ВЛИЯНИЕ ИСКАЖЕНИЙ В НАБЛЮДЕНИЯХ НА ХАРАКТЕРИСТИКИ ПОСЛЕДОВАТЕЛЬНЫХ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ

А.Ю. Харин, С.Ю. Чернов

Белорусский государственный университет, факультет прикладной математики и информатики
пр. Независимости, 4, г. Минск, Беларусь
телефон: + (375 17) 2095129; факс: +(375 17) 2095104;
e-mail: KharinAY@bsu.by

В работе рассмотрен последовательный критерий отношения вероятностей (ПКОВ) при абсолютно непрерывных распределениях вероятностей наблюдений. Исследуется влияние искажений в гипотетической модели на характеристики ПКОВ. Построен робастифицированный последовательный критерий.

Ключевые слова – искажение, последовательный критерий отношения вероятностей, робастность.

1 ПОСЛЕДОВАТЕЛЬНЫЙ КРИТЕРИЙ ОТНОШЕНИЯ ВЕРОЯТНОСТЕЙ

Пусть наблюдается последовательность независимых одинаково распределенных случайных величин $x_1, x_2, \ldots$, $x_t \in \mathbf{D} \subset \mathbf{R}$. Пусть эти случайные величины имеют плотность распределения вероятностей $f(x,\theta)$ с параметром $\theta \in \Theta = \{\theta_0, \theta_1\}$, истинное значение которого неизвестно. Пусть плотности распределения вероятностей $f(x,\theta)$ соответствует функция распределения $F(x,\theta)$. Относительно параметра θ имеются две простые гипотезы

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1. \tag{1}$$

Обозначим статистику:

$$\Lambda_n = \Lambda_n(x_1, \ldots x_n) = \sum_{t=1}^{n} \lambda_t, \tag{2}$$

$$\lambda_t = \lambda(x_t) = \ln \frac{f(x_t, \theta_1)}{f(x_t, \theta_0)} \quad - \tag{3}$$

логарифм статистики отношения правдоподобия, вычисленной по наблюдению x_t $t = 1, 2, \ldots$.

Для проверки гипотез (1) по $n = 1,2,\ldots$ наблюдениям выносится решение, основанное на последовательном критерии отношения вероятностей (ПКОВ) [1]:

$$N = \min\{n \in \mathbf{N} : \Lambda_n \notin (C_-, C_+)\}, \tag{4}$$

$$d = 1_{[C_+, +\infty)}(\Lambda_N), \tag{5}$$

где N – момент остановки, после которого принимается решение d в соответствии с (5). В (4) C_-, C_+ – заданные пороги (параметры критерия) [1]:

$$C_- = \ln \frac{\beta_0}{1-\alpha_0}, \quad C_+ = \ln \frac{1-\beta_0}{\alpha_0}, \tag{6}$$

где α_0, β_0 – допустимые значения вероятностей ошибок I и II рода.

Известно [2], что α_0, β_0 лишь приближенно равняются результирующим вероятностям ошибок первого и второго рода.

Для дальнейшего изложения сделаем следующие предположения:

П1) функция $f(x,\theta)$ имеет на множестве D конечные производные 1-го и 2-го порядка по переменной x, а также $f(x,\theta) \neq 0$, $\theta \in \Theta$;

П2) функция $\lambda(x)$, определенная (3), строго монотонна по переменной x на множестве D, а также имеет отличную от нуля производную 1-го порядка.

Данным предположениям удовлетворяют, например, представители экспоненциального семейства распределений, у которых плотность распределения вероятностей имеет вид $f(x,\theta) = a(x)b(\theta)exp\{c(x)d(\theta)\}$, где

1) $a(x), c(x)$ дважды дифференцируемы, а также $a(x) \neq 0, x \in \mathbf{D}$, и $b(0) \neq 0, \theta \in \Theta$;

2) $\operatorname{sign} c'(x) = \operatorname{invar}_{x \in \mathbf{D}}$, $d(\theta_0) \neq d(\theta_1)$.

Без ограничения общности будем считать, что истинной гипотезой является H_0 (случай H_1 рассматривается аналогично).

2 ПКОВ, ПРИМЕНЯЕМЫЙ ПРИ НАЛИЧИИ ИСКАЖЕНИЙ В НАБЛЮДЕНИЯХ

В рамках описываемой модели рассмотрим ситуацию, когда наблюдения $x_1, x_2, \ldots$ подвержены "выбросам" ("засорениям") Тьюки-Хьюбера [3], то есть случайные величины x_t, $t \geq 1$, имеют плотность распределения вероятностей

$$f_1(x,\theta) = (1-\varepsilon) f(x,\theta) + \varepsilon \tilde{f}(x,\theta), \tag{7}$$

где $\tilde{f}(x,\theta) \not\equiv f(x,\theta)$ – искажающая плотность распределения вероятностей, $\varepsilon \in [0, 0{,}5)$.

Пусть $\pi(\varepsilon)$, $P^{\pm}(\varepsilon)$ и $R^{\pm}(\varepsilon)$ – соответственно вектор вероятностей начальных состояний и матрицы вероятностей перехода цепей Маркова L_n^- и L_n^+, построение которых приводится в [6], в случае, когда наблюдения x_i имеют плотность распределения вероятностей (7). Пусть $\widetilde{\pi}$, $\widetilde{P}^{\pm}$ и $\widetilde{R}^{\pm}$ – соответственно вектор вероятностей начальных состояний и матрицы вероятностей перехода цепей Маркова L_n^- и L_n^+ в случае, когда наблюдения x_i имеют плотность распределения вероятностей $\widetilde{f}(x,\theta)$. Пусть π_{out} и $\widetilde{\pi}_{out}$ – вероятности попадания цепи Маркова L_n^- на первом шаге в поглощающее состояние, когда наблюдения x_i имеют плотность распределения вероятностей (7) и $\widetilde{f}(x,\theta)$ соответственно; $S^{\pm} = I - P^{\pm}$.

Теорема 1. Если для плотности распределения вероятностей (7) выполнены предположения П1 и П2, то вектор вероятностей начальных состояний и матрицы вероятностей перехода поглощающих цепей Маркова L_n^- и L_n^+ удовлетворяют соотношениям:

$$\begin{aligned} \pi(\varepsilon) &= (1-\varepsilon)\pi + \varepsilon\widetilde{\pi}, \\ P^{\pm}(\varepsilon) &= (1-\varepsilon)P^{\pm} + \varepsilon\widetilde{P}^{\pm}, \\ R^{\pm}(\varepsilon) &= (1-\varepsilon)R^{\pm} + \varepsilon\widetilde{R}^{\pm}. \end{aligned} \qquad (8)$$

Доказательство основано на результате Теоремы 1 в [6] и соотношении (7).

Пусть $\alpha(\varepsilon)$ – вероятность ошибки первого рода для ПКОВ, когда наблюдения имеют плотность распределения вероятностей $f_1(x,\theta)$; $\alpha_m^-(\varepsilon)$ и $\alpha_m^+(\varepsilon)$ вероятности ошибок первого рода последовательных критериев, основанных на цепях Маркова Λ_n^- и Λ_n^+, построение которых приводится в [6], соответственно, для модели наблюдений (7).

Теорема 2. Вероятности ошибок первого рода $\alpha_m(\varepsilon)$ и $\alpha_m^{'}(\varepsilon)$ допускают асимптотические разложения при $\varepsilon \to 0$:

$$\alpha_m^{\pm}(\varepsilon) = \alpha_m^{\pm} + a_1^{\pm}\varepsilon + O(\varepsilon^2),$$

где

$$\begin{aligned} a_1^{\pm} &= \widetilde{\pi}_{out} - \pi_{out} + (\widetilde{\pi} - \pi)(S^{\pm})^{-1}R^{\pm} + \\ &+ \pi(S^{\pm})^{-1}(\widetilde{P}^{\pm} - P^{\pm})(S^{\pm})^{-1}R^{\pm} + \\ &+ \pi(S^{\pm})^{-1}(\widetilde{R}^{\pm} - R^{\pm}). \end{aligned}$$

Доказательство основано на определениях величин $\alpha_m^-(\varepsilon)$ и $\alpha_m^{'}(\varepsilon)$ и соотношениях (8).

Следствие 1. Вероятности ошибок первого рода $\alpha_m(\varepsilon)$ и $\alpha_m^{'}(\varepsilon)$ допускают асимптотические разложения при $\varepsilon \to 0$:

$$\alpha_m^{\pm}(\varepsilon) = \alpha_m^{\pm} + \sum_{k=1}^{\infty} a_k^{\pm}\varepsilon^k,$$

где

$$a_k^{\pm} = \frac{d^k}{d\varepsilon^k}\alpha_m^{\pm}(\varepsilon)\bigg|_{\varepsilon=0}.$$

3 РОБАСТИФИЦИРОВАННЫЙ ПОСЛЕДОВАТЕЛЬНЫЙ КРИТЕРИЙ

Пусть случайные величины $x_1, x_2, \ldots$ имеют плотность распределения вероятностей $f_1(x,\theta)$, заданную (7), с параметром $\theta \in \Theta = \{\theta_0, \theta_1\}$, истинное значение которого неизвестно. Пусть плотности распределения вероятностей $f_1(x,\theta)$ соответствует функция распределения $F_1(x,\theta)$.

Для уменьшения влияния искажений (7) на точность принимаемых решений будем рассматривать семейство последовательных критериев, основанных на усеченных приращениях логарифмов статистик отношения правдоподобия.

Построим случайную последовательность

$$\Lambda_n^g = \sum_{t=1}^{n} \lambda_t^g,$$

$$\lambda_t^g = \lambda_t 1_{[g_-, g_+]}(\lambda_t) + g_- 1_{(-\infty, g_-)}(\lambda_t) + g_- 1_{(g_+, +\infty)}(\lambda_t),$$

где g_- и g_+ – параметры усечения статистик λ_t, $t \in \mathbf{N}$.

Аналогично тому, как в [6] для случайной последовательности Λ_n строятся граничные цепи Маркова Λ_n^- и Λ_n^+, построим цепи Маркова Λ_n^g и Λ_n^{g+} для последовательности Λ_n^g.

Разобьем область (C_-, C_-) между порогами теста на m подмножеств-полосок "толщиной" $h = \frac{C_+ - C_-}{m}$, где $m \in \mathbf{N}$ – параметр разбиения (аппроксимации).

Построим случайные последовательности

$$\Lambda_n^{g-} = \sum_{t=1}^{n} \lambda_t^{g-}, \; t \in \mathbf{N};$$

$$\lambda_1^g = C_- + \left[\frac{\lambda_1^g - C_-}{h}\right]h, \quad \lambda_t^{g-} = \left[\frac{\lambda_t^g}{h}\right]h, \quad t \geq 2;$$

$$\lambda_1^{g+} = C_- + \left[\frac{\lambda_1^g - C_-}{h}\right]h - h, \quad \lambda_t^{g-} = \left[\frac{\lambda_t^g}{h}\right]h + h, \quad t \geq 2.$$

Рассмотрим поглощающую цепь Маркова L_n^{g-}, имею-

щую состояния $0,1,\ldots,m+1$, из которых 0 и $m+1$ являются поглощающими:

$$L_n^{g^-} = \begin{cases} 0, & \Lambda_n^{g^-} \in (-\infty, C_- - h], \\ i, & \Lambda_n^{g^-} = C_- + (i-1)h, \quad i = \overline{1,m}, \\ m+1, & \Lambda_n^{g^-} \in [C_+, +\infty). \end{cases}$$

Аналогично строится цепь Маркова $L_n^{g^+}$:

$$L_n^{g^+} = \begin{cases} 0, & \Lambda_n^{g^+} \in (-\infty, C_-], \\ i, & \Lambda_n^{g^+} = C_- + ih, \quad i = \overline{1,m}, \\ m+1, & \Lambda_n^{g^+} \in [C_+ + h, +\infty). \end{cases}$$

Обозначим

$$G_-^{(0)} = \left[\frac{g_- - C_-}{h}\right] + 1, \qquad G_+^{(0)} = \left[\frac{g_+ - C_-}{h}\right] + 1.$$

Теорема 3. Если для рассмотренной модели (7) выполнены предположения П1 и П2, то вектор вероятностей начальных состояний поглощающей цепи Маркова $L_n^{g^-}$ вычисляется поэлементно следующим образом:

$$\pi_i^{g^-} = \pi_i, \quad i = \overline{G_-^{(0)}+1, G_+^{(0)}-1},$$
$$\pi_{G_-^{(0)}}^{g^-} = \sum_{k=0}^{G_-^{(0)}} \pi_k, \quad \pi_{G_+^{(0)}}^{g^-} = \sum_{k=G_+^{(0)}}^{m+1} \pi_k,$$
$$\pi_i^{g^-} = 0, \ i > G_+^{(0)}, \ i < G_-^{(0)},$$

где π – вектор вероятностей начальных состояний цепи Маркова L_n, построенной в [6].

Доказательство проводится аналогично доказательству Теоремы 1 в [6].

Обозначим

$$G_- = \left[\frac{g_-}{h}\right] + 1, \qquad G_+ = \left[\frac{g_+}{h}\right] + 1.$$

Теорема 4. Если для рассмотренной модели (7) выполнены предположения П1 и П2, то матрица вероятностей перехода поглощающей цепи Маркова $L_n^{g^-}$ вычисляется поэлементно следующим образом:

$$p_{i,i+j}^{g^-} = p_{i,j}^-, \quad j = \overline{G_- + 1, G_+ - 1},$$
$$p_{i,i+G_-}^{g^-} = \sum_{k=0}^{i+G_-} p_{i,k}^-,$$
$$p_{i,i+}^{g^-} = \sum_{k=i+G_+}^{m+1} p_{i,k}^-,$$
$$p_{i,i+j}^{g^-} = 0, \ j > G_+, \ j < G_-,$$
$$i = \overline{1,m}, \ j = \overline{0,m+1}.$$

где P^- – матрица вероятностей перехода для цепи Маркова L_n^-, построенной в [6].

Доказательство проводится аналогично доказательству Теоремы 1 в [6].

Введем матрицы $P^{g^-} \in \mathrm{R}^{m \times m}$ и $R^{g^-} \in \mathrm{R}^{m \times 2}$:

$$(P^{g^-})_{i,j} = p_{ij}^{g^-},$$
$$(R^{g^-})_{i,1} = p_{i,0}^{g^-},$$
$$(R^{g^-})_{i,2} = p_{i,m+1}^{g^-}, \quad i,j = \overline{1,m}.$$

Теорема 5. Цепи Маркова $\Lambda_n^{g^-}$, Λ_n^g и $\Lambda_n^{g^+}$ удовлетворяют следующему соотношению:

$$\forall n \in \mathbf{N}: \quad \Lambda_n^{g^-} \le \Lambda_n^g \le \Lambda_n^{g^+}.$$

Доказательство проводится аналогично доказательству Теоремы 3 в [6].

Пусть $\alpha_m^{g^-}$, α^g и $\alpha_m^{g^+}$ – вероятности ошибок первого рода последовательных критериев, основанных на цепях Маркова $\Lambda_n^{g^-}$, Λ_n^g и $\Lambda_n^{g^+}$ соответственно. Из построения цепей Маркова $\Lambda_n^{g^-}$ и $\Lambda_n^{g^+}$ следует, что

$$\alpha_m^{g^-} \le \alpha^g \le \alpha_m^{g^+}.$$

Поэтому в качестве оценки приближенного значения α^g принимается величина

$$\hat{\phi}_m^g = \frac{1}{2}\left(\phi_m^{g^+} + \phi_m^{g^-}\right). \tag{9}$$

Теорема 6. Если для рассмотренной модели искаженных наблюдений выполнены предположения П1 и П2, то величины $\alpha_m^{g^-}$ и $\alpha_m^{g^+}$, удовлетворяют соотношению

$$\alpha_m^{g^+} - \alpha_m^{g^-} = O(h).$$

Доказательство проводится аналогично доказательству Теоремы 4 в [6].

Следствие 2. Оценка (9) стремится к значению вероятности ошибки первого рода последовательного критерия, основанного на цепи Маркова Λ_n^g, со скоростью $O(h)$, причем отклонение от этого значения не превосходит половины длины отрезка $[\alpha_m^{g^-}, \alpha_m^{g^+}]$:

$$\left| \phi^g - \hat{\phi}_m^g \right| \le \frac{1}{2}\left(\phi_m^{g^+} - \phi_m^{g^-}\right).$$

Представителей семейства последовательных критериев, основанных на цепях Маркова Λ_n^g, будем называть робастифицированными критериями.

Проблему выбора параметров усечения g_- и g_+ можно решать, исходя из минимизации вероятностей ошибочных решений. Примеры представлены в разделе, посвященном вычислительным экспериментам.

4 ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Эксперименты проводились для случая $\alpha_0 = \beta_0 = 0.1$, $\theta_0 = 0$, $\theta_1 = 0.5$, $m = 2000$. Истинной гипотезой во всех экспериментах была гипотеза H_0.

Обозначим через $\hat{\phi}_{MC}$ и $\hat{\phi}^{g}_{MC}$, $\hat{n}_{MC}$ и $\hat{n}^{g}_{MC}$ – оценки вероятностей ошибок первого рода и математических ожиданий длительностей ПКОВ и последовательного критерия, основанного на цепи Маркова Λ^{g}_{n} соответственно, полученные имитационным моделированием (методом Монте-Карло) с числом "прогонов", равным 1000000. Вычислительные эксперименты проводились для случая

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta)^2}{2}}, \tilde{f}(x,\theta) = \frac{100}{\sqrt{2\pi}} e^{-5000(x-10)^2}.$$

Параметры усечения заданы следующим образом:

$$g_- = \frac{1}{2}\left(F^{-1}\left(\frac{g_\varepsilon}{2} \mid \theta_0\right) + F^{-1}\left(\frac{g_\varepsilon}{2} \mid \theta_1\right)\right),$$
$$g_+ = \frac{1}{2}\left(F^{-1}\left(1-\frac{g_\varepsilon}{2} \mid \theta_0\right) + F^{-1}\left(1-\frac{g_\varepsilon}{2} \mid \theta_1\right)\right), \quad (10)$$

параметр g_ε определяет используемые в (10) квантили.

Результаты вычислительных экспериментов представлены в таблицах 1 и 2, в которых приведена зависимость вероятностных характеристик ПКОВ и робастифицированного критерия от уровня искажения ε.

Таблица 1

Зависимость вероятностных характеристик ПКОВ от уровня искажения ε

ε	α_m^-	$\hat{\phi}_{MC}$	α_m^+	$\hat{n}_{MC}$
0.0	0.07384	0.07667	0.07966	17.194
0.005	0.14730	0.15120	0.15426	16.067
0.01	0.21379	0.21749	0.22189	15.179
0.02	0.32513	0.32868	0.33167	13.620
0.03	0.41458	0.41854	0.42206	12.328
0.05	0.55098	0.55419	0.55807	10.340
0.10	0.74854	0.75104	0.75311	7.298

Таблица 2

Зависимость вероятностных характеристик робастифицированного критерия от уровня искажения ε и параметров усечения

ε	g_ε	g_+	α_m^{g-}	$\hat{\phi}^{g}_{MC}$	α_m^{g+}	$\hat{n}_{MC}$
0.0	0.01	1.2879	0.07284	0.07537	0.07869	17.28
0.01	0.01	1.2879	0.10953	0.11301	0.11697	17.77
	0.10	0.8224	0.08334	0.08783	0.09069	19.85
	0.20	0.6408	0.06651	0.07195	0.07401	22.68
	0.40	0.4208	0.03649	0.04115	0.04381	31.43
	0.60	0.2622	0.01099	0.01381	0.01618	48.75
0.05	0.10	0.8224	0.19903	0.20624	0.21194	22.19
	0.20	0.6408	0.15598	0.16486	0.16984	25.99
	0.40	0.4208	0.09326	0.10388	0.10906	38.07
	0.60	0.2622	0.03680	0.04605	0.05212	62.39
	0.70	0.1927	0.01672	0.02062	0.02262	86.21
0.10	0.10	0.8224	0.40396	0.41115	0.42089	23.01
	0.20	0.6408	0.33818	0.35069	0.35896	28.76
	0.40	0.4208	0.24296	0.26272	0.27365	45.80
	0.50	0.3372	0.20508	0.21884	0.22392	60.59
	0.60	0.2622	0.15333	0.16977	0.17651	84.51
	0.70	0.1927	0.09286	0.11188	0.12023	127.6
	0.75	0.1593	0.06162	0.07961	0.08962	163.2
	0.80	0.1267	0.03188	0.04738	0.06182	216.8

ЛИТЕРАТУРА

[1] Вальд А. Последовательный анализ / А. Вальд. – М.: Наука, 1960.

[2] Харин А.Ю. Об одном подходе к анализу последовательного критерия отношения правдоподобия для различения простых гипотез / А.Ю. Харин // Вестник БГУ, Сер. 1, 2002, номер 1. – С. 92-96.

[3] Хьюбер П. Робастная статистика / П. Хьюбер. – М.: Мир, 1984 – 294 с.

[4] Kharin A. Robust sequential testing of hypotheses on discrete probability distributions / A. Kharin, D. Kishylau // Austrian Journal of Statistics, vol. 34, 2005, N 2. – P. 153-162.

[5] Ghosh B.K. Handbook of Sequential Analysis / B.K. Ghosh, P.K. Sen. – New York: Marcel Dekker, 1991. – 638 p.

[6] Kharin A. Error probabilities evaluation for sequential testing of simple hypotheses on data from continuous distribution / A. Kharin, S. Chernov // Pattern Recognition and Information Processing. Proceedings of the Conference. Minsk, 2009. P. 63-66.