

$$) \quad \frac{\sqrt{}}{\sqrt{}}$$

АНАЛИЗ ВЕРОЯТНОСТЕЙ ОШИБОЧНЫХ РЕШЕНИЙ ПОСЛЕДОВАТЕЛЬНОГО КРИТЕРИЯ ОТНОШЕНИЯ ВЕРОЯТНОСТЕЙ ПРИ НАЛИЧИИ ИСКАЖЕНИЙ В НАБЛЮДЕНИЯХ

С. Ю. Чернов, А. Ю. Харин

Белорусский государственный университет

Минск, Беларусь

E-mail: chernovsy@tut.by

Проводится анализ зависимости вероятностей ошибок последовательного критерия отношения вероятностей проверки двух простых гипотез о параметре непрерывного распределения при наличии в наблюдениях выбросов Тьюки –

Хьюбера. Предложен способ повышения устойчивости вероятностных характеристик последовательных критериев при наличии искажений в теоретической модели.

Ключевые слова: ПКОВ, выбросы Тьюки – Хьюбера, робастность.

ВВЕДЕНИЕ

Последовательный подход проверки гипотез широко используется при компьютерной обработке данных в медицине, статистическом контроле качества, биологии [3] и финансах [4]. Основным преимуществом последовательных процедур является в среднем меньшее число наблюдений по сравнению с фиксированным числом наблюдений, необходимых для принятия решения для эквивалентных классических процедур [1]. В работе рассматривается последовательный критерий отношения вероятностей (ПКОВ), предложенный А. Вальдом [1]. Одним из недостатков последовательных процедур является существенное влияние искажений на вероятностные характеристики критериев, что может привести к неадекватным решениям при компьютерном анализе данных на практике. В [2] рассмотрен случай, когда наблюдения имеют дискретное распределение. В данной работе исследовано воздействие выбросов Тьюки – Хьюбера на значения вероятностей ошибочных решений ПКОВ, когда наблюдения абсолютно непрерывны, и предложен способ повышения устойчивости указанных характеристик при наличии искажений. Теоретические результаты иллюстрируются выводами вычислительных экспериментов.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Пусть на измеримом пространстве $(\Omega, \mathfrak{S})$ наблюдается последовательность $x_1, x_2, \dots$ ($x_i \in D \in \mathbb{R}$) независимых, одинаково распределенных величин с плотностью распределения вероятностей $p(x, \theta)$, зависящей от параметра $\theta \in \Theta = \{\theta_0, \theta_1\}$. Истинное значение параметра θ неизвестно. Пусть функция распределения вероятностей $F(x, \theta)$ соответствует плотности $p(x, \theta)$.

Относительно значения параметра θ имеются две простые гипотезы:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1. \quad (1)$$

Обозначим статистику накопленного отношения правдоподобия:

$$\Lambda_n = \Lambda_n(x_1, \dots, x_n) = \sum_{k=1}^n \lambda_k, \quad (2)$$

$$\lambda_k = \lambda(x_k) = \log \frac{p(x_k, \theta_1)}{p(x_k, \theta_0)}, \quad x_0 = 0. \quad (3)$$

Отметим, что так как λ_k являются независимыми случайными величинами, то Λ_n – марковская случайная последовательность.

Для последовательной проверки гипотез (1) по n ($n \in N$) наблюдениям используется последовательный критерий отношения вероятностей (ПКОВ):

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

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

где N – момент остановки, после которого принимается решение d в соответствии с (5). В (4) пороги C_- , C_+ – параметры критерия, которые выбираются в соответствии с $0 < C_- = \log(\beta_0/(1 - \alpha_0))$, $C_+ = \log((1 - \beta_0)/\alpha_0)$, где α_0, β_0 – заданные значения вероятностей ошибок первого и второго рода соответственно. Известно [1], что α_0 и β_0 лишь приближенно равняются фактическим вероятностям первого и второго рода α и β ПКОВ.

Не нарушая общности, предположим, что верна гипотеза H_0 , поэтому в дальнейшем будем рассматривать вероятность ошибки первого рода α критерия (4), (5). Обозначим $p(x) = p(x, \theta_0)$, $F(x) = F(x, \theta_0)$, $F_\lambda(x) = P\{\lambda_1 \leq x\}$, $p_\lambda(x) = dF_\lambda(x)/dx$.

ПОСТРОЕНИЕ ГРАНИЧНЫХ МАРКОВСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Разобьем область (C_-, C_+) между порогами теста на m полосок «толщиной» h ($m \in N$ – параметр аппроксимации): $h = (C_+ - C_-)/m$.

Построим случайную последовательность

$$\Lambda_n^- = \sum_{k=1}^n \lambda_k^-, \quad \lambda_1^- = C_- + \left[\frac{\lambda_1 - C_-}{h} \right] h, \quad \lambda_k^- = \left[\frac{\lambda_k}{h} \right] h, \quad k \geq 2,$$

где $[\cdot]$ обозначает целую часть действительного числа.

Так как случайные величины λ_k^- являются независимыми, то Λ_n^- – однородная марковская последовательность.

Рассмотрим однородную цепь Маркова L_n^- с конечным пространством состояний $\{0, 1, \dots, m, m+1\}$, $\{L_n^- = i\} = \{\Lambda_n^- = C_- + (i-1)h\}$, $\{L_n^- = 0\} = \{\Lambda_n^- \leq C_- - h\}$, $\{L_n^- = m+1\} = \{\Lambda_n^- \geq C_+\}$, из которых состояния 0 и $m+1$ являются поглощающими.

Построим матрицы $\pi^- \in R^{1 \times m}$, $P^- \in R^{m \times m}$, $R^- \in R^{m \times 2}$, которые содержат начальные $P\{L_1^- = i\}$ и переходные вероятности $P\{L_{n+1}^- = j | L_n^- = i\}$ цепи Маркова L_n^- :

$$(\pi^-)_i = P\{L_1^- = i\}, \quad (P^-)_{ij} = P\{L_n^- = j | L_{n-1}^- = i\},$$

$$(R^-)_{i,k} = P\{L_n^- = k | L_{n-1}^- = i\}, \quad k \in \{0, m+1\}.$$

Теорема 1. Вектор начальных вероятностей и матрицы вероятностей переходов цепи Маркова L_n^- вычисляются поэлементно следующим образом:

$$\pi_i^- = F_\lambda(C_- + ih) - F_\lambda(C_- + (i-1)h), \quad \pi_0^- = F_\lambda(C_-), \quad \pi_{m+1}^- = 1 - F_\lambda(C_+);$$

$$P_{i,j}^- = F_\lambda((j-k+1)h) - F_\lambda((j-k)h),$$

$$R_{i,0}^- = F_\lambda((1-i)h), \quad R_{i,m+1}^- = 1 - F_\lambda((m-i+1)h), \quad i, j = \overline{1, m}.$$

Доказательство приведено в [5].

Марковская последовательность Λ_n^+ строится аналогично случайной последовательности Λ_n^- : $\Lambda_n^+ = \sum_{k=1}^n \lambda_k^+$, $\lambda_k^+ = \lambda_k^- + h$. Однородная цепь Маркова L_n^+ строится аналогично цепи Маркова L_n^- : $\{L_n^+ = i\} = \{\Lambda_n^+ = C_- + ih\}$, $\{L_n^+ = 0\} = \{\Lambda_n^+ \leq C_-\}$,

$\{L_n^+ = m+1\} = \{\Lambda_n^+ \geq C_+ + h\}$, состояния 0 и $m+1$ – поглощающие. Матрицы π^+ , P^+ , R^+ определяются аналогично матрицам π^- , P^- , R^- .

Теорема 2. Вектор начальных вероятностей и матрицы вероятностей переходов цепи Маркова L_n^+ вычисляются поэлементно следующим образом:

$$\begin{aligned}\pi_i^+ &= F_\lambda(C_- + ih) - F_\lambda(C_- + (i-1)h), \quad \pi_0^+ = F_\lambda(C_-), \quad \pi_{m+1}^+ = 1 - F_\lambda(C_+); \\ P_{i,j}^+ &= F_\lambda((j-k)h) - F_\lambda((j-k-1)h), \\ R_{i,0}^+ &= F_\lambda(ih), \quad R_{i,m+1}^+ = 1 - F_\lambda((m-i)h), \quad i, j = \overline{1, m}.\end{aligned}$$

Теорема 3. Марковская последовательность Λ_n ограничена сверху и снизу марковскими последовательностями Λ_n^- и Λ_n^+ : $\Lambda_n^- \leq \Lambda_n \leq \Lambda_n^+$.

Пусть вероятность ошибки 1-го рода критерия (4), (5) равна α ; α_m^- и α_m^+ – вероятности поглощения в состоянии $m+1$ цепей Маркова L_n^- и L_n^+ соответственно. По определению $\alpha_m^\pm = \pi_{m+1}^\pm + \pi(S^\pm)^{-1}R^\pm$, где $S^\pm = I - P^\pm$. Из способа построения цепей Маркова L_n^- и L_n^+ следует $\alpha_m^- \leq \alpha \leq \alpha_m^+$. В качестве приближения неизвестного фактического значения α рассмотрим величину

$$\hat{\alpha}_m = (\alpha_m^+ + \alpha_m^-)/2. \quad (6)$$

Теорема 4. Для рассмотренной модели наблюдений следующее равенство выполняется в асимптотике $h \rightarrow 0$: $\alpha_m^+ - \alpha_m^- = O(h)$.

Следствие 1. Приближение (6) сходится к значению вероятности ошибки 1-го рода α со скоростью $O(h)$, причем абсолютное отклонение от этого значения не превосходит половины длины отрезка $[\alpha_m^-, \alpha_m^+]$: $|\alpha - \hat{\alpha}_m| \leq (\alpha_m^+ - \alpha_m^-)/2$.

ПКОВ ПРИ НАЛИЧИИ ВЫБРОСОВ В НАБЛЮДЕНИЯХ

Пусть в рамках описываемой модели наблюдения $x_1, x_2, \dots$ ($x_i \in D \in R$) подвержены «выбросам» Тьюки – Хьюбера, т. е. плотность распределения вероятностей случайных величин x_i имеет вид $h(x, \theta) = (1 - \varepsilon)f(x, \theta) + \varepsilon v(x, \theta)$, где $v(x, \theta)$, $v(x, \theta) \neq f(x, \theta)$, – неизвестная искажающая плотность распределения вероятностей, а ε , $0 < \varepsilon < \varepsilon_0$, – уровень искажения, максимальное значение ε_0 ($\varepsilon_0 < 0.5$), которого известно заранее. Заметим, что статистики λ_k имеют плотность распределения вероятностей

$$h_\lambda(x, \theta) = (1 - \varepsilon)f_\lambda(x, \theta) + \varepsilon u(x, \theta), \quad (7)$$

где $u(x, \theta) = v_\lambda(x, \theta)$.

Пусть $\alpha(\varepsilon)$ – вероятность ошибки 1-го рода последовательного критерия отношения вероятностей для искаженной модели наблюдений (7). Пусть $\alpha_m^-(\varepsilon)$ и $\alpha_m^+(\varepsilon)$ вероятности поглощения цепи Маркова L_n^- и L_n^+ для модели наблюдений (7).

Теорема 5. Вероятности поглощения $\alpha_m^-(\varepsilon)$ и $\alpha_m^+(\varepsilon)$ допускают асимптотическое разложение при $\varepsilon \rightarrow 0$:

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

$$a_1^\pm = (\pi_u - \pi)(S^\pm)^{-1} R^\pm + \pi(S^\pm)^{-1} (P_u^\pm - P^\pm)(S^\pm)^{-1} R^\pm + \pi(S^\pm)^{-1} (R_u^\pm - R^\pm).$$

Таким образом, разности между значениями вероятностей ошибок 1-го рода $\alpha_m^-(\epsilon)$, $\alpha_m^+(\epsilon)$ для искаженной модели (11) и α_m^- , α_m^+ для теоретической модели имеют линейный порядок по уровню искажения ϵ , а линейные коэффициенты a_1^- и a_1^+ могут быть получены в явном виде.

РОБАСТИФИЦИРОВАННЫЙ ПОСЛЕДОВАТЕЛЬНЫЙ КРИТЕРИЙ

Рассмотрим «усеченную» плотность распределения вероятностей:

$$h_\lambda^g(x) = 1_{[g_-, g_+]}(x) h_\lambda(x) + \epsilon_- \delta(x - g_-) + \epsilon_+ \delta(x - g_+),$$

где g_- и g_+ – заданные параметры усечения статистики λ_k , имеющей плотность распределения вероятностей $h_\lambda(\cdot)$ и функцию распределения вероятностей $H_\lambda(\cdot)$, $\epsilon_- = H_\lambda(g_-)$, $\epsilon_+ = 1 - H_\lambda(g_+)$. Пусть $\bar{f}_\lambda^g(x) = (1 - \epsilon) f_\lambda^g(x) + \epsilon \delta(x - g_+)$.

Пусть $\alpha_m^+(h_\lambda, \epsilon)$ – вероятность поглощения цепи Маркова L_n^+ , если статистики λ_k имеют плотность распределения вероятностей $h_\lambda(x)$.

Теорема 6. Для фиксированного значения ϵ и плотности распределения вероятностей $h^g(x)$ выполняется следующее неравенство: $\alpha^+(h_\lambda^g, \epsilon) \leq \alpha^+(\bar{f}_\lambda^g, \epsilon)$.

Доказательство следует из определения величины $\alpha_m^+(h^g, \epsilon)$ и способа построения плотностей распределения вероятностей $h_\lambda^g(x)$ и $\bar{f}_\lambda^g(x)$.

Следствие 2. Вероятность ошибки первого рода $\alpha_m^+(\bar{f}^g, \epsilon)$ монотонно возрастает по параметру ϵ , $0 < \epsilon \leq \epsilon_0$, т. е. выполняется следующее неравенство:

$$\alpha_m^+(\bar{f}^g, \epsilon) \leq \alpha_m^+(\bar{f}^g, \epsilon_0).$$

Вычислительные эксперименты проводились для случая (рис. 1)

$$f(x; \theta) = (1/\sqrt{2\pi}) \exp\{-(x - \theta)^2 / 2\}, \quad \alpha_0 = \beta_0 = 0.1, \quad \theta_1 = 0.5, \quad \theta_0 = 0.$$

Для кривых «1», «2», «3» на рис. 2 параметры усечения соответственно равны: $g_+ = 1.8948$, $g_+ = 1.5316$, $g_+ = 0.7744$; $g_- = -g_+$. Заметим, что, меняя указанные параметры, можно уменьшить вероятность ошибки 1-го рода даже при наличии искажений в теоретической модели.

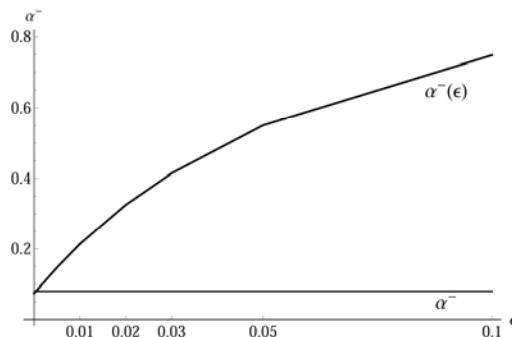


Рис. 1. Зависимость вероятности ошибки 1-го рода $\alpha_m^-(\epsilon)$ от уровня искажения ϵ

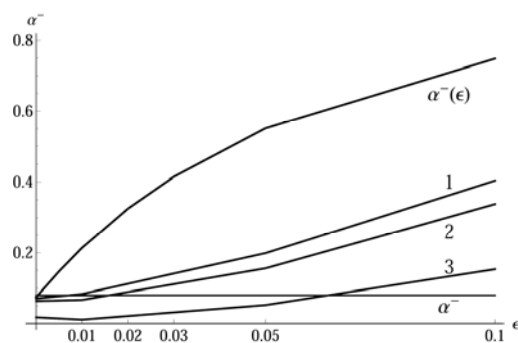


Рис.2. Зависимость вероятности ошибки 1-го рода $\alpha_m^-(\bar{f}^g, \epsilon)$ от уровня искажения ϵ и параметров усечения g_+

ЛИТЕРАТУРА

1. Вальд, А. Последовательный анализ / А. Вальд. М.: Наука, 1960.
2. Kharin, A. Robust sequential testing of hypotheses on discrete probability distributions / A. Kharin, D. Kishylau. Austrian Journal of Statistics, 2005, 34 (2). P. 153–162.
3. Lai, T. L. Sequential analysis: Some classical problems and new challenges / T. L. Lai. Statistica Sinica, 2001, 11. P. 303–408.
4. Mukhopadhyay, N. Applied Sequential Methodologies. / N. Mukhopadhyay, S. Datta, S. Chattopadhyay. Marcel Dekker, New York, 2004.
5. Харин, А. Ю. Оценивание вероятностей ошибок последовательного критерия отношения вероятностей / А. Ю. Харин, С. Ю. Чернов // Вестн. БГУ. 2011. № 1. С. 96–100.