УДК 519.2

А.Ю. ХАРИН

РОБАСТНОСТЬ ПОСЛЕДОВАТЕЛЬНЫХ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ

The problem of robustness in sequential statistical testing of parametric hypotheses under distortions of the model is considered. A survey of recent results is given. Robustness analysis is performed in case of composite hypotheses.

Последовательный анализ начал формироваться как направление в математической и прикладной статистике после выхода в свет в 1947 г. монографии [1] А. Вальда. Большой вклад в становление и развитие этого научного направления внесли А.Н. Ширяев [2], Д. Сигмунд [3], П.К. Сен, Б. Гош [4], Т. Лай [5]. Интенсивное развитие последовательного статистического анализа было обусловлено возрастающим применением статистических методов при решении практических задач [6], поскольку возникла потребность получать статистические выводы заданной точности с использованием минимального необходимого числа наблюдений. Такие задачи появились 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 [7], в этих случаях оптимальные свойства [8] последовательных статистических критериев (тестов) могут нарушаться. Поэтому задача анализа робастности [9] (устойчивости) таких критериев при наличии искажений гипотетической вероятностной модели является актуальной [10].

Робастность последовательной проверки простых параметрических гипотез

Вопрос о робастности последовательного критерия отношения вероятностей (критерия Вальда) для проверки простых параметрических гипотез был поставлен еще П. Хьюбером [9]. Аналитически эта задача решалась лишь в некоторых частных случаях (см., например, [11]). 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 [5], что затрудняет аналитическое исследование влияния искажений модели на эти характеристики.

В работах [12, 13] предложен подход к приближенному вычислению характеристик последовательного критерия отношения вероятностей, позволивший теоретически исследовать его робастность в случае искажений гипотетического распределения вероятностей наблюдений. В [13, 14] построены новые последовательные критерии, робастные при наличии «выбросов» для моделей дискретных независимых наблюдений, а также предложен подход, позволивший обобщить результаты для произвольного вероятностного распределения наблюдений. Для модели, когда наблюдения образуют цепь Маркова, подобные результаты были получены в [15], а в [16] исследована ситуация, когда искажения затрагивают модель зависимости наблюдений.

В [17] предложен и исследован подход к оцениванию характеристик последовательного критерия Вальда, основанный на так называемых граничных цепях Маркова, позволяющий исследовать робастность критерия при функциональных искажениях вероятностных распределений.

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 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:

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

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

$$\Lambda_n=\Lambda_n(x_1,...,x_n)=\ln\frac{\int_{\Theta} w_1(\theta)p_n(x_1,x_2,...,x_n\,|\,\theta)\,d\theta}{\int_{\Theta} w_0(\theta)p_n(x_1,x_2,...,x_n\,|\,\theta)\,d\theta}. \tag{2}$$

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 $\alpha_0,\beta_0\in(0,\frac{1}{2})$ – приемлемые значения 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 $h=\frac{C_+-C_-}{m}$; $p_{\Lambda_n}(u)$ – плотность распределения вероятностей статистики (2); $p_{\Lambda_{n+1}|\Lambda_n}(u\,|\,y)$ – 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, $n=0,1,2,...$, $Z_0^m=0$, с пространством состояний $V=\{0,1,...,m+1\}$:

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

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

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

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×2 и m×m соответственно, $\mathbf{I}_k$ – единичная матрица размерности k, $\mathbf{0}_{l\times m}$ – матрица размерности l×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:

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

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 принятия гипотезы 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 $p(\theta)$, $p_n(x_1, \ldots, x_n \mid \theta)$, однако фактически эти функции искажены. Фактическая плотность распределения вероятностей вектора параметров имеет вид

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

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

$$\overline{p}_n(x_1, ..., x_n \mid \theta) = (1 - \varepsilon_x) \cdot p_n(x_1, ..., x_n \mid \theta) + \varepsilon_x \cdot \tilde{p}_n(x_1, ..., x_n \mid \theta), \tag{6}$$

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

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

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

$$\tilde{W}_i = \int_{\Theta_i} \tilde{p}(\theta) d\theta;$$

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

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

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

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

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

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)*:

$$\overline{\alpha} = \alpha + \varepsilon_x \cdot \frac{1}{W_0} \cdot \int_{\Theta_0} F_1(\theta) \cdot p(\theta) d\theta +$$

$$+ \varepsilon_\theta \cdot \left(\frac{1}{W_0^2} \cdot \int_{\Theta_0} (\tilde{p}(u) - p(u)) du \cdot \int_{\Theta_0} \gamma_{H_1}(\theta) p(\theta) d\theta + \frac{1}{W_0} \cdot \int_{\Theta_0} \gamma_{H_1}(\theta) (\tilde{p}(\theta) - p(\theta)) d\theta \right) + O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h),$$

$$\overline{\beta} = \beta + \varepsilon_x \cdot \frac{1}{W_1} \cdot \int_{\Theta_1} F_0(\theta) \cdot p(\theta) d\theta +$$

$$+\varepsilon_\theta \cdot \left(\frac{1}{W_1^2} \cdot \int_{\Theta_1} (\tilde{p}(u) - p(u)) du \cdot \int_{\Theta_1} \gamma_{H_0}(\theta) p(\theta) d\theta + \frac{1}{W_1} \cdot \int_{\Theta_1} \gamma_{H_0}(\theta)(\tilde{p}(\theta) - p(\theta)) d\theta \right) + (\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h).$$

Доказательство использует результат работы [18], справедливый в условиях теоремы 1:

$$\overline{\gamma}_{H_i}(\theta) = \gamma_{H_i}(\theta) + \varepsilon_x \cdot F_i(\theta) + O(\varepsilon_x^2) + O(h), \quad \theta \in \Theta. \tag{7}$$

Строятся асимптотические разложения для $\int_{\Theta_i} \overline{p}(\theta)\, d\theta,\ i = 0{,}1.$ Результаты для $\overline{\alpha},\ \overline{\beta}$ получаются путем подстановки этих разложений в (7) и тождественных преобразований. ■

Следствие 1. Если в условиях теоремы 1 выполняется равенство

$$\int_{\Theta_0} \tilde{p}(\theta) d\theta = \int_{\Theta_0} p(\theta) d\theta, \tag{8}$$

то асимптотические разложения для вероятностей ошибок I и II рода принимают вид

$$\overline{\alpha} = \alpha + \varepsilon_x \cdot \frac{1}{W_0} \cdot \int_{\Theta_0} F_1(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \frac{1}{W_0} \int_{\Theta_0} \gamma_{H_1}(\theta) \big(\tilde{p}(\theta) - p(\theta) \big) d\theta + O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h),$$

$$\overline{\beta} = \beta + \varepsilon_x \cdot \frac{1}{W_1} \cdot \int_{\Theta_1} F_0(\theta) p(\theta)\ d\theta + \varepsilon_\theta \cdot \frac{1}{W_1} \int_{\Theta_1} \gamma_{H_0}(\theta) \big(\tilde{p}(\theta) - p(\theta) \big) d\theta + O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h).$$

Отметим, что выполнение условия (8) означает, что при искажении априорного распределения вероятностей (5) не изменяются априорные вероятности гипотез (1); происходит лишь «перераспределение» вероятностей внутри множеств Θ_0, Θ_1.

Следствие 2. Если в условиях теоремы 1 отсутствуют искажения распределения вероятностей наблюдений, т. е. $\varepsilon_x = 0$, то

$$\overline{\alpha} = \alpha + \varepsilon_\theta \cdot \frac{1}{W_0} \int_{\Theta_0} \Big(\big(\pi(\theta) \big)' B_{(2)}(\theta) + \pi_{m+1}(\theta) - \alpha \Big) \big(\tilde{p}(\theta) - p(\theta) \big) d\theta + O(\varepsilon_\theta^2) + O(h),$$

$$\overline{\beta} = \beta + \varepsilon_\theta \cdot \frac{1}{W_1} \int_{\Theta_1} \Big(\big(\pi(\theta) \big)' B_{(1)}(\theta) + \pi_0(\theta) - \beta \Big) \big(\tilde{p}(\theta) - p(\theta) \big) d\theta + O(\varepsilon_\theta^2) + O(h).$$

В частном случае, когда отсутствуют искажения априорной плотности распределения вероятностей, т. е. при $\varepsilon_\theta = 0$, результат теоремы 1 соответствует [18].

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:*

$$\overline{t_i} = t_i + \varepsilon_x \cdot \frac{1}{W_i} \cdot \int_{\Theta_i} A(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \left(t_i (\tilde{W}_i - W_i) + \frac{1}{W_i} \cdot \int_{\Theta_i} t(\theta)(\tilde{p}(\theta) - p(\theta)) \right) +$$

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

$$\overline{t} = t + \varepsilon_x \cdot \int_{\Theta} A(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \int_{\Theta} t(\theta)(\tilde{p}(\theta) - p(\theta)) + O(\varepsilon_x^2) + O(h).$$

Доказательство. В условиях теоремы 2 из [18] получаем

$$\overline{t}(\theta) = t(\theta) + \varepsilon_x \cdot A(\theta) + O(\varepsilon_x^2) + O(h), \quad \theta \in \Theta. \tag{9}$$

Используя представление $\overline{t_i} = \int_{\Theta} \overline{t}(\theta) \cdot \frac{1}{\int_{\Theta_i} \overline{p}(u)\, du} \cdot \overline{p}(\theta) \cdot \mathbf{1}_{\Theta_i}(\theta) d\theta,\ i = 0{,}1,$ и проводя тождественные преобразования, получаем разложения для условных математических ожиданий числа наблюдений. Разложение для безусловного математического ожидания следует из (9) и представления $\overline{t} = \int_{\Theta} \overline{t}(\theta) p(\theta)\ d\theta.$ ■

Следствие 3. В условиях следствия 1 имеют место следующие асимптотические разложения:

$$\overline{t_i} = t_i + \varepsilon_x \cdot \frac{1}{W_i} \cdot \int_{\Theta_i} A(\theta) p(\theta) d\theta + \varepsilon_\theta \cdot \frac{1}{W_i} \cdot \int_{\Theta_i} t(\theta)(\tilde{p}(\theta) - p(\theta)) + O(\varepsilon_x^2) + O(\varepsilon_\theta^2) + O(h), \ i = 0{,}1.$$

Следствие 4. В условиях следствия 2 справедливы следующие асимптотические разложения:

$$\overline{t_i} = t_i + \varepsilon_\theta \cdot \frac{1}{W_i} \cdot \int_{\Theta_i} \Big(1 + \big(\pi(\theta)\big)' S(\theta) \cdot \mathbf{1}_m - t_i\Big)\big(\tilde{p}(\theta) - p(\theta)\big) d\theta + O(\varepsilon_\theta^2) + O(h), \quad i = 0,1;$$

$$\overline{t} = t + \varepsilon_\theta \cdot \int_{\Theta} \big(\pi(\theta)\big)' S(\theta) \cdot \mathbf{1}_m \cdot \big(\tilde{p}(\theta) - p(\theta)\big) d\theta + O(h).$$

В случае $\varepsilon_\theta = 0$ результат теоремы 2 соответствует работе [18].

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 [13].

1. Wald A. Sequential analysis. New York, 1947.
2. Ширяев А.Н. Статистический последовательный анализ. М., 1976.
3. Siegmund D. Sequential analysis. Tests and confidence intervals. New York, 1985.
4. Handbook of sequential analysis / Ed. by B. Ghosh, P.K. Sen. New York, 1991.
5. Lai T. // Statistica Sinica. 2001. Vol. 11. P. 303.
6. Mukhopadhyay N., de Silva B. Sequential methods and their applications. New York, 2009.
7. Kharin A., Shlyk P. // Journal of Statistical Planning and Inference. 2009. Vol. 139. P. 3842.
8. Айвазян С.А. // Теория вероятностей и ее применения. 1959. № 4 (1). С. 86.
9. Хьюбер П. Робастность в статистике. М., 1984.
10. Huber P., Ronchetti E. Robust statistics. New York, 2009.
11. Quang P. // Annals of Statistics. 1985. Vol. 13 (2). P. 638.
12. Харин А.Ю. // Вестн. БГУ. Сер. 1. 2002. № 1. С. 92.
13. Kharin A. // Austrian Jornal of Statistics. 2002. Vol. 31 (4). P. 267.
14. Kharin A., Kishylau D. // Austrian Jornal of Statistics. 2005. Vol. 34 (2). P. 153.
15. Kharin A. // Theory and applications of recent robust methods. Basel, 2004. P. 165.
16. Кишилов Д.В. // Статистические методы оценивания и проверки гипотез. 2006. Вып. 19. С. 34.
17. Харин А.Ю., Чернов С.Ю. // Вестн. БГУ. Сер. 1. 2011. № 1. С. 96.
18. Kharin A. // Austrian Journal of Statistics. 2008. Vol. 37 (1). P. 51.

Поступила в редакцию 14.06.11.

Алексей Юрьевич Харин – кандидат физико-математических наук, доцент кафедры теории вероятностей и математической статистики. Область научных интересов – математическая и прикладная статистика. Автор более 100 научных работ, в том числе 3 учебных пособий.