УДК 621.391

Р.Б. Трегубов, М.В. Стремоухов

ЗАДАЧА ОЦЕНИВАНИЯ ПАРАМЕТРА БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПО ОГРАНИЧЕННОМУ ЧИСЛУ ОПЫТОВ

Цель настоящего исследования найти аппроксимацию интервала Клоппера–Пирсона при условии отказа от допущения о том, что частота появления события в n независимых опытах (схема Бернулли) распределена по нормальному закону. Практика показывает, что такое допущение обуславливает серьезные погрешности статистического анализа для редких событий в условиях ограниченного числа опытов. Указанная задача в работе решается путем аппроксимации точного решения уравнений Клоппера–Пирсона с помощью полинома шестой степени. В свою очередь для получения точного решения уравнений Клоппера–Пирсона в работе использовался численный метод бисекции (метод деления отрезка пополам), реализованный в среде математического моделирования Mathcad. Величина модуля ошибки предлагаемой полиномиальной аппроксимации (в условиях ограниченного числа n независимых опытов) не превышает значения $5 \cdot 10^{-3}$. В свою очередь, для известных аппроксимаций (в тех же условиях) величина модуля ошибки аппроксимации значительно больше, что подтверждается результатами математического моделирования в среде Mathcad. Основные результаты исследования представлены в виде таблиц коэффициентов аппроксимирующих полиномов для различных значений доверительных вероятностей – β (0,9; 0,95; 0,99; 0,995 и 0,999) и числа испытаний – n (10; 20; 30; …; 100), при этом значения коэффициентов полиномов для определения левой и правой границ параметра биномиального распределения совпадают. Отличительной особенностью предлагаемого в работе метода расчета границ доверительного интервала является то, что, во-первых, порядок аппроксимирующего полинома не зависит от числа испытаний, а его коэффициенты от того какая граница рассчитывается; а во-вторых, исключается необходимость в использовании таблиц биномиального распределения или аппроксимирующих его бета-распределения, F-распределения, нормального распределения и распределения Пуассона. Полученные результаты могут найти применение в задачах анализа вероятностно-временных характеристик (вероятности потерь протокольных блоков данных по перегрузкам, ошибкам, несвоевременности доставки и др.) инфокоммуникационных систем различного назначения или их имитационных моделей.

Вероятность; частота; точечная оценка; интервальная оценка; доверительный интервал; доверительная вероятность; биномиальное распределение; уравнения Клоппера–Пирсона.

R.B. Tregubov, M.V. Stremouhov

BOUNDED QUANTITATIVELY EXPERIMENT BINOMIAL DISTRIBUTION PARAMETER ESTIMATION PROBLEM

Aim of this study is to find an approximation interval Clopper–Pearson provided out of the assumption that the frequency of occurrence of an event in n independent experiments (Bernoulli scheme) normal distribution. The practice shows that this assumption causes a serious error for the statistical analysis of rare events in a limited number of experiments. This object is achieved in the work by approximation the exact solution of the equations Clopper–Pearson with the help of sixth – degree polynomial. In turn, to obtain an accurate solution of the equations Clopper–Pearson used in the numerical method of bisection (method of bisection of the interval), implemented in an environment of mathematical modeling Mathcad. The value of the module of the proposed polynomial approximation (in a limited number n of independent experiments) error doesn't exceed $5 \cdot 10^{-3}$. In turn, for the known approximation (in the same case) of the value of the module approximation error is much more, as evidence by the results of mathematical modeling Mathcad. The main results of the study are presented in tabular form coefficients of the approximating polynomials for different values of the confidence probability – β (0,9; 0,95; 0,99; 0,995 and 0,999) and number of tests – n (10; 20; 30 … 100), the values of the polynomial to

determine the left and right boundaries of the same parameter of the binominal distribution. A distinctive feature of the proposed method in the calculation of the boundaries of the confidence interval is that, firstly, the order of the approximating polynomial doesn't depend on the number of tests, and the coefficients of the boundary which is calculated; and secondly, eliminating the need to use the binomial distribution tables or approximating its beta distribution, F-distribution, normal distribution and Poisson distribution. The results can be applied for the analysis of probabilistic – temporal characteristics (loss probability PDUs transshipment, errors, not timely delivery etc.), communication systems for various purposes or their simulation models.

Probability;relative frequency; point estimate; interval estimate; confidence interval; confidence probability; binomial distribution; Clopper–Pearson equation.

Введение. В процессе обработки результатов измерений вероятностно-временных характеристик сложных систем или их имитационных моделей нередко исследователи работают со статистическим материалом весьма ограниченного объема. В этом случае, как правило, решается задача определения точечных и интервальных оценок для соответствующих показателей качества [1–6].

Рассматривается задача нахождения доверительного интервала параметра биномиального распределения по ограниченному числу опытов. Цель настоящего исследования найти аппроксимацию интервала Клоппера–Пирсона при условии отказа от допущения о том, что частота появления события в n независимых опытах (схема Бернулли) распределена по нормальному закону.

1. Постановка задачи. Если имеется реализация $X_1, X_2, \ldots, X_n$ из n испытаний, в которых событие A наблюдалось m раз (случайная величина X_i в каждом отдельном опыте принимает значение 1, если событие A появилось, и 0, если не появилось), то несмещенной и эффективной оценкой вероятности p события A является его частота p^* [7, 8]

$$p^* = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{m}{n}. \tag{1}$$

Известно, что интервальной оценкой (с доверительной вероятностью β) неизвестной вероятности p биномиального распределения по частоте p^* служит доверительный интервал (с приближенными границами p_1 и p_2) [7–10]

$$p_1 = \left(\frac{n}{t_\beta^2 + n}\right) \cdot \left(p^* + \frac{t_\beta^2}{2n} - t_\beta \sqrt{\frac{p^*(1-p^*)}{n} + \left(\frac{t_\beta}{2n}\right)^2}\right), \tag{2}$$

$$p_2 = \left(\frac{n}{t_\beta^2 + n}\right) \cdot \left(p^* + \frac{t_\beta^2}{2n} + t_\beta \sqrt{\frac{p^*(1-p^*)}{n} + \left(\frac{t_\beta}{2n}\right)^2}\right), \tag{3}$$

где

$$t_\beta = \arg\Phi^*\left(\frac{1+\beta}{2}\right), \tag{4}$$

в свою очередь $\arg\Phi^*$ – функция, обратная нормальной функции распределения [7]

$$\Phi^*(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt. \tag{5}$$

Учитывая, что с ростом n величины $\frac{t_\beta^2}{n}$ и $\left(\frac{t_\beta}{2n}\right)^2$ стремятся к нулю, в пределе формулы (2) и (3) принимают вид [7, 10–12]

$$p_1 = p^* - t_\beta \sqrt{\frac{p^*\left(1-p^*\right)}{n}}, \tag{6}$$

$$p_2 = p^* + t_\beta \sqrt{\frac{p^*\left(1-p^*\right)}{n}}. \tag{7}$$

В работах [13, 14] предложена модификация выражений (6) и (7), позволяющая учитывать разное смещение границ p_1 и p_2 относительно частоты p^*

$$p_1 = \frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2} - t_\beta \sqrt{\frac{\left(\frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2}\right) \cdot \left(1 - \frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2}\right)}{n + t_\beta^2}}, \tag{8}$$

$$p_2 = \frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2} + t_\beta \sqrt{\frac{\left(\frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2}\right) \cdot \left(1 - \frac{p^* \cdot n + 0{,}5 \cdot t_\beta^2}{n + t_\beta^2}\right)}{n + t_\beta^2}}. \tag{9}$$

Однако использование выражений (2)–(9) допустимо только в случае справедливости допущения о том, что частота p^* есть случайная величина, распределение которой близко к нормальному [7].

Также следует отметить, что границы p_1 и p_2 могут быть получены с помощью аппроксимации биномиального распределения бета-распределением [6, 8, 10, 15]

$$p_1 = B^{-1}\left(\frac{\alpha}{2}, p^* \cdot n + \frac{1}{2}, n - p^* \cdot n + \frac{1}{2}\right), \tag{10}$$

$$p_2 = B^{-1}\left(1 - \frac{\alpha}{2}, p^* \cdot n + \frac{1}{2}, n - p^* \cdot n + \frac{1}{2}\right), \tag{11}$$

где $B^{-1}\left(p, s_1, s_2\right)$ – это квантили обратного бета-распределения

$$f\left(x, s_1, s_2\right) = \frac{x^{s_1-1} \cdot \left(1-x\right)^{s_2-1}}{\int_0^1 x^{s_1-1} \cdot \left(1-x\right)^{s_2-1} d\,x} \tag{12}$$

с параметрами s_1 и s_2.

В процессе измерения вероятностно-временных характеристик инфокоммуникационных систем различного назначения или их имитационных моделей условия проводимых экспериментов (в ряде случаев) не позволяют воспользоваться допущением о том, что частота p^* есть случайная величина, распределение которой близко к нормальному (малое число опытов, вероятность p стремится к 0 или 1). В этом случае интервальная оценка (с доверительной вероятностью β) неизвестной вероятности p определяется путем решения уравнений Клоппера–Пирсона [1, 6, 10–16]

$$P(Y \le y, n, p_2) = \frac{1-\beta}{2}, \quad (13)$$

$$1 - P(Y \le y-1, n, p_1) = \frac{1-\beta}{2}, \quad (14)$$

где $P(Y \le y, n, p_2)$ – вероятность того, что в n испытаниях событие A (случающееся с вероятностью p_2) будет наблюдаться меньше или равно y раз

$$P(Y \le y, n, p_2) = \sum_{m=0}^{y} C_m^n p_2^m (1-p_2)^{n-m}, \quad (15)$$

в свою очередь $1 - P(Y \le y-1, n, p_1)$ – вероятность того, что в n испытаниях событие A (случающееся с вероятностью p_1) будет наблюдаться больше или равно y раз

$$1 - P(Y \le y-1, n, p_1) = \sum_{m=y}^{n} C_m^n p_1^m (1-p_1)^{n-m}. \quad (16)$$

На рис. 1 представлено графическое решение уравнений (8) и (9) для случая, когда $\beta = 0{,}9$, $n = 50$, и $y = 20$. В данном примере доверительный интервал ограничивается точками $p_1 = 0{,}283$ и $p_2 = 0{,}526$.

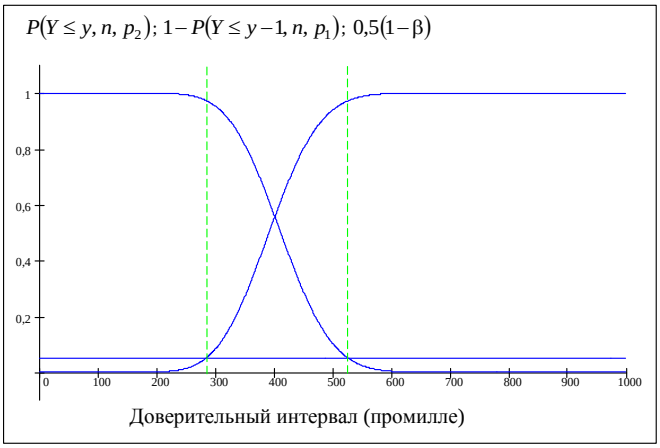

Рис. 1. Графическое решение уравнений Клоппера–Пирсона

Значения p_1 и p_2, соответствующие различным n и β, приведены в [17–20]. К сожалению, литература [17–20], содержащая такие справочные данные, зачастую оказывается недоступной, а значения p_1 и p_2 представлены лишь для некоторых значений доверительной вероятности (0,9; 0,95 и 0,995), чего может оказаться недостаточно для решения практических задач моделирования или проектирования инфокоммуникационных систем различного назначения.

2. Точный метод оценки параметра биномиального распределения. Для получения точного решения уравнений (8) и (9) в работе был применен численный метод бисекции (метод деления отрезка пополам), реализованный в среде математического моделирования *Mathcad*. Ниже представлены алгоритмы нахождения левой и правой границы параметра биномиального распределения.

Алгоритм нахождение правой границы p_2 *параметра биномиального распределения*

Вход:

параметр n – число независимых опытов;

параметр m – число появлений события A;

параметр β – доверительная вероятность;

параметр Δ – требуемая точность решения;

Выход:

параметр p_2 – значение правой границы доверительного интервала;

1: определить начальные границы и размер интервала

$p_2^{\text{н}} = \frac{m}{n},\ p_2^{\text{в}} = 1,\ \Delta p_2 = p_2^{\text{в}} - p_2^{\text{н}};$

2: **пока** $\Delta p_2 > \Delta$

3: определить новую границу интервала $p_2^{\text{пром}} = p_2^{\text{н}} + \frac{\Delta p_2}{2};$

4: **если** $P\left(Y \le m, n, p_2^{\text{пром}}\right) > \frac{1-\beta}{2}$ **то** $p_2^{\text{н}} = p_2^{\text{пром}};$

5: **если** $P\left(Y \le m, n, p_2^{\text{пром}}\right) < \frac{1-\beta}{2}$ **то** $p_2^{\text{в}} = p_2^{\text{пром}};$

6: **если** $P\left(Y \le m, n, p_2^{\text{пром}}\right) = \frac{1-\beta}{2}$ **то** $p_2^{\text{в}} = p_2^{\text{пром}}$ **выход** из цикла;

7: определить новый размер интервала $\Delta p_2 = p_2^{\text{в}} - p_2^{\text{н}};$

8: определить значение правой границы доверительного интервала $p_2 = p_2^{\text{в}}.$

Алгоритм нахождение левой границы p_1 *параметра биномиального распределения*

Вход:

параметр n – число независимых опытов;

параметр m – число появлений события A;

параметр β – доверительная вероятность;

параметр Δ – требуемая точность решения;

Выход:

параметр p_1 – значение левой границы доверительного интервала;

1: определить начальные границы и размер интервала

$p_1^{\text{н}} = 0,\ p_1^{\text{в}} = \frac{m}{n},\ \Delta p_1 = p_1^{\text{в}} - p_1^{\text{н}};$

2: пока $\Delta p_1 > \Delta$

3: определить новую границу интервала $p_1^{\text{пром}} = p_1^{\text{н}} + \frac{\Delta p_1}{2};$

4: если $1 - P\left(Y \le m-1, n, p_1^{\text{пром}}\right) > \frac{1-\beta}{2}$ то $p_1^{\text{в}} = p_1^{\text{пром}};$

5: если $1 - P\left(Y \le m-1, n, p_1^{\text{пром}}\right) < \frac{1-\beta}{2}$ то $p_1^{\text{н}} = p_1^{\text{пром}};$

6: если $1 - P\left(Y \le m-1, n, p_1^{\text{пром}}\right) = \frac{1-\beta}{2}$ то $p_1^{\text{н}} = p_1^{\text{пром}}$ выход из цикла;

7: определить новый размер интервала $\Delta p_1 = p_1^{\text{в}} - p_1^{\text{н}}$;

8: определить значение левой границы доверительного интервала $p_1 = p_1^{\text{н}}$.

На рис. 2 представлены результаты точного решения уравнений (8) и (9) для различных n (10, 50 и 100) при доверительной вероятности $\beta = 0{,}9$. По оси абсцисс откладывается частота p^* (промилле), по оси ординат – вероятность p. Точки одного цвета, лежащие на одной вертикали, определяют доверительный интервал вероятностей, отвечающий заданному значению частоты p^*.

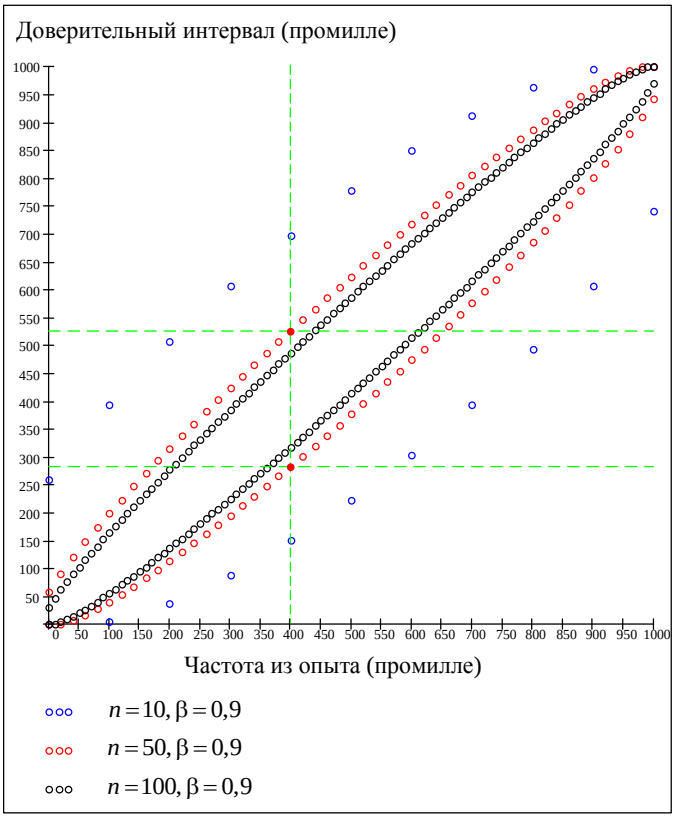

Рис. 2. Численное решение уравнений Клоппера–Пирсона

Разработанные алгоритмы позволили исследовать зависимость величины доверительного интервала от числа опытов (рис. 3). Из рис. 3 видно, что доверительный интервал, рассчитанный по выражениям (2) и (3) несколько меньше чем доверительный интервал, полученный в результате точного решения уравнений (8) и (9). Следовательно, можно сделать вывод о том, что применение существующего инструментария приведёт к получению неточного решения. Так при анализе вероятностно-временных характеристик инфокоммуникационных систем различного назначения это может способствовать неоправданно оптимистичным выводам о качестве обслуживания протокольных блоков данных.

3. Аппроксимация точного решения уравнений Клоппера–Пирсона. Точное решение уравнений (8) и (9) для различных условий функционирования инфокоммуникационных систем различного назначения требует значительных временных затрат для получения результата, что не всегда допустимо на практике. Для снижения временных затрат, связанных с получением границ доверительного интервала в работе была применена полиномиальная аппроксимация.

Так исследования показали, что точное решение уравнений Клоппера–Пирсона (выражения (8) и (9)) может быть достаточно хорошо аппроксимировано полином шестой степени. В табл. 1–10 представлены коэффициенты аппроксимирующего полинома шестой степени для различных значений параметра n и доверительной вероятности β.

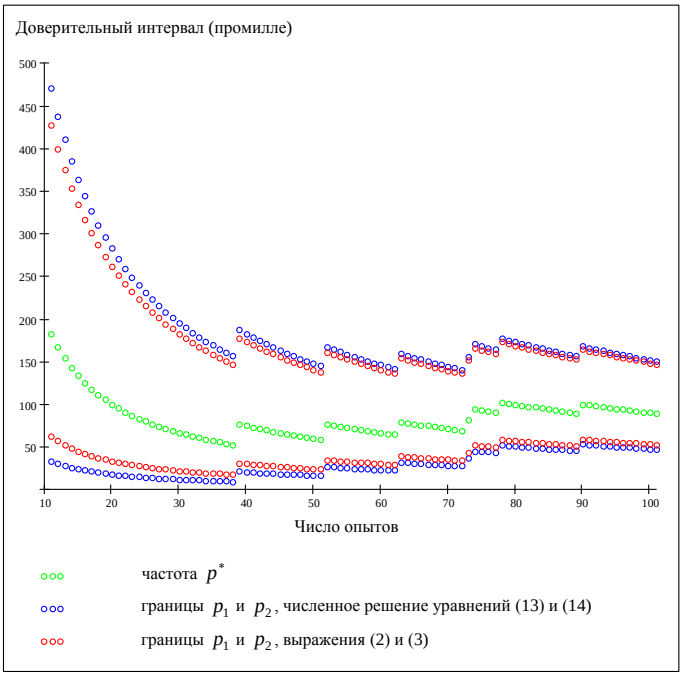

Рис. 3. Зависимость величины доверительного интервала от числа опытов

Таблица 1

Коэффициенты полинома для $\beta = 0,9$ (начало)

n	10	20	30	40	50
$k0_{\beta,n}$	0,2589	0,1393	0,0954	0,0728	0,0591
$k1_{\beta,n}$	0,5038	0,6452	0,6603	0,6445	0,6204
$k2_{\beta,n}$	-1,8121	-2,6847	-2,9435	-2,9665	-2,9032
$k3_{\beta,n}$	3,2114	6,9209	8,2349	8,5525	8,4869
$k4_{\beta,n}$	-4,1424	-11,277	-13,8516	-14,5094	-14,4355
$k5_{\beta,n}$	2,9607	9,5693	11,9163	12,5051	12,4343
$k6_{\beta,n}$	-0,9804	3,3128	-4,111	-4,2976	-4,26

Таблица 2

Коэффициенты полинома для $\beta = 0,9$ (окончание)

n	60	70	80	90	100
$k0_{\beta,n}$	0,0499	0,0433	0,0384	0,0345	0,0314
$k1_{\beta,n}$	0,5952	0,571	0,5487	0,5281	0,5095
$k2_{\beta,n}$	-2,8114	-2,7127	-2,6161	-2,5233	-2,438
$k3_{\beta,n}$	8,2784	8,0213	7,7542	7,4891	7,2419
$k4_{\beta,n}$	-14,0874	-13,6463	-13,1825	-12,721	-12,2903
$k5_{\beta,n}$	12,1197	11,725	11,3116	10,9021	10,5209
$k6_{\beta,n}$	-4,1418	-3,9987	-3,8511	-3,7063	-3,572

Таблица 3

Коэффициенты полинома для $\beta = 0,95$ (начало)

n	10	20	30	40	50
$k0_{\beta,n}$	0,3086	0,1688	0,1164	0,0891	0,0724
$k1_{\beta,n}$	0,5257	0,7102	0,7397	0,7297	0,7077
$k2_{\beta,n}$	-1,8964	-2,8517	-3,1953	-3,2716	-3,2401
$k3_{\beta,n}$	2,8526	6,7787	8,492	9,0846	9,1965
$k4_{\beta,n}$	-2,8849	-10,3541	-13,7729	-15,0229	-15,3363
$k5_{\beta,n}$	1,4206	8,3436	11,5324	12,7094	13,0236
$k6_{\beta,n}$	-0,3261	-2,7956	-3,912	-4,3175	-4,4224

Таблица 4

Коэффициенты полинома для $\beta = 0,95$ (окончание)

n	60	70	80	90	100
$k0_{\beta,n}$	0,0612	0,0532	0,0471	0,0424	0,0386
$k1_{\beta,n}$	0,6827	0,6577	0,634	0,612	0,5917
$k2_{\beta,n}$	-3,1661	-3,0752	-2,9811	-2,8887	-2,8009
$k3_{\beta,n}$	9,0997	8,907	8,6792	8,4397	8,204
$k4_{\beta,n}$	-15,2379	-14,9473	-14,5828	-14,1884	-13,796
$k5_{\beta,n}$	12,959	12,7177	12,4085	12,0704	11,7328
$k6_{\beta,n}$	-4,3967	-4,3108	-4,2023	-4,0844	-3,967

Таблица 5

Коэффициенты полинома для $\beta = 0,99$ (начало)

n	10	20	30	40	50
$k0_{\beta,n}$	0,4114	0,2333	0,163	0,1256	0,1025
$k1_{\beta,n}$	0,5088	0,7905	0,8557	0,8626	0,849
$k2_{\beta,n}$	-2,0894	-3,093	-3,5494	-3,7214	-3,7559
$k3_{\beta,n}$	2,81	6,4275	8,5786	9,6214	10,0698
$k4_{\beta,n}$	-2,1847	-8,6	-12,8854	-15,0842	-16,1205
$k5_{\beta,n}$	0,3189	6,065	10,1126	12,2327	13,2658
$k6_{\beta,n}$	0,2248	-1,8239	-3,2758	-4,037	-4,4106

Таблица 6

Коэффициенты полинома для $\beta = 0,99$ (окончание)

n	60	70	80	90	100
$k0_{\beta,n}$	0,0868	0,0754	0,0669	0,0601	0,0547
$k1_{\beta,n}$	0,8278	0,8043	0,7804	0,7575	0,7357
$k2_{\beta,n}$	-3,7237	-3,6608	-3,5822	-3,4989	-3,4148
$k3_{\beta,n}$	10,2073	10,1832	10,0663	9,9074	9,7242
$k4_{\beta,n}$	-16,5367	-16,6213	-16,5117	-16,3095	-16,0484
$k5_{\beta,n}$	13,7162	13,8517	13,8018	13,6617	13,4621
$k6_{\beta,n}$	-4,5771	-4,6315	-4,6201	-4,5765	-4,5115

Таблица 7

Коэффициенты полинома для $\beta = 0,995$ (начало)

n	10	20	30	40	50
$k0_{\beta,n}$	0,4508	0,2595	0,1823	0,1409	0,1151
$k1_{\beta,n}$	0,484	0,8059	0,8867	0,9015	0,8924
$k2_{\beta,n}$	-2,166	-3,1681	-3,6485	-3,8482	-3,9067
$k3_{\beta,n}$	2,9904	6,3416	8,5374	9,694	10,2542
$k4_{\beta,n}$	-2,4629	-8,1184	-12,4511	-14,8806	-16,1515
$k5_{\beta,n}$	0,5093	5,4149	9,5001	11,8492	13,1152
$k6_{\beta,n}$	0,1942	-1,536	-3,0078	-3,8575	-4,3189

Таблица 8

Коэффициенты полинома для $\beta = 0,995$ (окончание)

n	60	70	80	90	100
$k0_{\beta,n}$	0,0976	0,0848	0,0752	0,0677	0,0616
$k1_{\beta,n}$	0,8737	0,8515	0,8286	0,8058	0,784
$k2_{\beta,n}$	-3,8921	-3,8401	-3,7722	-3,6944	-3,6135
$k3_{\beta,n}$	10,4794	10,5165	10,4581	10,3352	10,178
$k4_{\beta,n}$	-16,7551	-16,9746	-16,9921	-16,8712	-16,6719
$k5_{\beta,n}$	13,752	14,0235	14,0997	14,0419	13,9061
$k6_{\beta,n}$	-4,5553	-4,661	-4,6963	-4,6835	-4,6426

Таблица 9

Коэффициенты полинома для $\beta = 0,999$ (начало)

n	10	20	30	40	50
$k0_{\beta,n}$	0,5324	0,3169	0,2253	0,1752	0,1436
$k1_{\beta,n}$	0,4023	0,8144	0,9311	0,9648	0,9669
$k2_{\beta,n}$	-2,2889	-3,3122	-3,8155	-4,0588	-4,1606
$k3_{\beta,n}$	3,4826	6,3107	8,42	9,7144	10,4553
$k4_{\beta,n}$	-3,499	-7,5697	-11,5837	-14,2604	-15,8977
$k5_{\beta,n}$	1,4935	4,5696	8,3016	10,8894	12,5167
$k6_{\beta,n}$	-0,1229	-1,1304	-2,48	-3,4257	-4,0253

Таблица 10

Коэффициенты полинома для $\beta = 0,999$ (окончание)

n	60	70	80	90	100
$k0_{\beta,n}$	0,122	0,1062	0,0942	0,0848	0,0772
$k1_{\beta,n}$	0,9551	0,9374	0,917	0,8959	0,875
$k2_{\beta,n}$	-4,1809	-4,1571	-4,1082	-4,046	-3,977
$k3_{\beta,n}$	10,8419	11,0177	11,0624	11,0263	10,94
$k4_{\beta,n}$	-16,839	-17,3534	-17,5966	-17,6661	-17,6241
$k5_{\beta,n}$	13,4884	14,0527	14,3559	14,4888	14,5112
$k6_{\beta,n}$	-4,3882	-4,6037	-4,7246	-4,7833	-4,8015

В этом случае интервальной оценкой (с доверительной вероятностью β) неизвестной вероятности p биномиального распределения служит доверительный интервал (с приближенными границами p_1 и p_2) определяемый следующими выражениями

$$p_1 = p^* - k0_{\beta,n} - (1-p^*)\cdot k1_{\beta,n} - (1-p^*)^2 \cdot k2_{\beta,n} - (1-p^*)^3 \cdot k3_{\beta,n} - (1-p^*)^4 \cdot k4_{\beta,n} - (1-p^*)^5 \cdot k5_{\beta,n} - (1-p^*)^6 \cdot k6_{\beta,n}. \quad (17)$$

$$p_2 = p^* + k0_{\beta,n} + p^* \cdot k1_{\beta,n} + (p^*)^2 \cdot k2_{\beta,n} + (p^*)^3 \cdot k3_{\beta,n} + (p^*)^4 \cdot k4_{\beta,n} + (p^*)^5 \cdot k5_{\beta,n} + (p^*)^6 \cdot k6_{\beta,n}, \quad (18)$$

Также следует отметить, что поскольку биномиальное распределение может быть достаточно точно аппроксимировано с помощью *F*-распределения, нормального распределения и распределения Пуассона [8] следовательно, значения границ p_1 и p_2 можно выразить и через квантили этих распределений [8, 21].

Заключение. В работе отражены результаты исследований, позволяющие найти доверительный интервал параметра биномиального распределения, по статистическому материалу ограниченного объема, в условиях отказа от допущения о том, что частота появления события в n независимых опытах (схема Бернулли) распределена по нормальному закону. Показана возможность снижения вычислительной сложности получения точного решения уравнений Клоппера–Пирсона за счет его аппроксимации полиномом шестой степени. Отмечено, что величина модуля ошибки предлагаемой полиномиальной аппроксимации (в условиях ограниченного числа n независимых опытов) не превышает значения $5\cdot10^{-3}$. В свою очередь, для известных аппроксимаций [9–15] (в тех же условиях) величина модуля ошибки аппроксимации значительно больше, что подтверждается результатами математического моделирования в среде *Mathcad* (рис. 4).

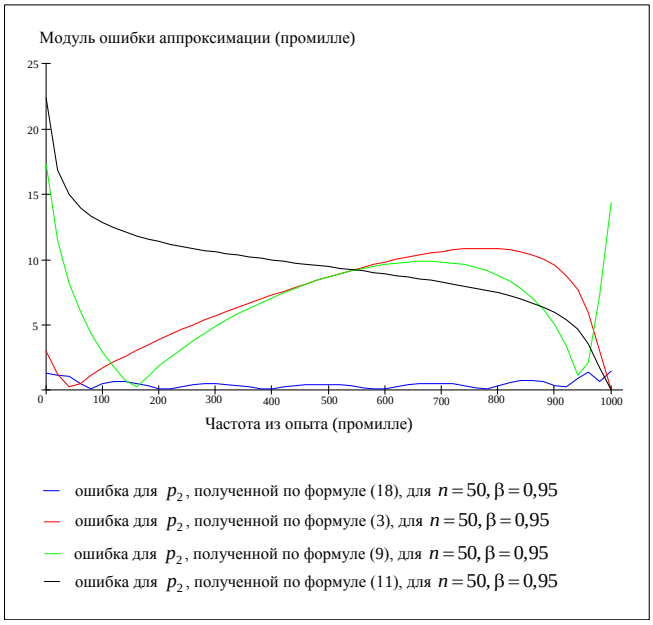

Рис. 4. Модуль ошибки аппроксимации доверительного интервала

Полученные результаты могут найти применения в задачах анализа вероятностно-временных характеристик (вероятности потерь протокольных блоков данных по перегрузкам, ошибкам, несвоевременности доставки и др.) инфокоммуникационных систем различного назначения или их имитационных моделей.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Thulin M.* The cost of using exact confidence intervals for a binomial proportion // Electronic Journal of Statistics. – 2014. – Vol. 8. – P. 817-840.
2. *Иванов Н.Н., Стрельников В.П.* Прогнозирование остаточной долговечности паяных соединений // Математичні машини і системи. – 2012. – № 3. – С. 162-165.
3. *Кузнецов А.Г. Александровская Л.Н.* Непараметрические методы "измерения" малых рисков в задачах оценки соответствия требований к безопасности автоматической посадки самолетов нормам летной годности // Труды Московского института электромеханики и автоматики (МИЭА). – 2011. – Вып. 3. – С. 2-11.
4. *Гусев Л.А.* Об интерпретации неразличимости в задаче интервальной оценки неизвестной вероятности // Автоматика и телемеханика. – 2010. – Вып. 8. – С. 38-48.
5. *Гусев Л.А.* О некоторых свойствах доверительных интервалов для неизвестных вероятностей // Автоматика и телемеханика. – 2007. – Вып. 12. – С. 70-84.
6. *Krishnamoorthy K., Peng J.* Some properties of the exact and score methods for binomial proportion and sample size calculation // Communications in Statistics – Simulation and Computation. – 2007. – Vol. 36. – P. 1171-1186.
7. *Вентцель Е.С.* Теория вероятностей: Учебник для вузов. – 7-е изд. стер. – М.: Высшая школа, 2001. – 575 с.
8. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. – М.: Физматлит, 2006. – 816 с.
9. *Agresti A.* Score and pseudo-score confidence intervals for categorical data analysis // American Statistical Association. Statistics in Biopharmaceutical Research. – 2011. – Vol. 3, No. 2. – P. 163-172.
10. *Brown L.D.,Cai T.T., DasGupta A.* Confidence intervals for a binomial proportion and asymptotic expansions // The Annals of Statistics. – 2002. – Vol. 30, No. 1. – P. 160-201.
11. *Reiczigel J.* Confidence intervals for the binomial parameter: some new considerations // Statistics in Medicine. – 2003. – Vol. 22. – P. 611-621.
12. *Boomsma A.* Confidence intervals for a binomial proportion // University of Groningen. Department of statistics and measurement theory. – 2005. – P. 1-9.
13. *Agresti A., Coull B.A.* Approximate is better than “exact” for interval estimation of binomial proportion // American Statistician. – 1998. – Vol. 52. – P. 119-125.
14. *Agresti A., Caffo B.* Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two // American Statistician. – 2000. – Vol. 54, No. 4. – P. 280-288.
15. *Robert C.P.* The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. – New York, Springer, 2007. – 602 p.
16. *Thulin M.* On split sample and randomized confidence intervals for binomial proportions // Statistics & Probability Letters. – 2014. – Vol. 92. – P. 65-71.
17. Справочник по надежности. В 3 т. Т. 1 = Reliability handbook / Под общ. ред. W.G. Ireson: Пер. с англ. / Под. ред. Б.Р. Левина. – М.: Мир, 1969. – 340 с.
18. *Оуэн Д.Н.* Сборник статистических таблиц: Пер. с англ. – М.: ВЦ АН СССР, 1966. – 568 с.
19. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука. Главная редакция физико-математической литературы 1983. – 416 с.
20. *Янко Я.* Математико-статистические таблицы: Пер. с чеш. – М.: Госстатиздат, 1961. – 244 с.
21. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и ее инженерные приложения. Учеб. пособие для втузов. – 2-е изд., стер. – М.: Высшая школа, 2000. – 480 с.

REFERENCES

1. *Thulin M.* The cost of using exact confidence intervals for a binomial proportion, *Electronic Journal of Statistics*, 2014, Vol. 8, pp. 817-840.

2. *Ivanov N.N., Strel'nikov V.P.* Prognozirovanie ostatochnoy dolgovechnosti payanykh soedineniy [Prediction of residual life of solder joints], *Matematichni mashini i sistemi* [Mathematical Machines and Systems], 2012, No. 3, pp. 162-165.
3. *Kuznetsov A.G. Aleksandrovskaya L.N.* Neparametricheskie metody "izmereniya" malykh riskov v zadachakh otsenki sootvetstviya trebovaniy k bezopasnosti avtomaticheskoy posadki samoletov normam letnoy godnosti [Nonparametric methods for measurement of small risks in the tasks of conformity assessment requirements for security of automatic landing aircraft airworthiness], *Trudy Moskovskogo instituta elektromekhaniki i avtomatiki (MIEA)* [Proceedings of the Moscow Institute of electromechanics and automation (MIEA)], 2011, Issue 3, pp. 2-11.
4. *Gusev L.A.* Ob interpretatsii nerazlichimosti v zadache interval'noy otsenki neizvestnoy veroyatnosti [About the interpretation of fuzzy in the problem of interval estimation of unknown probability], *Avtomatika i telemekhanika* [Avtomatika i Telemekhanika], 2010, Issue 8, pp. 38-48.
5. *Gusev L.A.* O nekotorykh svoystvakh doveritel'nykh intervalov dlya neizvestnykh veroyatnostey [On some properties of confidence intervals for unknown probabilities], *Avtomatika i telemekhanika* [Avtomatika i Telemekhanika], 2007, Issue 12, pp. 70-84.
6. *Krishnamoorthy K., Peng J.* Some properties of the exact and score methods for binomial proportion and sample size calculation, *Communications in Statistics – Simulation and Computation*, 2007, Vol. 36, pp. 1171-1186.
7. *Venttsel' E.S.* Teoriya veroyatnostey: Uchebnik dlya vuzov [Probability theory: the Textbook for high schools]. 7 th ed. Moscow: Vysshaya shkola, 2001, 575 p.
8. *Kobzar' A.I.* Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov [Applied mathematical statistics. For engineers and scientists]. Moscow: Fizmatlit, 2006, 816 p.
9. *Agresti A.* Score and pseudo-score confidence intervals for categorical data analysis, *American Statistical Association. Statistics in Biopharmaceutical Research*, 2011, Vol. 3, No. 2, pp. 163-172.
10. *Brown L.D.,Cai T.T., DasGupta A.* Confidence intervals for a binomial proportion and asymptotic expansions, *The Annals of Statistics*, 2002, Vol. 30, No. 1, pp. 160-201.
11. *Reiczigel J.* Confidence intervals for the binomial parameter: some new considerations, *Statistics in Medicine*, 2003, Vol. 22, pp. 611-621.
12. *Boomsma A.* Confidence intervals for a binomial proportion, *University of Groningen. Department of statistics and measurement theory*, 2005, pp. 1-9.
13. *Agresti A., Coull B.A.* Approximate is better than “exact” for interval estimation of binomial proportion, *American Statistician*, 1998, Vol. 52, pp. 119-125.
14. *Agresti A., Caffo B.* Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two, *American Statistician*, 2000, Vol. 54, No. 4, pp. 280-288.
15. *Robert C.P.* The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. New York, Springer, 2007, 602 p.
16. *Thulin M.* On split sample and randomized confidence intervals for binomial proportions, *Statistics & Probability Letters*, 2014, Vol. 92, pp. 65-71.
17. Spravochnik po nadezhnosti. V 3 vol. Vol. 1 = Reliability handbook, Under ed. W.G. Ireson: translation from English, Under ed. B.R. Levina. Moscow: Mir, 1969, 340 p.
18. *Ouen D.N.* Sbornik statisticheskikh tablits [The collection of statistical tables]: translation from English. Moscow: VTs AN SSSR, 1966, 568 p.
19. *Bol'shev L.N., Smirnov N.V.* Tablitsy matematicheskoy statistiki [Tables of mathematical statistics]. Moscow: Nauka. Glavnaya redaktsiya fiziko-matematicheskoy literatury 1983, 416 p.
20. *Yanko Ya.* Matematiko-statisticheskie tablitsy [Mathematical-statistical tables]: translation from Czech. Moscow: Gosstatizdat, 1961, 244 p.
21. *Venttsel' E.S., Ovcharov L.A.* Teoriya veroyatnostey i ee inzhenernye prilozheniya. Ucheb. posobie dlya vtuzov [Probability theory and its engineering applications. Textbook for technical colleges]. 2 nd ed. Moscow: Vysshaya shkola, 2000, 480 p.

Статью рекомендовал к опубликованию д.т.н., профессор В.Т. Еременко.

Трегубов Роман Борисович – Академия Федеральной службы охраны Российской Федерации; e-mail: reba@list.ru; 302034, г. Орел, ул. Приборостроительная, 35; к.т.н.; сотрудник.

Стремоухов Михаил Владимирович – e-mail: smv_57@bk.ru; сотрудник.

Tregubov Roman Borisovich – Academy of the Federal Guard Service of the Russian Federation; e-mail: smv_57@bk.ru; 35, Priborostroitelnaya street, Orel, 302034, Russia; cand. of eng. sc.; member.

Stremouhov Mihail Vladimirovich – e-mail: smv_57@bk.ru; member.

УДК 004.89

С.С. Алхасов, А.Н. Целых

ОСНОВНЫЕ ПОДХОДЫ К ПОСТРОЕНИЮ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ МОДЕЛИРОВАНИЯ ОТТОКА КЛИЕНТОВ УСЛУГ СВЯЗИ

Кратко рассмотрены важнейшие функциональные модули информационной системы прогнозирования оттока клиентов телекоммуникационного предприятия. Определены основные подходы к предварительной обработке архивных данных и моделированию оттока клиентов. Заданы базовые требования для практической реализации прогностической системы. Отдельное внимание обращено на преодолении сильной коррелированности между переменными в массиве входных данных. Предложено использовать метод главных компонент, предполагающий декомпозицию входного массива на вектора счетов и нагрузок. Рассмотренный алгоритм NIPALS имеет итеративный характер. Вектор счетов, вычисленный на некоторой итерации, является соответствующей главной компонентой. Определение главных компонент дальних порядков, как правило, лишено смысла, поскольку их значения обусловлены наличием некоторой погрешности во входных данных. Указаны основные критерии для определения эффективного числа главных компонент: объясненная дисперсия и нормированное собственное значение вектора счетов. В качестве примера сформирован экспериментальный массив входных данных размера 9×2000, в который специально подобраны разнородные переменные (технология подключения, тип населенного пункта, скорость подключения, стоимость услуги, трафик в 1-ом месяце, трафик во 2-м месяце, трафик в 3-м месяце и др.). Отмечено, что данная методика позволяет преодолеть разнородность входной информации и сильную коррелированность переменных, а также снижает размерность входного массива. Графически показано, как число используемых главных компонент влияет на объясненную дисперсию и величину нормированного собственного значения. Все эти аспекты свидетельствуют, что данный подход перспективен для применения в прогностической системе, содержащей кластеризирующие и нейросетевые модули.

Прогнозирование; отток клиентов; Интернет; метод главных компонент; снижение размерности; кластеризация.

S.S. Alkhasov, A.N. Tselykh

THE MAIN APPROACHES TO THE CREATION OF THE INFORMATION SYSTEM FOR MODELING OF TELECOMMUNICATION CLIENTS OUTFLOW

In the present article most important functional modules of the information system for the prediction of clients outflow from a telecom company are briefly considered. The basic approaches to preprocessing of archived data and clients outflow modeling are defined. The main requirements for the practical implementation of the prognostic system are introduced. Special attention is focused on overcoming of the strong correlation between the variables in the array of input data. It's offered to use principal components method, implying the decomposition of the input array to the score and the loading vectors. Considered algorithm NIPALS has iterative character. The score vector calculated on some iteration is the corresponding principal component. The determination of the principal components of the long-range orders doesn't have the sense typically because their values caused availability of some error in the input data. The basic criteria for definition of efficient number of principal components are