ON ESTIMATES FOR THE ENTROPY OF A LANGUAGE ACCORDING TO SHANNON

A. P. SAVCHUK (*MOSCOW*)

(*Summary*)

C. E. Shannon proposed the upper and lower estimates for the entropy of a language. It is proved in this paper that in order to attain the lower estimate, it is necessary and sufficient that some letters be equally probable, the probability of the rest being equal to zero after any combination b_i^N of N letters such that $p(b_i^{N-1}) > 0$. In order to attain the upper estimate, it is necessary and sufficient that the probability that the k-th letter appears after b_i^N be dependent on k and N, and independent of i, for a sequence of language letters arranged in descending order as to the probability that they appear after b_i^N (note that the arrangement of letters in this sequence, however, depends on i). In the latter result it is supposed that no letter occurs with the probability $= 1$, which is true for every real language.

ON ESTIMATING REGRESSION

E. A. NADARAYA

(*Translated by Bernard Seckler*)

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be a sample of independent observations of a two-dimensional random variable (X, Y) with distribution function

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\,du\,dv.$$

Let $f(x)$ and $g(y)$ be the density functions of X and Y, and let $\bar{y}(x)$ denote the regression curve for Y with respect to X.

If $\bar{y}(x)$ is of a known analytic form and contains a certain number of unknown parameters, there exist well-known methods for estimating these parameters from empirical data, for example, by the method of least squares. This note gives a statistic of general form for estimating regression curves and studies certain of its properties for $n \to \infty$.

As an approximation to the regression curve $\bar{y}(x)$ from empirical data, we take the statistic

$$\bar{y}_n(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h(n)}\right)},$$

where $K(x)$ is a density function satisfying the conditions:

a) $K(x) < C < \infty$ and b) $\lim_{x\to\pm\infty}|xK(x)| = 0$ and $h(n) \to 0$ when $n \to \infty$.

Let

$$\frac{1}{nh(n)}\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h(n)}\right) = \varphi_n(x), \qquad \frac{1}{nh(n)}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h(n)}\right) = \psi_n(x).$$

Our results can be stated in the form of the following three theorems.

Theorem 1. *If Y is a bounded random variable and $nh^2(n) \to \infty$, then*

$$\mathbf{E}\bar{y}_n(x) = \frac{\mathbf{E}\varphi_n(x)}{\mathbf{E}\psi_n(x)} + O\left(\frac{1}{nh^{3/2}(n)}\right)$$

and

$$\mathbf{P}\left\{\sqrt{nh(n)}\left[\bar{y}_n(x) - \frac{\mathbf{E}\varphi_n(x)}{\mathbf{E}\psi_n(x)}\right] < \lambda\right\} \to N(0, \sigma(x)),$$

where $\sigma^2(x) = \mathbf{E}(Y^2|X = x)\int_{-\infty}^{\infty} K^2(u)du/f(x)$.

Theorem 2. *Let* $nh^2(n) \to \infty$ *and* $\int_{-\infty}^{\infty} y^2 g(y)dy < \infty$. *Then* $\bar{y}_n(x)$ *is a consistent estimate for* $\bar{y}(x)$ *for any* x *for which* $\bar{y}(x)$ *and* $f(x)$ *are continuous and* $f(x)$ *is positive.*

Theorem 3. *Let the characteristic function* $\chi(t) = \int_{-\infty}^{\infty} e^{itx}K(x)dx$ *be absolutely integrable,* $\bar{y}(x)$ *and* $f(x)$ *be continuous on the finite interval* $[a, b]$, *and*

$$\min_{a\leqq x\leqq b}\{f(x)\} = \mu > 0, \qquad \int_{-\infty}^{\infty} y^4 g(y)dy < \infty, \qquad \sum_{n=1}^{\infty} n^{-2}h^{-4}(n) < \infty.$$

Then with probability 1,

$$\sup_{a\leqq x\leqq b} |\bar{y}_n(x) - \bar{y}(x)| \to 0$$

when $n \to \infty$.

The proofs of Theorems 1 and 2 are based on Theorem IA of [1], and that of Theorem 3 on the use of the Fourier transform of $K(x)$.

V. A. Steklov Mathematics Institute of the USSR Academy of Sciences

Received by the editors September 27, 1963

REFERENCE

[1] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist., **33**, 3, 1962, pp. 1065–1076.

ON ESTIMATING REGRESSION

E. A. NADARAYA *(TBILISI)*

(Summary)

A study is made of certain properties of an approximation to the regression line on the basis of sampling data when the sample size increases unboundedly.

ON LIMIT DISTRIBUTIONS FOR ORDER STATISTICS

D. M. CHIBISOV

(Translated by Bernard Seckler)

Let $X_1, X_2, \cdots, X_n$ be independent random variables with the same distribution function $F(x)$, and let $\xi_1^{(n)} \leqq \xi_2^{(n)} \leqq \cdots \leqq \xi_n^{(n)}$ be the corresponding order statistics. In [1], N. V. Smirnov studied the limit distributions for the $\xi_k^{(n)}$ and their domains of attraction for the cases where $k/n \to \lambda$, $0 < \lambda < 1$ and k (or $n-k$) is a constant. This paper considers the case where $k \to \infty$ and $k/n \to 0$ (or $n-k \to \infty$ and $(n-k)/n \to 0$).

Let $G_{kn}(x) = \mathbf{P}\{\xi_k^{(n)} < x\}$. Denote by $\Phi(x)$, the normal (0, 1) distribution function.

Lemma 1. *If* $n \to \infty$, $k \to \infty$, *and* $k/n \to 0$, *then*

$$G_{kn}(a_n x + b_n) - \Phi\big(u_n(x)\big) \to 0 \tag{1}$$

uniformly in x, *where*

$$u_n(x) = \frac{nF(a_n x + b_n) - k}{\sqrt{k}}. \tag{2}$$

This lemma results from Lemma 2, Part I of [1] taking into consideration the condition $k/n \to 0$.