}{\Gamma(z)}\frac{d\Gamma(z)}{dz} = \int_0^\infty \left(\frac{e^{-x}}{x} - \frac{e^{-zx}}{1-e^{-x}}\right)dx,$$

which holds for all complex numbers z such that $\operatorname{Re} z > 0$, to (16) we receive after simple modifications

$$\frac{d\psi(t)}{dt} = i(n+1)\int_0^\infty \frac{e^{-itx}}{e^x - 1}(e^{-nx-itnx} - 1)\,dx.$$

Hence

$$\left.\frac{d\psi(t)}{dt}\right|_{t=0} = i(n+1)\int_0^\infty \frac{e^{-nx}-1}{e^x-1}\,dx$$

and substituting y for e^x we find successively

(17) $$\left.\frac{d\psi(t)}{dt}\right|_{t=0} = i(n+1)\int_0^\infty \frac{1-y^n}{y-1} y^{-n-1}\,dy = -i(n+1)\int_0^\infty \sum_{k=2}^{n+1} y^{-k} = -i(n+1)\sum_{j=1}^{n} j^{-1}.$$

Further differentiating (16), we obtain

$$\frac{d^2\psi(t)}{dt} = -(n+1)\frac{d^2 \log \Gamma(z_1)}{dz_1^2} + (n+1)\frac{d^2 \log \Gamma(z_2)}{dz_2^2}.$$

Using here the equality (see [5] p. 241)

$$\frac{d^2 \log \Gamma(z)}{dz^2} = \sum_{j=0}^{\infty} \frac{1}{(z+j)^2}$$

we receive after simple modifications

$$\frac{d^2\psi(t)}{dt^2} = -(n+1)\Big(\sum_{j=0}^{\infty}(1+it+j)^{-2} - \sum_{j=0}^{\infty}(n+1)\left[(n+1)(1+it)+j\right]^{-2}.$$

Hence

$$(18) \qquad \left.\frac{d^2\psi(t)}{dt^2}\right|_{t=0} = (n+1)\Big(\sum_{j=0}^{\infty}(1+j)^{-2} - \sum_{j=0}^{\infty}(n+1)(n+1+j)^{-2}\Big) =$$

$$= (n+1)\Big(n\sum_{j=0}^{\infty}(1+j)^{-2} - (n+1)\sum_{j=0}^{n-1}(1+j)^{-2}\Big) =$$

$$= n(n+1)\frac{\pi^2}{6} - (n+1)^2\sum_{j=1}^{n} j^{-2}.$$

Using (14) and (15) we obtain from (17) and (18)

$$EH_0 = -(n+1)\sum_{j=1}^{n} j^{-1},$$

$$DH_0 = (n+1)^2\sum_{j=1}^{n} j^{-2} - n(n+1)\frac{\pi^2}{6}.$$

Considering (4), we receive from here

$$E\tilde{H} = \log(n+1) + EH_0/(n+1) = \log(n+1) - \sum_{j=1}^{n} j^{-1},$$

and

$$D\tilde{H} = (n+1)^{-2} DH_0 = \sum_{j=1}^{n} j^{-2} - \frac{n}{n+1}\frac{\pi^2}{6}.$$

Consequently, the corollary is proved.

The distribution of the statistic $\tilde{H}$ is then given by Theorem 1. By Corollary 1 there are given basic characteristics of this distribution. To find the density function $f_H(x)$ relevant to the characteristic function $\varphi(t)$, we must use the Fourier transformation and calculate according to the formula

$$(19) \qquad f_H(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt.$$

But to solve this integral is a matter by no means easy. Further we shall derive an approximation $f_A(x)$ of density function $f_H(x)$. First we shall approximate the characteristic function $\varphi_0(t)$ of the statistic $H_0 = (n+1)\tilde{H} - (n+1)\log(n+1)$ by means of the function $\psi_A(t)$ so that we replace $\Gamma(z)$ in (10) by the approximation $\sqrt{2\pi}\, z^{z-1/2}\, e^{-z}$ given by Stirling's formula (see [1] p. 552). In this way we obtain after simple modifications

$$\psi_A(t) = c_n(1 + it)^{-\frac{1}{2}n}\, e^{-it(n+1)\log(n+1)},$$

where

$$c_n = n!(2\pi)^{\frac{1}{2}n}/(n+1)^{n+\frac{1}{2}}.$$

The function $\psi_A(t)$ is not a characteristic function because $\psi_A(0) = c_n \neq 1$. The deviation $\psi_A(0)$ from 1 is caused by the approximation by Stirling's formula. To remove this deviation, we shall further deal with function

$$\varphi_A(t) = c_n^{-1}\, \psi_A(t),$$

which is an approximation of the characteristic function $\varphi_0(t)$. Hence using (4), we can write an approximation $\tilde{\varphi}_A(t)$ of the characteristic function $\varphi(t)$ of the statistic $\tilde{H}$ in the form

$$\tilde{\varphi}_A(t) = e^{it\log(n+1)}\, \varphi_A\left(\frac{t}{n+1}\right) = \left(1 + \frac{it}{n+1}\right)^{-\frac{1}{2}n}.$$

From here and by (19) we can express the density function $f_A(x)$ being found approximation of density function $f_H(x)$ and a density function corresponding to the characteristic function $\tilde{\varphi}_A(t)$ as follows

$$f_A(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \left(1 + \frac{it}{n+1}\right)^{-\frac{1}{2}n} dt.$$

Substituting $-2(n+1)\,s$ for t in the last integral we obtain

$$f_A(x) = \frac{n+1}{\pi} \int_{-\infty}^{\infty} e^{2i(n+1)sx} (-2is)^{-\frac{1}{2}n}\, ds. \tag{20}$$

Now, if we consider that the characteristic function $\chi_n(s)$ of the Pearson's χ^2 distribution with n degree of freedom is given by

$$\chi_n(s) = (1 - 2is)^{-\frac{1}{2}n} \qquad \text{for } s \in (-\infty, \infty)$$

and the density function $h_n(x)$ corresponding to this characteristic function is given by

$$h_n(x) = x^{\frac{1}{2}n-1}\, e^{-\frac{1}{2}x} \Big/ \left[2^{\frac{1}{2}n}\Gamma\left(\frac{n}{2}\right)\right] \qquad \text{for } x \geqq 0, \tag{21}$$

$$h_n(x) = 0 \qquad \text{for } x < 0,$$

then we can reduce (20) to the final form as follows

$$(22) \qquad f_A(x) = \frac{n+1}{\pi} \int_{-\infty}^{\infty} e^{2i(n+1)sx} \chi_n(s)\, ds = 2(n+1)\, h_n(-2(n+1)\, x)$$

and then using (21) we obtain

$$(23) \qquad \begin{aligned} f_A(x) &= (n+1)^{\frac{1}{2}n} (-x)^{\frac{1}{2}n-1} e^{(n+1)x} && \text{for } x \leqq 0 \\ f_A(x) &= 0 && \text{for } x > 0 \end{aligned}$$

which is the found approximation of the density function $f_H(x)$ of the statistic $\tilde{H}$ under conditions of Theorem 1.

It follows from (22) and (23) that the distribution of the statistic $K = -2(n+1)\tilde{H}$ can be approximated by Pearson's χ^2 distribution with n degree of freedom. A comparison of the approximation given by (23) and that based on the normal distribution

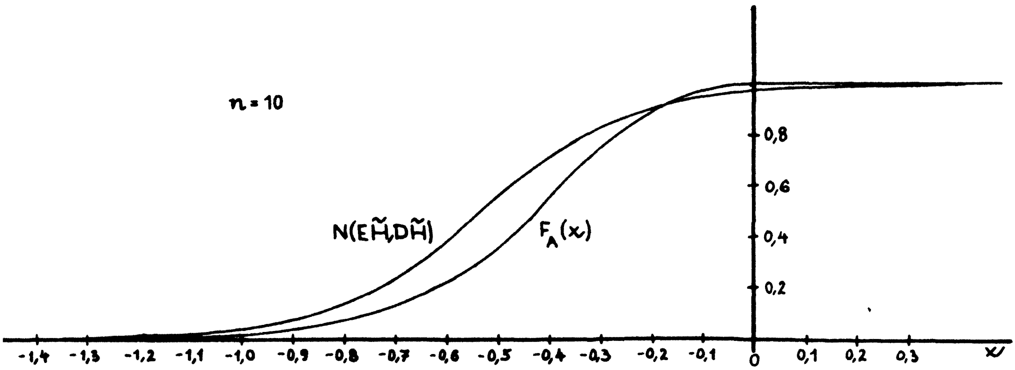

Fig. 1

given by Tarasenko in [4] is in Figure 1 for a random sample of size $n = 10$. In this figure there is given the distribution function $F_A(x) = \int_{-\infty}^{x} f_A(y)\, dy$ and that of normal distribution $N(E\tilde{H}, D\tilde{H})$. We can see that the distribution given by density function $f_A(x)$ has for large size of random sample the expected value greater than $E\tilde{H}$. Really

$$(24) \qquad \lim_{n\to\infty} E\tilde{H} = \lim_{n\to\infty} \left(\log(n+1) - \sum_{j=1}^{n} j^{-1}\right) = -C = -0{,}5772,$$

where C is Euler's constant, and the expected value distribution with density function $f_A(x)$ converges for $n \to \infty$ to the $-0{,}5$, because the expected value of Pearson's χ^2 distribution is equal to the degree of freedom. Then the expected value of the distribution with density function $f_A(x)$ is $\dfrac{-n}{2(n+1)}$.

From (24) it can be seen that $\tilde{H}$ has a bias asymptotically equal to $-C = -0{,}577$ and further that a statistic which has density function $f_A(x)$ and which can be approximated by the statistic $\tilde{H}$, has a bias asymptotically equal to $-0{,}5$ (thus smaller than statistic $\tilde{H}$).

All of the foregoing enables us to propose nonparametric entropy test of goodness-of-fit. For testing the hypothesis: "the random sample $Y_1, Y_2, \ldots, Y_n$ is from distribution with distribution function $G(y)$", it is necessary:

a) to transform the random sample $Y_1, Y_2, \ldots, Y_n$ into random sample $X_1, X_2, \ldots, X_n$ taken of uniformly distributed random variable over an interval $[0, 1]$ under the condition that the hypothesis is true. This transformation is

$$X_i = G(Y_i), \qquad i = 1, 2, \ldots, n.$$

b) to calculate the order random sample $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$; to calculate the statistic $\tilde{H}$ by (2) and statistic $K = -2(n + 1)\,\tilde{H}$.

c) since the uniform distribution has a maximum value of entropy under interval $[0, 1]$ of possible values of x's we reject the hypothesis if $\tilde{H}$ is "small enough". It means, we reject the hypothesis on the significant level α if

$$K > \chi^2_{1-\alpha}(n),$$

where $\chi^2_{1-\alpha}(n)$ is $(1 - \alpha)$ – quantile of Perason's χ^2 distribution with n degree of freedom. In the case $K \leqq \chi^2_{1-\alpha}(n)$ we accept it.

REFERENCES

[1] Макрушевич, А. И: *Теория аналитических функций*, изд. технико-теор. лит. Москва—Ленинград 1950.

[2] Rao, C. R.: *Linear Statistical Inference and Its Applications*, John Wiley Sons, New York, London, Sydney, 1965.

[3] Sarkan, A. E. and Greenborg, B. G.: *Contributions to Order Statistics*, John Wiley, New York 1962.

[4] Tarasenko, F. P.: *On the Evaluation of an Unknown Probability Density Function, the Direct Estimation of the Entropy from Independent Observations of a Continuous Random Variable, and the Distribution—Free Entropy Test of Goodnes-of-Fit.* Proceedings of the IEEE, November 1968.

[5] Whittaker, E. P. and Watson, G. M.: *A Course of Modern Analysis.* At the Univ. Press, Cambridge 1920.

[6] Wilks, S. S.: *Mathematical Statistics*, John Wiley, New York 1962.

J. Michálek
662 95 Brno, Janáčkovo nám. 2a
Czechoslovakia