ДЭУ-метод проверки гипотезы пуассоновости выборки

1.Пусть $y = (y_1, \ldots, y_n)$, y_i – независимые случайные величины, $i = \overline{1,n}$. Далее будем проверять гипотезу

$$\Gamma_1 : y_i \stackrel{d}{=} POIS(a_i\theta_0),\ \theta_0 > 0,\ a_i > 0, \tag{1}$$

где a_i – известные величины, $i = \overline{1,n}$, против альтернативы

$$\Gamma_2 : y_i \stackrel{d}{\neq} POIS(a_i\theta_0),\ \theta_0 > 0,\ i = \overline{1,n}.$$

Проблема проверки Γ_1 против альтернативы Γ_2 при $n \to \infty$ детально рассмотрена в статье Большева, см. [1]. Приведем основные результаты этой работы.

Положим $a = \sum_{i=1}^{n} a_i$, $\quad \pi_i = \dfrac{a_i}{a}$, $\quad i = \overline{1,n}$, $\quad \overrightarrow{\pi} = (\pi_1, \ldots, \pi_n)^T$.

Как известно, семейство пуассоновских распределений при гипотезе Γ_1 обладает достаточной статистикой $S(y) = \sum_{i=1}^{n} y_i$, причем

$$S(y) \stackrel{d}{=} POIS(a\theta_0).$$

Непосредственно показывается, что

$$y|(S(y) = s) \stackrel{d}{=} MULTI(s; n, \overrightarrow{\pi}), \tag{2}$$

т.е. условное распределение случайной величины $y|\,(S(y) = s)$ "свободно т.е. не зависит от мешающего параметра θ_0.

Как показано в работе [1], соотношение (2) является характеристическим для гипотезы Γ_1 и напомним, что это означает, что если условное распределение случайной величины $y|(S(y) = s)$ является полиномиальным конкретного вида, то $y_1, \ldots, y_n$ – н.о.р. пуассоновские случайные величины вида (1). Характеристическое свойство (2) следует проверять. Тем самым для проверки гипотезы Γ_1 будет строится подобный критерий. Теория подобных критериев подробно рассматривается в монографии [2].

Для проверки гипотезы полиномиальности с известным вектором вероятностей $\overrightarrow{\pi_n}$ обычно рекомендуется использовать статистику хи-квадрат вида

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - s\pi_i)^2}{s\pi_i} = \sum_{i=1}^{n} \frac{y_i^2}{s\pi_i} - s. \tag{3}$$

Как отмечено в [1], при $n \to \infty$, $s \to \infty$ и $\min\limits_i s\pi_i \geq 10$

$$X^2 \stackrel{d}{\approx} \chi^2_{n-1}. \tag{4}$$

Для приложений заведомо интересен случай, когда $n \to \infty$, но $s \ll n$, т.е. когда $\theta_0 \ll 1$. Именно с такой ситуацией приходится сталкиваться при анализе данных о числе ДТП в автостраховании. Так при объеме портфеля $n = 1749$ общее число ДТП оказывается равным $s = 57$, или при $n = 2389$ число $s = 67$. Приведенные примеры соотношений между n и s являются характерными для практики автострахования. В этом случае условное распределение статистики $X^2|(S(y) = s)$ при $n \to \infty$ будет асимптотически нормальным со средним

$$E\{X^2|S = s\} = n - 1 \tag{5}$$

и дисперсией

$$D\{X^2|S=s\} = 2(n-1) + \frac{1}{s}\left(\sum_{i=1}^{n}\frac{1}{\pi_i} - n^2 - 2n + 2\right), \tag{6}$$

см. [1]. Таким образом, и при $s \ll n$ статистику X^2 можно с учетом ее условной асимптотической нормальности и соотношений (5) и (6) использовать для проверки гипотезы полиномиальности.

Если в качестве альтернативной рассматривается гипотеза типа $Ey_i < Dy_i$ (соответственно, $Ey_i > Dy_i$)), то критическую область следует брать в виде $\{y : X^2 > c\}$ (соответственно, $\{y : X^2 < c\}$). Если же подобного рода априорные представления об альтернативе отсутствуют, то критическая область выбирается двусторонней, т.е. $\{y : (X^2 < c_1) \cup (X^2 > c_2)\}$. Постоянные c , c_1 и c_2 выбираются на основе значений асимптотической ошибки первого рода, см. [1].

2. Рассмотренная выше теория не охватывает случай малых и умеренных значений n. Несомненный интерес представляет также использование для проверки гипотезы Γ_1 и других статистик, отличных от статистики (3), но чувствительных к отклонениям данных от гипотезы Γ_1. К таким статистикам относится, в частности,

$$T(y) = \max_{1\le i\le n} |\frac{y_i}{n} - \pi_i|. \tag{7}$$

Однако, при $n \ll \infty$ найти

$$\Psi(u) = P\{T(y) < u | S(y) = s\} \tag{8}$$

– условную функцию распределения статистики $T(y)$ на гиперповерхности $\{y : S(y) = s\}$ не удается, т.к. это весьма сложная многомерная переборная задача. Аналогичное утверждение справедливо и относительно функции распределения статистики X^2. Однако, оценить величину

$$\alpha_{\text{набл}}(s) = 1 - \Psi(T(y)) \tag{9}$$

– условный наблюденный уровень значимости методом достаточного эмпирического усреднения (ДЭУ-методом) относительно просто.

Действительно, пусть $z = (z_1, \dots, z_s)^T$, где z_i – независимые одинаково распределенные н.о.р. одномерные дискретные с.в., $z_i \in \mathsf{N}$, $i = \overline{1,s}$, с функцией распределения $F_0(u) = \sum_{i=1}^{n} \mathbb{I}(i < u)\pi_i$ где $\mathbb{I}(A)$ индикатор события A. Тогда, положив $y_i^* = \sum_{j=1}^{s} \mathbb{I}(i = z_j)$ для $i = \overline{1,n}$, $y^* = (y_1^*, \dots, y_n^*)$, получим , что $y^* \stackrel{d}{=} y|(S(y) = s)$. Случайную величину y^* назовчм *вариантом данных* y, т.к. безусловное распределение y^* совпадает при Γ_1 с распределением y, т.е. содержит одинаковое с y количество информации о гипотезе Γ_1. Напомним, что

$$\alpha_{\text{набл}}(s) = \mathbf{P}\{T(y^*) \ge T(y) \mid S(y) = s\}. \tag{10}$$

Если породить независимую выборку вариантов данных $\{y^*(1), \dots, y^*(B)\}$, объема B, где $y^*(j) \stackrel{d}{=} y \mid (S(y) = s)$ для $j = \overline{1,B}$, то при гипотезе Γ_1 величина

$$\hat{\alpha}_{\text{набл}}(s) = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}(T(y^*(j)) \ge T(y)) \tag{11}$$

является состоятельной несмещенной оценкой для $\alpha_{\text{набл}}(s)$, [3]. Более того, поскольку $y^*(j)$ н.о.р. случайные величины, то из (11) вытекает, что

$$B\hat{\alpha}_{\text{набл}}(s) \stackrel{d}{=} BIN(B; \alpha_{\text{набл}}(s)), \tag{12}$$

где $BIN(B,q)$ – биномиальная случайная величина, т.е. число успехов в B испытаниях Бернулли с вероятностью успеха в отдельном испытании равным q, $0 < q < 1$. Заметим,

что значение B можно выбрать сколь угодно большим: его выбор ограничен лишь компьютерным временем вычислений и мощностью датчика псевдо случайных равномерно распределенных чисел.

Но ни первая, ни вторая причины реально не являются препятствием для оценки $\alpha_{\text{набл}}(s)$. С одной стороны, существуют датчики случайных чисел практически неограниченной мощности, см. [4]. С другой стороны, проблема оценки нахождения $\hat{\alpha}_{\text{набл}}(s)$ идеально структурируется для проведения вычислений ее на мощных многопроцессорных компьютерах.

Что касается выбора величины B, то ее можно осуществить или на основе двустадийной процедуры построения γ-доверительного интервала заранее заданной ширины для $\alpha_{\text{набл}}(s)$, см. [5], или на основе очевидного асимптотического равенства, следующего из теоремы Муавра-Лапласа,

$$\mathbf{P}\left\{\sqrt{B}\left|\frac{\hat{\alpha}_{\text{набл}}(s)-\alpha_{\text{набл}}(s)}{\sqrt{\alpha_{\text{набл}}(s)(1-\alpha_{\text{набл}}(s))}}\right|\le u_{\frac{1+\gamma}{2}}\right\}=\gamma+O\left(\frac{1}{\sqrt{B}}\right), \tag{13}$$

где $\Phi(u_\varepsilon)=\varepsilon$, $0<\varepsilon<1$, $0<\gamma<1$, $\Phi(u)=\int\limits_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz$, см [6].

Из последнего равенства следует, что ошибка в оценивании $\alpha_{\text{набл}}(s)$ с вероятностью γ не превосходит $\Delta=\dfrac{u_{\frac{1+\gamma}{2}}}{2\sqrt{B}}$. Задав Δ и γ получим, что

$$B\approx\left(\frac{u_{\frac{1+\gamma}{2}}}{2\Delta}\right)^2. \tag{14}$$

Заметим, что изложенная методология не "привязана"к конкретной статистике критерия. Ее можно реализовать для любой статистики критерия, не являющейся S-измеримой.

Для построения критерия размера (приблизительно) α_1, в качестве решающего правила следует принять следующее

$$\delta(y)=\begin{cases}\text{принимать } \Gamma_1 \text{ при } \hat{\alpha}_{\text{набл}}(s)\ge\alpha_1,\\ \text{отвергать } \Gamma_1 \text{ при } \hat{\alpha}_{\text{набл}}(s)<\alpha_1.\end{cases}$$

Для ряда альтернатив (в том числе и сложных) возможно также оценить и мощность принятого критерия. Необходимая для этого ДЭУ-методология изложена в [3] и [7].

3.Рассмотренный в п.2 метод построения критериев естественным образом переносится на проблему решения проверки гипотез об однородности и стохастической упорядоченности потоков ДТП. Именно с этими проблемами приходится сталкиваться при тарификации различных контингентов страхователей.

4.Пуассоновский поток является базовой стохастической моделью в ряде исследований по теории страхования. Принятые гипотезы пуассоновости (или смешанной пуассоновасти) во многом облегчает математическое моделирование этого важнейшего компонента тарификационной системы страхования. Однако, приведенный анализ статистических данных, в частности, российской системы ОСАГО, свидетельствует в пользу отвержения как гипотезы пуассоновости, так и ряда широко используемых смешанных пуассоновских семейств распределений.

Список литературы

[1] Большев Л.Н. *Избранные труды. Теория вероятностей и математическая статистика.* Наука, М., 1987.

[2] Леман Э. *Проверка статистических гипотез.* Наука, М., 1979.

[3] Chepurin E.V. On analytic-computer methods of statistical inferences of small data samples. *Proc. of the Intern. Conf. Prob. Analysis of rare events, Riga: Aviation University*, pages 180–194, 1999.

[4] Deng L.V., Lin D. Random numder generation forthe new century. *The American Statistician*, 54,2:180–194, May 2000.

[5] Закс Ш. *Теория статистических выводов.* Мир, М., 1975.

[6] Гнеденко Б.В. *Курс теории вероятностей.* УРСС, М., 2001.

[7] Андронов А.М., Гаджиев А.Г., Чепурин Е.В. *Тезисы PCI*, 2006.