РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ
№ 4 2000

УДК 519.816

С. Н. Моисеев
(Воронеж)

РАЗЛИЧЕНИЕ ГИПОТЕЗ О НОРМАЛЬНОМ ИЛИ КОШИ-РАСПРЕДЕЛЕНИИ ВЫБОРКИ

По методу максимального правдоподобия синтезирован алгоритм различения двух альтернативных гипотез о нормальном или Коши-распределении выборки при неизвестных параметрах распределений. Получены и подтверждены моделированием выражения для вероятностей ошибок алгоритма.

Нормальное распределение при определенных условиях является предельным для сумм независимых одинаково распределенных случайных величин (НОРСВ). В общем случае предельными для сумм НОРСВ являются устойчивые распределения. Среди них особо следует выделить законы с целым показателем устойчивости (нормальный и Коши), которые занимают важное место в аналитическом плане в теории устойчивых распределений. По наблюдениям автора работы [1] эти законы чаще других устойчивых распределений появляются в различных приложениях. Поэтому, когда возникает задача подбора распределения к наблюдаемым данным, получающимся в результате суммирования большого числа НОРСВ с неизвестными статистическими характеристиками, наиболее естественным первым шагом представляется проверка их на принадлежность нормальному или Коши-закону распределения. Подобного рода задачи идентификации наблюдений появляются при анализе телекоммуникационных трафиков [2], эконометрических данных [3], геофизических наблюдений (профильтрованные значения электронной концентрации слоя E_s ионосферы [4]) и в других приложениях, где встречаются распределения с тяжелыми хвостами.

Настоящая работа посвящена синтезу и анализу алгоритмов различения гипотез о нормальном или Коши-распределении выборки в условиях полной априорной неопределенности относительно параметров распределений.

Задачу различения сформулируем следующим образом. Пусть проверяется сложная гипотеза H_0 о распределении независимых выборочных значений $\mathbf{x} = \|x_1, \ldots, x_n\|$ по нормальному закону с плотностью вероятностей

$$W_0(x; m, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x-m)^2}{2\theta}\right\}, \qquad \theta > 0, \tag{1}$$

против сложной альтернативы H_1 о распределении **x** по закону Коши с плотностью вероятностей

$$W_1(x; a, b) = \frac{b}{\pi[b^2 + (x-a)^2]}, \qquad b > 0. \tag{2}$$

Синтез. Правило принятия решения, построенное по методу максимального правдоподобия (МП) аналогично работе [5], имеет вид

$$\Lambda = \frac{\omega_0(\mathbf{x} \mid \hat{m}, \hat{\theta})}{\omega_1(\mathbf{x} \mid \hat{a}, \hat{b})} \underset{H_1}{\overset{H_0}{\gtrless}} 1, \tag{3}$$

где $\omega_0(\mathbf{x} \mid m, \theta) = \prod_{i=1}^{n} W_0(x_i; m, \theta)$ и $\omega_1(\mathbf{x} \mid a, b) = \prod_{i=1}^{n} W_1(x_i; a, b)$ – функции правдоподобия распределений (1) и (2) соответственно; $\hat{m}, \hat{\theta}$ и $\hat{a}, \hat{b}$ – оценки МП-параметров распределений (1) и (2), вычисленные при справедливости соответствующих гипотез H_0 и H_1:

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{m})^2, \tag{4}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{a})W_1(x_i; \hat{a}, \hat{b}) = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} W_1(x_i; \hat{a}, \hat{b}) = \frac{1}{a\pi\hat{b}}. \tag{5}$$

Представим алгоритм МП (3) в наиболее удобном для анализа виде:

$$L = \hat{S}_1 - \hat{S}_0 = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{W_0(x_i; \hat{m}, \hat{\theta})}{W_1(x_i; \hat{a}, \hat{b})} \underset{H_1}{\overset{H_0}{\gtrless}} 0, \tag{6}$$

где $\hat{S}_0 = -\frac{1}{n}\sum_{i=1}^{n} \ln W_0(x_i; \hat{m}, \hat{\theta}) = \frac{1}{2}\ln(2\pi e\hat{\theta})$, $\hat{S}_1 = -\frac{1}{n}\sum_{i=1}^{n} \ln W_1(x_i; \hat{a}, \hat{b})$ – выборочные оценки энтропий распределений (1), (2), проминимизированные по неизвестным параметрам, а статистика L представляет собой выборочную оценку расстояния Кульбака – Лейблера между распределениями W_0 и W_1 [6].

Практически использовать алгоритм МП (6) сложно из-за необходимости решать систему трансцендентных уравнений (5). Получим более простой квазиправдоподобный (КП) алгоритм, подставив в (6) вместо оценок МП $\{\hat{a}, \hat{b}\}$ оценки $\{a^*, b^*\}$, найденные по методу квантилей: $a^* = \mathrm{med}(x)$, $b^* = E^* = (x_{(0,75)} - x_{(0,25)})/2$, где $\mathrm{med}(x), E^*$ и $x_{(p)}$ – соответственно выборочные медиана, срединное отклонение и квантиль порядка p. В результате имеем КП-алгоритм, формально выраженный через статистики $a^*, b^*, \hat{\theta}$:

$$Q = \ln\left[b^*\sqrt{\frac{\pi}{2e\hat{\theta}}}\right] + \frac{1}{n}\sum_{i=1}^{n} \ln\left[1 + \left(\frac{x_i - a^*}{b^*}\right)^2\right] \underset{H_1}{\overset{H_0}{\gtrless}} 0. \tag{7}$$

Заменим оценку энтропии $\hat{S}_1$ в (6) на более простую $S_1^* = \ln(4\pi b^*)$, полученную на основании вида теоретической энтропии распределения Коши $S_1 = \ln(4\pi b)$. В результате имеем второй КП-алгоритм:

Ма МП (6) и контроля ошибок при его практическом применении необходимо найти аналитические выражения для вероятностей ошибок двух родов: вероятности $P(H_1 \mid H_0)$ принятия гипотезы H_1 при условии, что справедлива гипотеза H_0, и вероятности $P(H_0 \mid H_1)$ принятия гипотезы H_0, когда верна гипотеза H_1.

Пусть справедлива гипотеза H_0 о нормальном распределении выборки. В этом случае, разлагая статистику L (6) в ряд Тейлора по степеням $\hat{m}, \hat{\theta}, \hat{a}, \hat{b}$ в точке их средних значений, можно показать [6], что L будет асимптотически нормальной при $n \to \infty$ и для ее анализа достаточно ограничиться нулевым членом разложения. Иными словами, имеет место следующая сходимость по распределению:

$$L \xrightarrow{d} L_0 \xrightarrow{d} N[\mathrm{M}(L_0), \mathrm{D}(L_0)], \tag{9}$$

где $N(m,\theta)$ – нормальная случайная величина из распределения (1);

$$L_0 = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{W_0(x_i;\, m,\, \theta)}{W_1(x_i;\, m,\, z\sqrt{\theta})}, \tag{10}$$

$z = 0{,}61200\ldots$ – корень уравнения $z = -\frac{1}{2}\frac{d}{dz}\ln(1-\Phi(z))$,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \exp(-x^2/2)\,dx$$

– интеграл вероятностей. Среднее и дисперсия статистики L_0 не зависят от параметров распределения (1) (как, впрочем, и другие моменты) и вычисляются по формулам

$$\mathrm{M}(L_0) = \int_{-\infty}^{\infty} W_0(x;0,1) \ln \frac{W_0(x;\,0,\,1)}{W_1(x;\,0,\,z)}\,dx = 0{,}182758\ldots,$$

$$\mathrm{D}(L_0) = \frac{1}{n}\int_{-\infty}^{\infty} W_0(x;0,1)\left[\ln \frac{W_0(x;\,0,\,1)}{W_1(x;\,0,\,z)} - \mathrm{M}(L_0)\right]^2 dx = \frac{0{,}122276\ldots}{n}.$$

Таким образом, на основании (9) можно написать асимптотически точное с ростом n выражение для вероятности ошибки МП-алгоритма $P(H_1 \mid H_0) = P(L<0 \mid H_0)$:

$$P(H_1 \mid H_0) \to \Phi(-\mathrm{M}(L_0)/\sqrt{\mathrm{D}(L_0)}) = \Phi(-c_0\sqrt{n}), \tag{11}$$

где $c_0 = 0{,}5226\ldots$.

Пусть справедлива гипотеза H_1 о распределении выборки по закону Коши. В этом случае асимптотическое поведение статистики L будет отличаться от нормального. Представим L в виде

$$L = \eta_0 - \frac{1}{2}\ln(\eta_1 - \eta_2) + \frac{1}{2}\ln\frac{\pi}{2e n},$$

где $\eta_0 = \frac{1}{n}\sum_{i=1}^{n}\ln(1+\mathring{x}_i^2)$; $\eta_1 = \frac{1}{n^2}\sum_{i=1}^{n}\mathring{x}_i^2$, $\eta_2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathring{x}_i\right)^2$; $\mathring{x}_i = \frac{x_i - \hat{a}}{\hat{b}}$. Срединные (вероятные) отклонения $E(\cdot)$ статистик η_0, η_1, η_2, характеризующие ширину их распределений, при $n \to \infty$ находятся в соотношениях $E(\eta_2) = o(E(\eta_1))$, $E(\eta_0) = o(E(\ln \eta_1))$. Эти соотношения легко получить, учитывая, что при $n \to \infty$ $\eta_0 \xrightarrow{d} N[\ln 4, O(n^{-1})]$, $\frac{1}{n}\sum_{i=1}^{n}\mathring{x}_i \xrightarrow{d} x_1$. Следовательно, асимптотическое поведение L будет определяться статистикой L_1 следующего вида:

$$L \xrightarrow{d} L_1 = \frac{1}{2}\ln\frac{8\pi}{e n \eta}, \tag{12}$$

где

$$\eta = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{x_i - a}{b}\right)^2. \tag{13}$$

Замена η_1 на η в (12) допустима, поскольку $\eta_1 \xrightarrow{d} \eta$, $n \to \infty$. Отметим, что распределение случайной величины η, а следовательно, и L_1 не зависит от параметров a и b распределения (2).

Найдем асимптотическое распределение статистики η. Для линейно нормированной суммы НОРСВ предельное распределение, если оно существует, может быть только устойчивым [7]. Поэтому сразу можно написать четырехпараметрическую характеристическую функцию η при $n \to \infty$:

$$\Theta_\eta(u) \to \exp\{i\gamma u - \lambda|u|^\alpha[1 + i\beta u\omega(u,\alpha)/|u|]\}, \quad \alpha \in [0,2],\ |\beta| < 1,\ \lambda > 0, \tag{14}$$

где

$$\omega(u,\alpha) = \begin{cases} \operatorname{tg}(\pi\alpha/2), & \text{если } \alpha \neq 1; \\ \frac{2}{\pi}\ln|u|, & \text{если } \alpha = 1. \end{cases}$$

Для определения показателя устойчивости α в (14) воспользуемся известным результатом Б. В. Гнеденко, изложенным, например, в [8], об областях притяжения предельного устойчивого закона, согласно которому хвост функции распределения $F(x)$ случайной величины $\left(\frac{x_i - a}{b}\right)^2$ в сумме (13),

имеющей предельное распределение (14) с $\alpha < 2$, ведет себя с ростом x как $1 - F(x) = O(x^{-\alpha})L(x)$, где $L(x)$ – медленно меняющаяся функция

$$(L(\varepsilon x)/L(x) \to 1,\ x \to \infty,\ \forall \varepsilon > 0).$$

Поскольку $1 - F(x) = 1 - \frac{2}{\pi}\operatorname{arctg}\sqrt{x} = O(x^{-1/2})$, то в (14) $\alpha = 1/2$.

Очевидно, что вероятностная мера величины η сосредоточена на всей положительной полуоси. Для устойчивого закона это выполняется, только если $\alpha < 1, \beta = 1$ [1]. Поэтому положим в (14) $\beta = 1$.

Среди устойчивых распределений значению параметров $\alpha = 1/2$, $\beta = 1$ соответствует так называемое семейство распределений Леви с плотностями [1]

$$W_\eta(x) \to \begin{cases} \frac{\lambda}{\sqrt{2\pi}} x^{-3/2} \exp\left(-\frac{\lambda^2}{2(x-\gamma)}\right), & x > \gamma; \\ 0, & x \le \gamma. \end{cases} \tag{15}$$

В нашем случае $\gamma = 0$, так как только тогда распределение Леви сосредоточено на всей положительной полуоси.

При $\gamma = 0$ распределение (15) будет строго устойчивым. Для НОРСВ $\xi_1, \ldots, \xi_n$ из строго устойчивого распределения с показателем устойчивости α при любом n справедливо [8] $\frac{1}{n^{1/\alpha}} \sum_{i=1}^{n} \xi_i \stackrel{d}{=} \xi_1$. Поэтому параметр λ в (15) асимптотически не зависит от объема выборки n. Проще всего рассчитать λ методом Монте-Карло, используя связь между Леви случайной величиной и нормальной:

$$\eta \xrightarrow{d} \frac{\lambda^2}{N^2(0,1)}.$$

Отсюда получаем формулу для расчета λ:

$$\lambda = \frac{\Phi^{-1}(3/4)}{\sqrt{\operatorname{med}(1/\eta)}} \approx 0{,}8,$$

где $\Phi^{-1}(\cdot)$ – функция, обратная интегралу вероятностей $\Phi(x)$.

Функция распределения статистики L_1 (12), однозначно связанной с η, при $n \to \infty$ имеет вид:

$$F_{L_1}(x) \to 2\Phi\left[\frac{\lambda\sqrt{n}\exp(x + 1/2)}{\sqrt{8\pi}}\right] - 1, \quad x \in (-\infty, \infty). \tag{16}$$

Из (12) и (16) следует асимптотически точное с ростом n выражение для вероятности ошибки МП-алгоритма $P(H_0 \mid H_1) = P(L > 0 \mid H_1)$:

$$P(H_0 \mid H_1) \to 2\Phi\left(-\lambda\sqrt{\frac{ne}{8\pi}}\right) = 2\Phi(-c_1\sqrt{n}), \tag{17}$$

где $c_1 \approx 0{,}26$.

Большие уклонения. Из равномерной сходимости функции распределения $F_n(x)$ нормированной суммы НОРСВ к нормальной функции распределения $\Phi(x)$, вытекающей из центральной предельной теоремы, следует [7], что соотношение

$$\frac{F_n(-x)}{\Phi(-x)} \to 1, \quad n \to \infty, \tag{18}$$

имеет место равномерно по x, когда x попадает в зону нормальной сходимости $[0,\ o(\sqrt{n})]$. В нашем случае, см. формулу (11), $x = c_0\sqrt{n}$ и сходимость (18) может не иметь места. Чтобы удовлетворить (18), воспользуемся формулой [7]

$$\frac{F_n(-x)}{\Phi(-x)} = g(x,n) = \exp\left\{-\frac{x^3}{\sqrt{n}} R\left(-\frac{x}{\sqrt{n}}\right)\right\}\left[1 + O\left(\frac{x+1}{\sqrt{n}}\right)\right]. \tag{19}$$

Здесь $R(x) = s_0 + s_1 x + s_2 x^2 + \ldots$ – ряд Крамера. Подставляя $x = c_0\sqrt{n}$ в (19), получаем $g(x,n) = \exp(c_{10}n + c_{20})$, где c_{10} и c_{20} – константы, не зависящие от параметров распределения (1), так как они определяются всеми семиинвариантами случайной величины L_0 (10) при $n = 1$. Наиболее просто рассчитать эти константы методом Монте-Карло по формуле

$$c_{10}n + c_{20} = \ln \frac{P_n^*(H_1 \mid H_0)}{\Phi(-c_0\sqrt{n})},$$

где $P_n^*(H_1 \mid H_0)$ – полученные моделированием значения вероятности ошибки $P(H_1 \mid H_0)$.

Уточненная для больших уклонений асимптотически точная формула для вероятности ошибки $P(H_1 \mid H_0)$ окончательно принимает вид:

$$P(H_1 \mid H_0) \to \Phi(-c_0\sqrt{n})\exp(c_{10}n), \tag{20}$$

где $c_{10} \approx 0{,}06$.

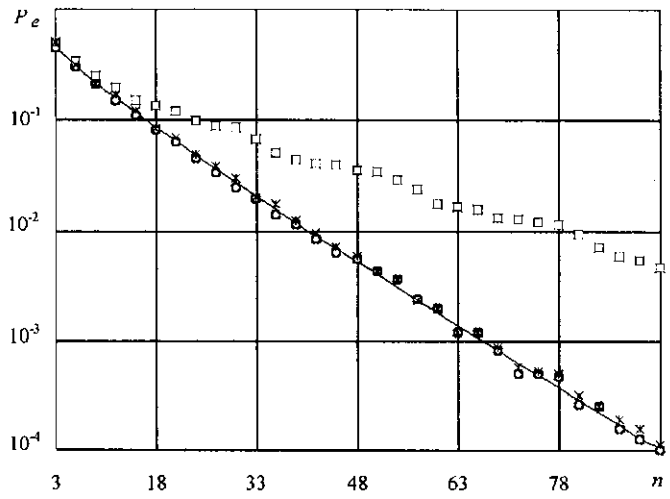
Аналогичное уточнение для вероятности ошибки $P(H_0 \mid H_1)$, основанное на похожести формул (11) и (17), приводит к асимптотически точному выражению

$$P(H_0 \mid H_1) \to 2\Phi(-c_1\sqrt{n})\exp(c_{11}n + c_{21}), \tag{21}$$

где $c_{11} \approx -0{,}05$, $c_{21} \approx 0{,}18$.

Моделирование. Для проверки точности полученных асимптотических формул проведено моделирование на ЭВМ алгоритмов различения (6)–(8). Объем испытаний для каждого n составил $1{,}6 \cdot 10^4$. Моделирование при различных наборах параметров распределений (1), (2) подтвердило вывод о независимости характеристик алгоритма различения от них. На рисунке сплошная кривая соответствует теоретической средней вероятности ошибки МП-алгоритма (6)

$$P_e = \frac{1}{2}[P(H_1 \mid H_0) + P(H_0 \mid H_1)], \tag{22}$$



рассчитанной по формулам (20), (21). Полученные моделированием значения средней вероятности ошибки МП-алгоритма (6) показаны на рисунке кружками, КП-алгоритма (7) – крестиками, КП-алгоритма (8) – квадратиками. Видно, что теоретическая кривая P_e (22) хорошо соответствует модельным значениям для произвольных $n \geq 3$. Заметим, что минимальный объем выборки n, начиная с которого возможно различение с нетривиальной $P_e < 0{,}5$ распределений с неизвестными параметрами сдвига и масштаба, равен 3. Моделирование показало, что МП-алгоритм (6) различает гипотезы H_0 и H_1 при $n = 3$ с $P_e \approx 0{,}44$, тогда как КП-алгоритмы (7), (8) – гипотезы с $P_e < 0{,}5$, начиная лишь с $n = 4$.

Из рисунка видно, что КП-алгоритм (8) заметно менее эффективен, чем алгоритмы (6) и (7). В то же время КП-алгоритм (7) почти не уступает МП-алгоритму (6). Так, при $n > 10$ средняя вероятность ошибки КП-алгоритма (7) превышает аналогичную характеристику МП-алгоритма (6) не более чем на 5 %. Поэтому для практического использования можно порекомендовать простой КП-алгоритм (7), синтез которого намного проще синтеза МП-алгоритма.

В качестве асимптотически точного выражения для вероятности ошибки $P(H_0 \mid H_1)$ КП-алгоритма (7) допустимо использовать (21), так как из проведенного выше анализа следует, что $L \xrightarrow{d} Q$, $n \to \infty$ при справедливости гипотезы H_1. Вывод асимптотически точной формулы для вероятности ошибки $P(H_1 \mid H_0)$ КП-алгоритма (7) полностью идентичен выводу формул (11) и (20) для МП-алгоритма за исключением того, что в (10) необходимо заменить z на $z = \Phi^{-1}(3/4) = 0{,}67448\ldots$. Следовательно, асимптотически точная формула для вероятности ошибки $P(H_1 \mid H_0)$ КП-алгоритма (7) совпадает с выражением (20), где следует положить $c_0 = 0{,}59168\ldots$, $c_{10} \approx 0{,}098$.

СПИСОК ЛИТЕРАТУРЫ

1. **Золотарев В. М.** Одномерные устойчивые распределения. М.: Наука, 1983.
2. **Resnick S. I.** Heavy tail modeling and teletraffic data. Ithaca, N. Y., 1995. (Prepr. /School of ORIE, Cornell University).

3. **Mandelbrot B. B.** The Pareto-Levy law and the distribution of income // Internat. Econom. Rev. 1960. N 1. P. 79.

4. **Моисеев С. Н.** О нарушении центральной предельной теоремы для электронной концентрации слоя E_s из-за ее распределения по закону Коши // Геомагнетизм и аэрономия. 1998. **38**, № 6. С. 181.

5. **Моисеев С. Н.** Различение гипотез о функции распределения частоты экранирования спорадического слоя E ионосферы // Автометрия. 1997. № 3. С. 76.

6. **Боровков А. А.** Математическая статистика. М.: Наука, 1984.

7. **Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.** Справочник по теории вероятностей и математической статистике. М.: Наука, 1985.

8. **Хохлов Ю. С.** Псевдоустойчивые распределения и их области притяжения // Фундаментальная и прикладная математика. 1996. **2**, № 4. С. 1143.

Воронежский государственный университет,
E-mail: mois@rf.main.vsu.ru

Поступила в редакцию
15 марта 1999 г.