CONSTRUCTION OF CONFIDENCE INTERVALS FOR MATHEMATICAL EXPECTATIONS OF RANDOM VARIABLES OF A CERTAIN TYPE

K. Kachiashvili, D. Melikjanian

ABSTRACT

A method of finding the confidence interval for mathematical expectation of a random variable is suggested. The given result is used in problems of restoration of nonlinear functional dependences. The introduced auxiliary functions are studied and tabulated.

CONFIDENCE INTERVAL FOR EXPECTATION

Confidence intervals for parameters of probability distribution laws of the studied phenomena are widely used when solving numerous theoretical and applied problems. Below one of the methods of finding the confidence interval for mathematical expectation of a random variable is given; it is based on the usage of ordinal statistics [2,3,12]. In contrast to the classical methods, this method doesn't require knowledge of other parameters of distribution laws (e.g., variance at normal probability distribution law) when constructing the confidence interval of mathematical expectation.

Let x be a random variable defined in interval $(-\infty, +\infty)$ with mathematical expectation m and variance σ^2; $\Phi(\dot{x})$ and let $p(\dot{x})$ be a function and a distribution density of the corresponding normalized random variable $\dot{x} = (x-m)/\sigma$; $[x_1, x_2, ..., x_N]$ - N is a size sample; x_{min} and x_{max} are accordingly, minimum and maximun sample elements.

Theorem 1. *The confidence interval for parameter m with confidence probability $1-\alpha$ is*

$$[x_{min} - h_N(\alpha)\cdot(x_{max} - x_{min});\ x_{max} + H_N(\alpha)\cdot(x_{max} - x_{min})],$$

where functions $h = h_N(\alpha)$ and $H = H_N(\alpha)$ are defined by solving equations

$$\psi(h) = \alpha/2 \quad \text{and} \quad \Psi(H) = \alpha/2$$

in region $h > -1$ and $H > -1$, at

$$\begin{aligned} \psi(h) &= \int_0^\infty N\cdot p(u)\cdot \Big(\Phi(u) - \Phi\big(u\cdot h/(1+h)\big)\Big)^{N-1} du; \qquad (0.1)\\ \Psi(H) &= \int_0^\infty N\cdot p(-u)\cdot \Big(\Phi\big(-u\cdot H/(1+H)\big) - \Phi(-u)\Big)^{N-1} du; \end{aligned}$$

Proof. Let's consider random variables $u = (x_{min} - m)/\sigma$ and $v = (x_{max} - m)/\sigma$. According to [2,15], density of their joint distribution is

$$p_w(u,v) = N(N-1)\cdot p(u)\cdot p(v)\cdot \Big(\Phi(v) - \Phi(u)\Big)^{N-2}, \quad \text{at } u \le v.$$

Confidence interval limits are defined by the following conditions

$$\begin{aligned} \mathcal{P}\{x_{min} - h_N(\alpha)\cdot(x_{max}-x_{min}) > m\} &= \alpha/2; \\ \mathcal{P}\{x_{max} + H_N(\alpha)\cdot(x_{max}-x_{min}) < m\} &= \alpha/2 \end{aligned}$$

(symbol $\mathcal{P}$ denotes probability of event). Inequalities

$$x_{min} - h\cdot(x_{max}-x_{min}) > m \quad \text{and} \quad x_{max} + H\cdot(x_{max}-x_{min}) < m$$

are equivalent to inequalities $v\cdot h/(1+h) < u$ and $v < u\cdot H/(1+H)$. Critical regions D_1 and D_2 on plane (uv) correspond to them; region D_1 is in the upper half-plane between straight lines $v = u$ and $v = u\cdot(1+h)/h$, and region D_2 – in the lower half-plane between straight lines $v = u$ and $v = u\cdot H/(1+H)$.

Let's denote probabilities of these inequalities $\psi(h)$ and $\Psi(H)$. These functions are equal to integrals of $p_w(u,v)$ over regions D_1 and D_2, which coincide with right parts of the equations (1).

Note, that equations (1) do not contain parameter σ. This takes place because a change in σ corresponds to the change in plot scale in fig.**??**, when each of regions D_0, D_1 and D_2 doesn't change its position with respect to axes of reference.

For positive argument values, functions $\psi(h)$ and $\Psi(H)$ may also be defined by formulas

$$\begin{aligned} \psi(h) &= \int_0^\infty N\cdot p(u)\cdot\Big(\Phi\Big(u(1+h)/h\Big)-\Phi(u)\Big)^{N-1} du; \\ \Psi(H) &= \int_0^\infty N\cdot p(-u)\cdot\Big(\Phi(-u)-\Phi\Big(-u(1+H)/H\Big)\Big)^{N-1} du. \end{aligned}$$

MAIN PROPERTIES OF FUNCTIONS $\psi(h)$ AND $h_N(\alpha)$

If function $p(u)$ is even, i.e. symmetric with respect to mathematical expectation, then $\psi(h) = \Psi(h)$ and, correspondingly, $h_N(\alpha) = H_N(\alpha)$.

The results given below are true for symmetric densities of probability distribution.

Let's give values of $h_N(\alpha)$ for confidence probability limiting values at fixed N: at $\alpha \to 0 \quad h_N(\alpha) \to \infty$; and at $\alpha \to 1 \quad h_N(\alpha) \to -1/2$.

At the fixed α the function $h_N(\alpha)$ decreases when sample size N grows, and becomes negative from a certain value of N, namely at $N > \log_2(2/\alpha)$. If distribution density $p(u)$ is continuous, and there exists a positive number u_B, such that at $u > u_B$ function $p(u)/p(bu)$ decreases and

$$\lim_{u\to\infty} p(u)/p(bu) = 0 \quad \forall b,\ 0 \le b < 1.$$

It is obvious, that these conditions, in particular, are true for normal distribution. In this case

$$\lim_{N\to\infty} h_N(\alpha) = -\frac{1}{2}.$$

Table 0.1: Values of coefficients $h_N(\alpha)$ for normal distribution

$N\backslash\alpha$	0.10	0.05	0.02
2	2.6569	5.8531	15.410
4	0.0538	0.2385	0.5409
6	-0.1485	-0.0612	0.0609
8	-0.2224	-0.1616	-0.0830
10	-0.2616	-0.2126	-0.1521
20	-0.3345	-0.3035	-0.2678
30	-0.3596	-0.3337	-0.3042
40	-0.3732	-0.3499	-0.3234

$N\backslash\alpha$	0.10	0.05	0.02
50	-0.3821	-0.3603	-0.3356
60	-0.3884	-0.3678	-0.3444
70	-0.3933	-0.3735	-0.3510
80	-0.3971	-0.3781	-0.3563
90	-0.4003	-0.3818	-0.3606
100	-0.4030	-0.3849	-0.3643
200	-0.4176	-0.4020	-0.3839
300	-0.4243	-0.4097	-0.3929

Table 1 gives values of coefficients $h_N(\alpha) = H_N(\alpha)$ for different N and α at normal probability distributions with a random variance.

For the uniform distribution, simple explicit expressions for functions $\psi_N(h)$ and $h_N(\alpha)$ may be obtained:

$$\psi_N(h) = \frac{1}{2} \cdot \Big(2 \cdot (1+h)\Big)^{-N+1}; \quad h_N(\alpha) = \frac{1}{2} \cdot \alpha^{-1/(N-1)} - 1.$$

DELIMITATION OF SEARCH OF PARAMETERS AT RESTORATION OF NONLINEAR FUNCTIONAL DEPENDENCES

The identification nonlinear regression is an urgent task of mathematical statistics and calculus mathematics at processing of experimental data for the quantitative description of unknown dependences between observable variables. It has numerous application in the most different areas of science and manufacture, being by an irreplaceable means of the quantitative description of cause-effect dependences. Many works are devoted to the decision of the given problem, among which we shall note [6]-[14].

At restoration of functional dependences a method of the least squares frequently is used because of its relative simplicity and reliability. It consists of the following: function $y = f(x)$, given on discrete set of incoincident points $x_1, ..., x_N,\ \ y_l = F(x_l)\ \ (l = 1, ..., N)$ is approximated by function of the given class $f(a_1, ..., a_m,\, x)$; by this values of parameters $a_1, ..., a_m$ are found, for which weighted sum of error squares

$$S = \sum_{l=1}^{N} \lambda_l \cdot |f(a_1, ..., a_m,\, x_l) - y_l|^2$$

takes the minimal value; here λ_l are given weight factors.

If the dependence of approximated function from parameters is nonlinear, for the decision of considered optimization task different iteration methods are used, for example, algorithm of Hook-and-Jivs [12,16].

Let c be one of approximation parameters. When for finding the parameter c the iterative algorithm is used, it is necessary previously to determine an interval of finding of given parameter $[c_{lo}, c_{hi}]$, length of this interval should be in an allowable degree small, and at the same time the probability that value c is included in this interval should be close to unit.

One of universal, not dependent from a concrete type of restored functional dependence, method of determination of intervals of finding of estimations of parameters of approximation is the method offered below which will be called *a method of tests.*

The set of all points (x_j, y_j), $j = 1, ..., N$, *of the plane* (x, y), *appropriate to the measured values, is divided on* L *groups* n *points each in everyone; here* L *is an integral part of the number* N/n. *For each of the given groups, if it is possible, interpolation of function* $f(c_1, ..., c_n, x)$ *is carried out, i.e. such values of parameters* $c_1, ..., c_n$ *are determined, at which plot of function* $f(c_1, ..., c_n, x)$ *passes through all off* n *points of considered group. The sequence of values of parameter* c $[c^{(1)}, ..., c^{(\nu)}]$ *obtained in this way, will be called name as a sequence of tested values of this parameter; each tested value* $c^{(k)}$ *corresponds to one of groups of points* (x_j, y_j), *for which the interpolation of function* $f(c_1, ..., c_n, x)$ *is possible, i.e. the appropriate system of equations has a solution. Length of this sequence, obviously, satisfies relation* $0 \le \nu \le L$. *As bounds of required region* C_{lo} C_{hi} *the minimum and maximum values of elements of sequence* $c^{(k)}$, $k = 1, ..., \nu$.

The described method of determination of region $[C_{lo}, C_{hi}]$, *which should includ the parameter of approximation* c, *gives satisfactory result, if number* ν *is sufficiently big; otherwise, given method requires amendments. A method of determination of region* $[C_{lo}, C_{hi}]$ *described below, will be called the modified method of tests.*

Let C_{min} *and* C_{max} *be accordingly minimum and maximum tested values of parameter* c. *Let's assume, that a sequence of tested values* $[c^{(1)}, ..., c^{(\nu)}]$ *can be considered as the sample appropriate defined on interval* $(-\infty, +\infty)$ *random value* c *with mathematical expectation* $\check{c}$. *As a required interval of finding the parameter* c *the confidential interval undertakes to which includes* $\check{c}$ *with the given level of probability* $1 - \alpha$. *According to the theorem 1 bounds of this interval are equal to*

$$\begin{aligned} C_{lo} &= C_{min} - h_\nu(\alpha) \cdot (C_{max} - C_{min}); \\ C_{hi} &= C_{max} + H_\nu(\alpha) \cdot (C_{max} - C_{min}), \end{aligned} \tag{0.2}$$

where the parameters $h = h_\nu(\alpha)$ *and* $H = H_\nu(\alpha)$ *are defined by formulas 0.1;* $\Phi(c)$ $p(c)$ *are function and density of distribution appropriate* c *normalized random value.*

It is possible to show, that under the rather general assumptions the value c *is normally distributed, that at rather big values of* ν *the function* $h_\nu(\alpha)$ *becomes negative means, that the interval* $[C_{lo}, C_{hi}]$, *determined by modified* *method of tests, does not expand the interval* $[C_{min}, C_{max}]$*, but narrows it. At* $\nu \to \infty$ *the interval* $[C_{lo}, C_{hi}]$ *degenerates in a point which is in the middle of the interval* $[C_{min}, C_{max}]$.

R e f e r e n c e s

[1] Kendall M., Stuart A., The Advanced Theory of Statistics. Moscow: Nauka, 1973, 899.

[2] David G., Ordinal Statistics. Moscow: Nauka, 1979, 335.

[2] Kachiashvili K.J., Stepanishvili V.A., Estimating of Unknown Parameters of Some Probability Distribution Irregular Parameters. Autometria, N42, 1988, 109-111.

[3] Koroliuk V.S., Petrenko N.E., Skorokhod A.V., Turbin A.F., Reference Book on Probability Theory and Mathematical Statistics. Moscow: Nauka, 1985, 640.

[4] Primak A.V., Kafarov V.V., Kachiashvili K.J., Systems Analysis of Control and Management of Air and Water Quality. Kiev: Naukova Dumka, 1991, 360.

[5] Macevich D.A., Petrovich M.L., Fedorov V.V., Comparison of Exactness of Evalution of Regression Model Parameters in the Case of Errors in Independent Variables. //Zavodskaa Laboratoria, N7, 1984, 48-53.

[6] Mosteller F., Tuki Dj., Analysis of Data and Regression. /Per. s angl. M.: Finansi i Statistika, 1982, 239 s.

[7] Kachiashvili K.I., Computation of Real Multi-dimensional Static Gaussian Markovs Series with the Given Depth of Connectedness. // V kn.: Mnogomerniy Statistiqeskiy Analiz i Veroatnostnoe Modelirovanie Real'nix Processov: Uqenie Zapiski po Statistike, 54, M.: Nauka, 1990, 250-254.

[8] Kachiashvili K.I., Improvement of Metrological Characteristic Devices of Control of Environment Pollution Using Microprocessors. Zagraznenia Okrujauwey Sredi s Ispol'zovaniem Mikroprocessorov. Problemi Kontrola Zawiti Atmosferi ot Zagraznenia, 14, 1988, 47-53.

[9] Kachiashvili K.I., Khuchua V.I., Restoration of Polynomial Regression on the Basis of Active Experimants. Zavodskaa Laboratoria N 10, 1990, 87-90.

[10] Kachiashvili K.I., Stepanishvili V.A., Cotoidze D.I., Evsteeva O.A., Automatized Working Places of Ecologists. Problems of Idustry of Ecology and Security. International Conference MP "DIADA", M., 1993, 43.

[11] Primak A.V., Kafarov V.V., Kachiashvili K.I., System Analysis of Control and Administration of Quality of Air and Woter. Kiev: Nauk. dumka, 1991, 360.

[12] Vatutin V.A., Televinova T.M., Chistakov V.P., Methods of Probability in Physical Investigations. M: Nauka, 1985, 207.

[13] Minkov D.P., Mitkov A.L., Todorov T.C., Algorithm for the Construction of one Class of Nonlinear Regressive Models. Zavodskaa Laboratoria, N25, 1986, 64-67.

[14] Koroluk V.S., Portenko N.I., Skorokhod A.V., Turbin A.F., Handbook of Theory of Probability and Mathematical Statistics. M.: Nauka, 1985, 640.

[15] Bazara M., Wetti K., Nelineynoe Programmirovanie: Teoria i Algoritmi. M.: Mir, 1982, 583.

[16] Kachiashvili K., Melikjanian D., Stepanishvili V., Application Package for Experimental Date Processing.// Georgian symposium for project development and conversion, Collection of reparts, Tbilisi, 1995, 143-145.

[17] Kachiashvili K.I., Melikdjanian D.U., Identifikacia Nelineynix Zavisimostey Mejdu Parametrami Ob'ektov Okrujauwey Sredi. // Tezisi dokladov VIII konferencii stran SNG "Problemi ekologii i ekspluatacii ob'ektov energetiki", Kiev, 1998, 41-43.

[18] Kachiashvili K.I., Cotoidze D.I., Algorithms and Program of Computation of Many-Dimensional Integral on the Domain of Layer Configuration from the Normal Density of the Probability distribution by Means of the Monte-Carlo Method. DR 2907 pr – 85. Dep. Bibl. Ukazatel' VINITI, "Deponirovannie Nauchnie Raboti", N9, 1985, 125.