CONSTRUCTION OF CONFIDENCE INTERVAL FOR MATHEMATICAL EXPECTATION OF RANDOM VARIABLES OF A CERTAIN TYPE

K.Kachiashvili, D.Melikdzhanian

Georgian National Center of Environmental Monitoring
150 Agmashenebeli St., 380035 Tbilisi, Georgia

(Received: December 10, 1998; revised: February 15, 1999.)

Abstract

A method of finding the confidence interval for mathematical expectation of random variable is suggested. The given result is compared with classical formulas. The introduced auxiliary functions are studied and tabulated.

Key words and phrases: Confidence intervals,random variable, probability distribution, ordinal statistics, critical regions.

AMS subject classification: 62G05, 62G10

Confidence intervals for parameters of probability distribution laws of the studied phenomena are widely used when solving numerous theoretical and applied problems. The quality of a confidence interval is defined by its width for a given confidence coefficient. There are three principal methods of finding confidence intervals [1]; they are based on: 1) probability frequency theory; 2) fiducial distributions; 3) Bayes' theorem. The first method uses the asymptotic normality of the first derivative of the likelihood function logarithm. According to Wilks' theorem for great-size samples, this method uses fiducial distributions corresponding to the distribution, being considered. In the third method the confidence interval limits are determined on the considered parameter.

Below is given one of the methods of finding the confidence interval for mathematical expectation of a random variable; it is based on usage of ordinal statistics [2,3,5]. In contrast to the above-described ones, this method doesn't require knowledge of other parameters of distribution laws (e.g., variance at normal probability distribution law) when constructing the confidence 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 $p(\dot{x})$ - a function and a distribbution density of the corresponding normalized random variable $\dot{x} = (x-m)/\sigma$; $[x_1, x_2, ..., x_N]$ - N-size sample;

$$x_{min} = \min_{1 \le j \le N}\{x_j\}; \quad x_{max} = \max_{1 \le j \le N}\{x_j\}.$$

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

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

at

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

Proof. let's consider random variables

$$u = (x_{min} - m)/\sigma \quad \text{and} \quad (x_{max} - m)/\sigma.$$

According to [2,4], density of their joint distribution is

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

Confidence interval limits are defined by the following conditions

$$\mathcal{P}\{x_{min} - h_N(\alpha) \cdot (x_{max} - x_{min}) > m\} = \alpha/2;$$

$$\mathcal{P}\{x_{max} + H_N(\alpha) \cdot (x_{max} - x_{min}) < m\} = \alpha/2$$

(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

$$v \cdot h/(1+h) < u \quad \text{and} \quad v \cdot (1+H)/H < u.$$

Regions D_1 and D_2 on plane (uv), shown in fig.1, correspond to them.

Let's denote probabilities of these inequalities $\psi(h)$ and $\Psi(H)$, then

$$\psi(h) = \int_0^{\infty} \left(\int_{vh/(1+h)}^{v} p_w(u,v)\, du \right) dv =$$

$$= \int_0^{\infty} N \cdot p(v) \cdot \Big(\Phi(v) - \Phi(vh/(1+h)) \Big)^{N-1} dv$$

and

$$\Psi(H) = \int_{-\infty}^{0} \left(\int_{u}^{uH/(1+H)} p_w(u,v)\, dv \right) du =$$

$$= \int_0^{\infty} N \cdot p(-u) \cdot \Big(\Phi(-uH/(1+H)) - \Phi(-u) \Big)^{N-1} du,$$

which was to be proved. □

Fig.1. Critical regions D_1, D_2 and hypothesis-acceptance region D_0; $\tan\varphi_1 = h/(1+h)$; $\tan\varphi_2 = H/(1+H)$. Axes of reference are denoted by u and v; region D_0 is bounded by straight lines $v = u\cdot(1+h)/h$ and $v = u \cdot H/(1+H)$; 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 $v = u \cdot H/(1+H)$.

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.1, when each of regions D_0, D_1 and D_2 doesn't change its position with respect to axes of reference.

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

$$\psi(h) \;=\; \int_0^{\infty} N \cdot p(u) \cdot \Big(\Phi(u(1+h)/h) - \Phi(u) \Big)^{N-1} du; \qquad (2)$$

$$\Psi(H) \quad = \quad \int_0^\infty N \cdot p(-u) \cdot \Big(\Phi(-u) - \Phi(-u(1+H)/H)\Big)^{N-1} du.$$

These relations may be obtained in the following way; let $\psi'(h)$ denote the function, which is equal to the right side of the first equation of (2), then

$$\psi(h) - \psi'(h) =$$

$$= \int_0^\infty N \cdot \Big(p(u) - \frac{h}{1+h} \cdot p(uh/(1+h))\Big) \cdot \Big(\Phi(u) - \Phi(uh/(1+h))\Big)^{N-1} du =$$

$$= \Big(\Phi(u) - \Phi(uh/(1+h))\Big)^N \Big|_0^\infty = 0.$$

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 $\alpha \to 0$ $h_N(\alpha) \to \infty$; and at $\alpha \to 1$ $h_N(\alpha) \to -1/2$.

The latter relation may be proved in the following way: at $h = -1/2$

$$\psi(-1/2) = \int_0^\infty N \cdot p(u) \cdot (2\Phi(u) - 1)^{N-1} du = \frac{1}{2} \cdot (2\Phi(u) - 1)^N \Big|_0^\infty = \frac{1}{2}.$$

Table 1 gives values of coefficients $h_N(\alpha) = H_N(\alpha)$ for different N and α for the normal and uniform probability distribution with a random variance. As can be seen from the table, at the fixed α the function $h_N(\alpha)$ decreases when sample size N grows, and becomes negative beginning from a certain value of N.

The value of N, passing through which $h_N(\alpha)$ changes its sign, may be calculated in the following way: at $h = 0$

$$\psi(0) = \int_0^\infty N \cdot p(u) \cdot (\Phi(u) - \Phi(0))^{N-1} du = (\Phi(u) - \Phi(0))^N \Big|_0^\infty = \left(\frac{1}{2}\right)^N.$$

It is obvious, that $\Phi(h)$ is an increasing function. Hence, when $\alpha/2 < 2^{-N}$ $h_N(\alpha) > 0$, and when $\alpha/2 > 2^{-N}$ $h_N(\alpha) < 0$; i.e. the function $h_N(\alpha)$ is negative at $N > \log_2(2/\alpha)$.

Let's consider the limit of $h_N(\alpha)$ at $N \to \infty$ for the fixed α.

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

	Normal distribution			Uniform distribution		
$N\backslash\alpha$	0.10	0.05	0.02	0.10	0.05	0.02
2	2.6569	5.8531	15.410	4.0000	9.0000	24.000
3	0.3968	0.8133	1.6172	0.5811	1.2361	2.5355
4	0.0538	0.2385	0.5409	0.0772	0.3572	0.8420
5	-0.0784	0.0388	0.2133	-0.1109	0.0574	0.3296
6	-0.1485	-0.0612	0.0609	-0.2076	-0.0897	0.0934
7	-0.1922	-0.1213	-0.0265	-0.2661	-0.1762	-0.0403
8	-0.2224	-0.1616	-0.0830	-0.3053	-0.2329	-0.1257
9	-0.2445	-0.1906	-0.1227	-0.3332	-0.2729	-0.1847
10	-0.2616	-0.2126	-0.1521	-0.3542	-0.3025	-0.2278
11	-0.2752	-0.2300	-0.1749	-0.3705	-0.3254	-0.2606
12	-0.2863	-0.2441	-0.1932	-0.3836	-0.3435	-0.2865
13	-0.2956	-0.2557	-0.2082	-0.3942	-0.3582	-0.3073
14	-0.3035	-0.2656	-0.2207	-0.4031	-0.3704	-0.3244
15	-0.3104	-0.2741	-0.2314	-0.4106	-0.3807	-0.3388
16	-0.3164	-0.2815	-0.2406	-0.4170	-0.3895	-0.3510
17	-0.3217	-0.2880	-0.2487	-0.4226	-0.3970	-0.3615
18	-0.3264	-0.2937	-0.2558	-0.4275	-0.4037	-0.3706
19	-0.3306	-0.2989	-0.2621	-0.4318	-0.4095	-0.3786
20	-0.3345	-0.3035	-0.2678	-0.4356	-0.4146	-0.3857
21	-0.3380	-0.3078	-0.2730	-0.4390	-0.4192	-0.3920
22	-0.3412	-0.3116	-0.2777	-0.4421	-0.4233	-0.3976
23	-0.3441	-0.3152	-0.2820	-0.4448	-0.4271	-0.4027
24	-0.3468	-0.3184	-0.2859	-0.4474	-0.4304	-0.4073
25	-0.3493	-0.3214	-0.2895	-0.4497	-0.4335	-0.4115
26	-0.3516	-0.3242	-0.2929	-0.4518	-0.4363	-0.4153
27	-0.3538	-0.3268	-0.2960	-0.4537	-0.4389	-0.4188
28	-0.3559	-0.3293	-0.2989	-0.4555	-0.4413	-0.4220
29	-0.3578	-0.3316	-0.3017	-0.4571	-0.4435	-0.4250
30	-0.3596	-0.3337	-0.3042	-0.4587	-0.4456	-0.4278
40	-0.3732	-0.3499	-0.3234	-0.4696	-0.4601	-0.4472
50	-0.3821	-0.3603	-0.3356	-0.4759	-0.4685	-0.4584
60	-0.3884	-0.3678	-0.3444	-0.4801	-0.4740	-0.4657
70	-0.3933	-0.3735	-0.3510	-0.4830	-0.4778	-0.4708
80	-0.3971	-0.3781	-0.3563	-0.4852	-0.4807	-0.4746
90	-0.4003	-0.3818	-0.3606	-0.4869	-0.4829	-0.4775
100	-0.4030	-0.3849	-0.3643	-0.4882	-0.4846	-0.4798
200	-0.4176	-0.4020	-0.3839	-0.4942	-0.4924	-0.4901
300	-0.4243	-0.4097	-0.3929	-0.4961	-0.4950	-0.4934

Theorem 2 *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*

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

then

$$\lim_{N\to\infty} h_N(\alpha) = -\frac{1}{2}. \tag{3}$$

It is obvious, that the theorem conditions are valid for normal distribution.

Proof. In accordance with the conditions of the theorem, function $p(u)$ is strictly monotonic at $u > u_B$. Let's denote

$$s_N = 1 + \frac{h_N(\alpha)}{1 + h_N(\alpha)} \quad \text{and} \quad u_N = \Phi^{-1}(1 - N^{-1/2}),$$

where $\Phi^{-1}(.)$ is the reverse function of $\Phi(.)$. Assume that $N > (1 - \Phi(u_B))^{-1/2}$ and, consequently, $u_N > u_B$, and also assume that $h_N(\alpha) < 0$ and, consequently, $0 < s_N < 1$. Let's divide the integration interval in the expression defining $\psi(h)$ into two subintervals; we shall have

$$\alpha = 2\cdot\psi(h_N(\alpha)) = \int_0^{u_N} 2N\cdot p(u)\cdot(\Phi(u)+\Phi(u-u\cdot s_N)-1)^{N-1}\,du+$$

$$+\int_{u_N}^{\infty} 2N\cdot p(u)\cdot(\Phi(u)+\Phi(u-u\cdot s_N)-1)^{N-1}\,du \le$$

$$\le \int_0^{u_N} 2N\cdot p(u)\cdot(2\Phi(u)-1)^{N-1}\,du+$$

$$+\int_{u_N}^{\infty} \frac{2p(u)}{p(u)+(1-s_N)\cdot p((1-s_N)\cdot u)}\cdot\frac{d}{du}(\Phi(u)+\Phi(u-u\cdot s_N)-1)^N\,du \le$$

$$\le \int_0^{u_N} \frac{d}{du}(2\Phi(u)-1)^N\,du+$$

$$+\frac{2p(u_N)}{p(u_N)+(1-s_N)\cdot p((1-s_N)\cdot u_N)}\cdot\int_{u_N}^{\infty}\frac{d}{du}(\Phi(u)+\Phi(u-u\cdot s_N)-1)^N\,du =$$

$$= (2\Phi(u_N)-1)^N+$$

$$+\frac{2p(u_N)}{p(u_N)+(1-s_N)\cdot p((1-s_N)\cdot u_N)}\cdot\Big(1-(\Phi(u_N)+\Phi(u_N-u_N\cdot s_N)-1)^N\Big)\le$$

$$\le(2\Phi(u_N)-1)^N+\frac{2}{1+(1-s_N)\cdot p((1-s_N)\cdot u_N)/p(u_N)}.$$

The first summand in the right side of this equation

$$(2\Phi(u_N)-1)^N=(1-2N^{-1/2})^{-N}\to 0 \quad \text{at} \quad N\to\infty.$$

Let's consider the second summand. Taking into consideration that sequence u_N infinitely increases when $N\to\infty$, we see, that, if s_N has a different from zero limit, the considered value tends to zero when $N\to\infty$. However, parameter α must fulfil the condition $0<\alpha<1$. Hence, sequence s_N must converge to zero and

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

which was to be proved. □

Note, that the theorem condition is sufficient but not necessary, because for some widely spread distributions, such as, for example, the uniform distribution, the theorem condition is not filfilled, but (3) still holds.

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

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

If value x is distributed normally, it is reasonable to use Laguerre quadrature formula for calculation of function $\psi(h)=\Psi(h)$

$$\psi(h)=\int_0^\infty e^{-x^2/2}\cdot f(x)\,dx=\int_0^\infty e^{-y}\cdot(2y)^{-1/2}f(\sqrt{2y})\,dy\approx$$

$$\approx\sum_{k=1}^{n} w_k\cdot\frac{f(\sqrt{2\xi_k})}{\sqrt{2\xi_k}},$$

where

$$f(x)=N\cdot\frac{1}{\sqrt{2\pi}}\cdot\Big(\Phi_{norm}(x)-\Phi_{norm}(xh/(1+h))\Big)^{N-1};$$

$\Phi_{norm}(x)$ – a function of standardized normal distribution; $\xi_k, k = 1, ..., n$ – zeroes of Laguerre polynomial $L_n^{(0)}(x)$ of degree n; $w_k = \frac{1}{\xi_k \cdot \left(\dot{L}_n^{(0)}(\xi_k)\right)^2}$ – weight coefficients of the quadrature formula.

If the number of nodal points n is taken to be equal to 16, it will give the accuracy of calculation of function $h_N(\alpha)$ negative values, with which they are presented in table 1.

Figures 2 and 3 show the length of confidence interval, calculated according to the classical method and the method given above, plotted versus, correspondingly, the sample size at fixed variance of the observed results and the standard deviation at fixed sample size for the normal distribution. Figures 4 and 5 show the analogous plots for the uniform probability distribution law.

Fig.2. Length of confidence interval versus sample size for the normal probability distribution law, at $m = 0$, $\sigma = 1$; 1 – the classical method; 2 – the new method.

Fig.3. Length of confidence interval versus standard deviation for the normal

probability distribution law, at sample size $N = 50$; 1 – the classical method; 2 – the new method.

Fig.4. Length of confidence interval versus sample size for the uniform probability distribution law, at $a = 0$, $b = 5$; 1 – the classical method; 2 – the new method.

Fig.5. Length of confidence interval versus length of the interval, on which a random variable is defined, for the uniform probability distribution law, at sample size $N = 50$; 1 – the classical method; 2 – the new method.

Calculation of length of one confidence interval value was performed by averaging of three-fold calculation of its value on the base of three independent samples of necessary size for the given parameters.

It may be seen from these plots, that the classical method gives a better result for the normal distribution. This was bound to be so according to Wilks theorem [1], as the likelihood function in this case is distributed normally. On the other hand, this assumption does not correspond to reality in case of the uniform probability distribution law. On account of this, the proposed method, in which coefficients $h_N(\alpha)$ are calculated from probability distribution laws of random variables, gives much better result that the classical one.

Thus, the proposed method, besides simplicity of calculations, gives a better result than the classical one for construction of confidence interval of mathematical expectation of a random variable, if the latter has the probability distribution law other than the normal one.

Acknowledgement

The work is fulfilled within project G-047 of the International Science and Technology Center.

References

1. Kendall M., Stuart A. *The Advanced Theory of Statistics.* Nauka, 1973.
2. David G. *Ordinal Statistics.* Moscow, Nauka 1979.
3. Kachiashvili K.J., Stepanishvili V.A. *Estimating of Unknown Parameters of Some Probability Distribution Irregular Parameters.* Autometria, 42 (1988), 109-111.
4. Koroliuk V.S., Petrenko N.E., Skorokhod A.V., Turbin A.F. *Reference Book on Probability Theory and Mathematical Statistics.* Moscow, Nauka, 1985.
5. Primak A.V., Kafarov V.V., Kachiashvili K.J. *Systems analysis of Control and Management of Air and Water Quality.* Kiev, Naukova Dumka, 1991.