РОБАСТНОСТЬ ПОСЛЕДОВАТЕЛЬНОЙ СТАТИСТИЧЕСКОЙ ПРОВЕРКИ ГИПОТЕЗ ПРИ ИСКАЖЕНИЯХ ПАРАМЕТРИЧЕСКОЙ МОДЕЛИ

А. Ю. Харин

Белорусский государственный университет
Минск, Беларусь
E-mail: KharinAY@bsu.by

Рассмотрена проблема робастности в последовательной статистической проверке гипотез при искажениях модельных предположений. Проведён анализ робастности для случая проверки сложных гипотез.

Ключевые слова: последовательный тест, параметр, искажения, робастность.

ВВЕДЕНИЕ

Последовательный анализ начал формироваться как направление в математической и прикладной статистике после публикации А. Вальдом в 1947 г. монографии [1]. Большой вклад в становление и развитие этого научного направления внесли А.Н. Ширяев [2], Д. Сигмунд [3], П.К. Сен, Б. Гош [4]. Интенсивное развитие последовательного статистического анализа было обусловлено возрастающим применением статистических методов при решении практических задач [5]. В таких задачах возникла потребность получать статистические выводы заданной точности с использованием лишь минимального числа наблюдений, необходимого для обеспечения требуемой точности. Такие задачи возникли во многих приложениях, но наибольшее стимулирующее воздействие на развитие вероятностно-статистических методов последовательного анализа оказали техника, статистический контроль качества, медицина, где требуемое число наблюдений является важнейшей характеристикой наряду с точностью принятия решений.

Поскольку наблюдаемые в практических задачах статистические данные не всегда адекватно описываются гипотетической вероятностной моделью [6], при таких искажениях оптимальные свойства [7] последовательных статистических критериев (тестов) могут нарушаться. Поэтому задача анализа робастности [8] (устойчивости) таких критериев при наличии искажений гипотетической вероятностной модели является актуальной [9].

РОБАСТНОСТЬ В ПОСЛЕДОВАТЕЛЬНОЙ ПРОВЕРКЕ ПРОСТЫХ ПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Вопрос о робастности последовательного критерия отношения вероятностей (критерия Вальда) для проверки простых параметрических гипотез был поставлен ещё П. Хьюбером [8]. Эта задача решалась теоретически лишь в некоторых частных случаях. Основная сложность теоретического решения этой задачи состоит в том, что вероятностные характеристики последовательных критериев не могут быть записаны в явном виде даже для простейших вероятностных моделей наблюдений, что затрудняет аналитическое исследование влияния искажений модели на эти характеристики.

В работах [10], [11] предложен подход к приближенному вычислению характеристик последовательного критерия отношения вероятностей, позволивший теоретически исследовать его робастность при наличии искажений гипотетического распределения вероятностей наблюдений. В [11], [12] построены новые последовательные критерии, робастные при наличии «выбросов» для моделей дискретных независимых наблюдений, а также предложен поход, позволивший обобщить результаты на случай произвольного вероятностного распределения наблюдений. Для модели, когда наблюдения образуют цепь Маркова, подобные результаты были получены в [13]. В работе [14] исследована ситуация, когда искажения затрагивают модель зависимости наблюдений.

В [15] предложен и исследован подход к оцениванию характеристик последовательного критерия Вальда, основанный на так называемых граничных цепях Маркова, позволяющий исследовать робастность критерия при наличии функциональных искажений вероятностных распределений. Подход к анализу робастности проверки сложных гипотез представлен в [16].

На практике, однако, модель, в которой гипотезы формулируются как простые, не всегда применима. Приведем некоторые результаты, касающиеся анализа робастности последовательных критериев проверки сложных гипотез.

ПОСЛЕДОВАТЕЛЬНАЯ ПРОВЕРКА СЛОЖНЫХ ГИПОТЕЗ ПРИ ОДНОВРЕМЕННЫХ ИСКАЖЕНИЯХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ПАРАМЕТРОВ И НАБЛЮДЕНИЙ

Пусть на измеримом пространстве $(\Omega, \Im)$ наблюдается последовательность случайных величин $x_1, x_2, \ldots \in \mathbf{R}$, имеющих n-мерную плотность распределения вероятностей $p_n(x_1,\ldots,x_n \mid \theta)$, $n \in \mathbf{N}$, $\theta \in \Theta \subseteq \mathbf{R}^k$ – неизвестное значение случайного вектора параметров, плотность распределения вероятностей которого $p(\theta)$ предполагается известной. Относительно значения параметра θ имеются две сложные гипотезы:

$$H_0\colon \theta \in \Theta_0,\ H_1\colon \theta \in \Theta_1;\ \ \Theta_0 \cup \Theta_1 = \Theta,\ \ \Theta_0 \cap \Theta_1 = \varnothing. \tag{1}$$

Примем обозначения:

$$\mathbf{1}_S(s) = \begin{cases} 1, s \in S, \\ 0, s \notin S; \end{cases}$$

$$W_i = \int\limits_{\Theta_i} p(\theta) d\theta,\ \ w_i(\theta) = \tfrac{1}{W_i} \cdot p(\theta) \cdot \mathbf{1}_{\Theta_i}(\theta),\ \ \theta \in \Theta,\ \ i = 0,\ 1;$$

$$\Lambda_n = \Lambda_n(x_1,\ldots,x_n) = \ln \frac{\int\limits_{\Theta} w_1(\theta) p_n(x_1, x_2, \ldots, x_n \mid \theta)\, d\theta}{\int\limits_{\Theta} w_0(\theta) p_n(x_1, x_2, \ldots, x_n \mid \theta)\, d\theta}. \tag{2}$$

Для проверки гипотез (1) используется параметрическое семейство последовательных статистических критериев:

$$N = \min\{n \in \mathbf{N} : \Lambda_n \notin (C_-, C_+)\},\ \ d = \mathbf{1}_{[C_+, +\infty)}(\Lambda_N), \tag{3}$$

где N – случайный номер наблюдения, момент остановки, после которого принимается решение d; решение $d = i$ означает принятие гипотезы H_i, i = 0,1. В (3) $C_- < 0$, $C_+ > 0$ –

параметры, называемые порогами критерия; на практике их значения полагают равными $C_- = \ln\frac{\beta_0}{1-\alpha_0}$, $C_+ = \ln\frac{1-\beta_0}{\alpha_0}$, где $\alpha_0, \beta_0 \in (0, \frac{1}{2})$ – величины, близкие к приемлемым значениям вероятностей ошибок I и II рода [1]. Фактические значения α, β вероятностей ошибок I и II рода могут существенно отличаться от α_0, β_0 [12].

Для вычисления α, β и математического ожидания случайного числа наблюдений *N* построим стохастическую аппроксимацию статистики Λ_n, $n \in \mathbf{N}$, с параметром $m \in \mathbf{N}$. Пусть $h = (C_+ - C_-)/m$; $p_{\Lambda_n}(u)$ – плотность распределения вероятностей статистики (2); $p_{\Lambda_{n+1}|\Lambda_n}(u \mid y)$ – условная плотность распределения вероятностей, $n \in \mathbf{N}$; $[x]$ – целая часть *x* (наибольшее целое число, меньшее либо равное *x*). Построим дискретную случайную последовательность Z_n^m, $n = 0,1,2,\ldots$, $Z_0^m = 0$, с пространством состояний $V = \{0,1,\ldots,m+1\}$:

$$Z_n^m = \begin{cases} 0, & Z_{n-1}^m = 0, \\ m+1, & Z_{n-1}^m = m+1, \\ \left(\left[\frac{\Lambda_n - C_-}{h}\right]+1\right)\cdot \mathbf{1}_{(C_-,C_+)}(\Lambda_n) + (m+1)\cdot \mathbf{1}_{(C_+,+\infty)}(\Lambda_n), & \text{иначе.} \end{cases} \tag{4}$$

Для случайной последовательности (4) рассмотрим $((m+2)\times(m+2))$-матрицу условных вероятностей

$$P^{(n)}(\theta) = (p_{ij}^{(n)}(\theta)) = \left(\mathrm{P}\left\{Z_{n+1}^m = j \mid Z_n^m = i\right\}\right), \; i, j \in V, \; n \in \mathbf{N}.$$

Случайную последовательность Z_n^m аппроксимируем цепью Маркова $z_n^m \in V$, $n \in \mathbf{N}$, с таким же начальным распределением вероятностей, соответствующим моменту $n = 1$, и матрицей переходных вероятностей $P^{(n)}(\theta)$ в момент *n*. С использованием перенумерации состояний $V = \{\{0\},\{m+1\},\{1\},\ldots,\{m\}\}$ матрицу $P^{(n)}(\theta)$ можно представить в блочном виде:

$$P^{(n)}(\theta) = \begin{pmatrix} \mathbf{I}_2 & | & \mathbf{0}_{2\times m} \\ --- & | & --- \\ R^{(n)}(\theta) & | & Q^{(n)}(\theta) \end{pmatrix}, \; \theta \in \Theta,$$

где блоки $R^{(n)}(\theta)$ и $Q^{(n)}(\theta)$ имеют размерность $m \times 2$ и $m \times m$ соответственно, $\mathbf{I}_k$ – единичная матрица размерности *k*, $\mathbf{0}_{l\times m}$ – матрица размерности $l \times m$, все элементы которой равны 0. Пусть $\pi(\theta) = (\pi_i(\theta))$ – вектор начальных вероятностей состояний 1,…,*m* для случайной последовательности (4); $\pi_0(\theta)$, $\pi_{m+1}(\theta)$ – начальные вероятности поглощающих состояний, $\mathbf{1}_m$ – вектор размерности *m*, все компоненты которого равны 1. Обозначим матрицы:

$$S(\theta) = \mathbf{I}_m + \sum_{i=1}^{\infty}\prod_{j=1}^{i} Q^{(j)}(\theta); \quad B(\theta) = R^{(1)}(\theta) + \sum_{i=1}^{\infty}\prod_{j=1}^{i} Q^{(j)}(\theta) R^{(i+1)}(\theta).$$

Пусть $B_{(j)}(\theta)$ – столбец номер *j* матрицы $B(\theta)$, $j = 1,2$; $t_i = \mathrm{E}\{N \mid \theta \in \Theta_i\}$, $i = 0,1$, $t = \mathrm{E}\{N\}$; $\gamma_{H_i}(\theta)$ и $t(\theta)$ – вероятность принять гипотезу H_i и математическое ожидание случайной величины *N* соответственно при условии, что вектор параметров принял значение θ.

Пусть для проверки гипотез (1) используется последовательный статистический критерий (2), (3), построенный с использованием гипотетических плотностей распределения вероятностей $p(\theta)$, $p_n(x_1,\dots,x_n\,|\,\theta)$, однако фактически эти функции искажены. Фактическая плотность распределения вероятностей вектора параметров имеет вид:

$$\bar{p}(\theta)=(1-\varepsilon_\theta)p(\theta)+\varepsilon_\theta\cdot\tilde{p}(\theta),\ \ \theta\in\Theta, \tag{5}$$

где $\varepsilon_\theta\in[0,\frac{1}{2}]$ – вероятность появления «засорения» [9], а $\tilde{p}(\theta)$ – «засоряющая» плотность распределения вероятностей, отличная от $p(\theta)$. Наблюдения подвержены «выбросам» [9], то есть фактическая условная плотность распределения вероятностей наблюдений представляет собой смесь

$$\bar{p}_n(x_1,\dots,x_n\,|\,\theta)=(1-\varepsilon_x)\cdot p_n(x_1,\dots,x_n\,|\,\theta)+\varepsilon_x\cdot\tilde{p}_n(x_1,\dots,x_n\,|\,\theta), \tag{6}$$

$$\theta\in\Theta,\ \ x_1,x_2,\ldots\in\mathbf{R},\ \ n\in\mathbf{N},$$

где $\varepsilon_x\in[0,\frac{1}{2}]$ – вероятность появления «выброса» в наблюдениях, $\tilde{p}_n(x_1,\dots,x_n\,|\,\theta)$ – «засоряющая» условная плотность распределения вероятностей наблюдений.

Пусть $\tilde{\pi}(\theta)$, $\tilde{\pi}_0(\theta)$, $\tilde{\pi}_{m+1}(\theta)$, $Q^{(n)}(\theta)$, $\tilde{R}^{(n)}(\theta)$ вычислены аналогично $\pi(\theta)$, $\pi_0(\theta)$, $\pi_{m+1}(\theta)$, $Q^{(n)}(\theta)$, $R^{(n)}(\theta)$ заменой плотности распределения вероятностей $p_n(x_1,\dots,x_n\,|\,\theta)$ на «засоряющую» плотность $\tilde{p}_n(x_1,\dots,x_n\,|\,\theta)$ в распределении вероятностей случайной последовательности (4); $\Delta\pi_0(\theta)=\tilde{\pi}_0(\theta)-\pi_0(\theta)$, $\Delta\pi_1(\theta)=\tilde{\pi}_{m+1}(\theta)-\pi_{m+1}(\theta)$; $\bar{t}(\theta)$, $\bar{t}_i(\theta)$, $\bar{\gamma}_{H_i}(\theta)$ – соответствующие характеристики последовательного критерия (2), (3), вычисленные при искажениях (5), (6). Обозначим: $\tilde{W}_i=\int\limits_{\Theta_i}\tilde{p}(\theta)d\theta$;

$$A(\theta)=\left((\tilde{\pi}(\theta)-\pi(\theta))'S(\theta)+(\pi(\theta))'\sum_{i=1}^{\infty}\sum_{j=1}^{i}\prod_{k=1}^{j-1}Q^{(k)}(\theta)(\tilde{Q}^{(j)}(\theta)-Q^{(j)}(\theta))\prod_{k=j+1}^{i}Q^{(k)}(\theta)\right)\cdot\mathbf{1}_m,$$

$$F_i(\theta)=\Delta\pi_i(\theta)+(\tilde{\pi}(\theta)-\pi(\theta))'(B_{(i+1)}(\theta)+\tilde{R}^{(1)}_{(i+1)}(\theta)-R^{(1)}_{(i+1)}(\theta)+$$

$$+\sum_{l=1}^{\infty}\sum_{j=1}^{l}\prod_{k=1}^{j-1}Q^{(k)}(\theta)(\tilde{Q}^{(j)}(\theta)-Q^{(j)}(\theta))\prod_{k=j+1}^{l}Q^{(k)}(\theta))R^{(l+1)}_{(i+1)}(\theta)+$$

$$+\prod_{j=1}^{l}Q^{(j)}(\theta)(\tilde{R}^{(l+1)}_{(i+1)}(\theta)-R^{(l+1)}_{(i+1)}(\theta))),\ i=0,1.$$

Теорема 1. Пусть случайная последовательность (2) является марковской, $\forall\theta\in\Theta$, плотности распределения вероятностей $p_{\Lambda_1}(u)$, $p_{\Lambda_{n+1}|\Lambda_n}(u\,|\,y)$ – дифференцируемые функции переменной $u\in[C_-,C_+]$, причем существует $C>0$:

$$\left|p_{\Lambda_1}{}'(u)\right|\le C,\ \ \left|p_{\Lambda_{n+1}|\Lambda_n}{}'(u\,|\,y)\right|\le C,\ \ u,y\in[C_-,C_+],\ \ n\in\mathbf{N}.$$

Тогда при искажениях (5), (6) гипотетической модели в асимптотике $\varepsilon_\theta\to 0$, $\varepsilon_x\to 0$, $h\to 0$ имеют место следующие разложения вероятностей ошибок I и II рода для последовательного статистического критерия (2), (3):

$$\bar{\alpha}=\alpha+\varepsilon_x\cdot\tfrac{1}{W_0}\cdot\int\limits_{\Theta_0}F_1(\theta)\cdot p(\theta)d\theta+$$

$$+\varepsilon_\theta\cdot\left(\tfrac{1}{W_0^2}\cdot\int\limits_{\Theta_0}(\tilde{p}(u)-p(u))du\cdot\int\limits_{\Theta_0}\gamma_{H_1}(\theta)p(\theta)d\theta+\tfrac{1}{W_0}\cdot\int\limits_{\Theta_0}\gamma_{H_1}(\theta)(\tilde{p}(\theta)-p(\theta))d\theta\right)+$$

$$+O(\varepsilon_x^2)+O(\varepsilon_\theta^2)+O(h),$$

$$\overline{\beta} = \beta + \varepsilon_x \cdot \tfrac{1}{W_1} \cdot \int\limits_{\Theta_1} F_0(\theta) \cdot p(\theta) d\theta +$$

$$+ \varepsilon_\theta \cdot \left(\tfrac{1}{W_1^2} \cdot \int\limits_{\Theta_1} (\widetilde{p}(u) - p(u)) du \cdot \int\limits_{\Theta_1} \gamma_{H_0}(\theta) p(\theta) d\theta + \tfrac{1}{W_1} \cdot \int\limits_{\Theta_1} \gamma_{H_0}(\theta)(\widetilde{p}(\theta) - p(\theta)) d\theta \right) +$$

$$+ O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h).$$

Теорема 2. Если выполнены условия Теоремы 1, то в асимптотике $\varepsilon_\theta \to 0$, $\varepsilon_x \to 0$, $h \to 0$ справедливы разложения для условных и безусловного математических ожиданий случайного числа наблюдений N:

$$\bar{t}_i = t_i + \varepsilon_x \cdot \tfrac{1}{W_i} \cdot \int\limits_{\Theta_i} A(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \left(t_i (\widetilde{W}_i - W_i) + \tfrac{1}{W_i} \cdot \int\limits_{\Theta_i} t(\theta)(\widetilde{p}(\theta) - p(\theta)) \right) +$$

$$+ O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h), \quad i = 0{,}1;$$

$$\bar{t} = t + \varepsilon_x \cdot \int\limits_{\Theta} A(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \int\limits_{\Theta} t(\theta)(\widetilde{p}(\theta) - p(\theta)) + O(\varepsilon_x^2) + O(h).$$

Полученные результаты позволяют оценить, насколько сильно искажения (5), (6) влияют на вероятностные характеристики последовательного критерия (2), (3), и могут быть использованы при построении минимаксных робастных последовательных статистических критериев в соответствии со схемой, аналогичной разработанной в [11].

ЛИТЕРАТУРА

1. *Wald, A.* Sequential analysis / A. Wald. N.Y. : Sringer, 1947.
2. *Ширяев, А. Н.* Статистический последовательный анализ / А.Н. Ширяев. М. : Наука, 1976.
3. *Siegmund, D.* Sequential analysis. Tests and confidence intervals / D. Siegmund. N.Y. : Sringer, 1985.
4. Handbook of sequential analysis / Ed. by B. Ghosh, P. K. Sen. N.Y., 1991.
5. *Mukhopadhyay, N.* Sequential methods and their applications / N. Mukhopadhyay, B. de Silva. N.Y. : 2009.
6. *Kharin, A.* Robust multivariate Bayesian forecasting under functional distortions in the chi-square metric / A. Kharin, P. Shlyk // Journal of Statistical Planning and Inference. 2009. V. 139. P. 3842–3846.
7. *Айвазян, С. А.* Сравнение оптимальных свойств критериев Неймана – Пирсона и Вальда / С. А. Айвазян // Теория вероятностей и ее применения. 1959. № 4 (1). С. 86–93.
8. *Хьюбер, П.* Робастность в статистике / П. Хьюбер. М. : Мир, 1984.
9. *Huber, P.* Robust statistics / P. Huber, E. Ronchetti. N.Y. : Wiley, 2009.
10. *Харин, А. Ю.* Об одном подходе к анализу последовательного критерия отношения правдоподобия для различения простых гипотез / А. Ю. Харин // Вестник Белорусского государственного университета. Сер. физ., мат., инф. 2002, № 1. С. 92–96.
11. *Kharin, A.* On robustifying of the sequential probability ratio test for a discrete model under “contaminations” / A. Kharin // Austrian Journal of Statistics. 2002. V. 31 (4). P. 267–278.
12. *Kharin, A.* Robust sequential testing of hypotheses on discrete probability distributions / A. Kharin, D. Kishylau // Austrian Journal of Statistics. 2005. V. 34 (2). P. 153–162.
13. *Kharin, A.* Robustness in sequential discrimination of Markov chains under "contamination" / A. Kharin // Theory and applications of recent robust methods. Basel : Springer, 2004. P. 165–172.
14. *Кишилов, Д. В.* Об устойчивости последовательного теста Вальда к нарушению предположения о независимости наблюдений / Д. В. Кишилов // Статистические методы оценивания и проверки гипотез. 2006. Вып. 19. С. 34–42.
15. *Харин, А. Ю.* Оценивание вероятностей ошибок последовательного критерия отношения вероятностей / А. Ю. Харин, С. Ю. Чернов // Вестник БГУ. Сер. 1. 2011. № 1. С. 96–100.
16. *Kharin, A.* Robustness analysis for Bayesian sequential testing of composite hypotheses under simultaneous distortion of priors and likelihoods / A. Kharin // Austrian Journal of Statistics. 2011. V. 40 (1&2). P. 65–73.