МАТЕМАТИКА

УДК 517.948

О НЕКОТОРЫХ СВОЙСТВАХ ОЦЕНКИ КОЭФФИЦИЕНТА ВАРИАЦИИ ДЛИН ПОСЛЕДОВАТЕЛЬНЫХ ИНТЕРВАЛОВ ТОЧЕЧНОГО ПРОЦЕССА

Заляпин И.В., Писаренко В.Ф.

zaliapin@mitp.ru

Международный Институт Теории Прогноза Землетрясений и Математической Геофизики РАН, Москва, Россия

Статья поступила в редакцию 10 января 1999 года

Коэффициентом вариации R_ξ случайной величины ξ называют отношение ее стандартного отклонения σ_ξ к матожиданию m_ξ:

$$R_\xi = \frac{\sigma_\xi}{m_\xi}. \tag{1}$$

Статистику R_ξ обычно рассматривают применительно к положительным случайным величинам. Легко проверить, что она принимает значение 1 в случае экспоненциального распределения. Это свойство делает (1) чрезвычайно полезной при рассмотрении точечных процессов. В этом случае в качестве случайных величин выступают интервалы между последовательными событиями t_i точечного процесса $N(dt)$.:

$$\tau_i = t_i - t_{i-1}, \tag{2}$$

Известно, что если процесс $N(dt)$ – стационарен, то τ_i из (2) одинаково распределены для любого i. Таким образом, в этом случае использование статистики R_τ оказывается законным.

Значение R_τ является характеристикой группируемости точечного процесса. В случае однородного Пуассоновского процесса $R_\tau = 1$, что указывает на полную случайность или отсутствие группируемости. В случае наличия группируемости событий R_τ возрастает, принимая большие значения для более кластеризованного процесса. При наличии тенденции к периодичности величина R_τ становится меньше 1 и, наконец, $R_\tau = 0$ для полностью периодического процесса – процесса, события которого отстоят друг от друга на детерминированный интервал Δ.

Статистика R_ξ широко используется на практике для определения степени кластеризации событий точечного процесса и на основании ее значений нередко делаются качественные выводы о природе процесса. В связи с этим необходимо уметь определять статистическую значимость результатов, полученных по выборке конечного объема.

Рассмотрим задачу проверки гипотезы о распределении длин интервалов между событиями. Нулевая гипотеза состоит в том, что это распределение экспоненциально, то есть рассматриваемый процесс – Пуассоновский:

$$H_0: \quad \tau_i \sim f(x) = \begin{cases} ae^{-\alpha x}; & x \geq 0 \\ 0; & x < 0 \end{cases}$$

Коэффициент вариации R_ξ может использоваться для построения критерия значимости при проверке такой гипотезы.

Нам будет удобнее работать не с самой статистикой R, а с ее квадратом

$$R_\xi^2 = \frac{\sigma_\xi^2}{m_\xi^2}. \tag{3}$$

Наблюдая процесс на конечном промежутке времени, мы получаем в распоряжение конечное количество интервалов $\tau_i,\ i = 1,\dots,n$ между событиями. По ним могут быть вычислены эмпирические дисперсия $\hat\sigma_\tau$ и матожидание $\hat m_\tau$:

$$\hat\sigma_\tau = \frac{1}{n}\sum_i(\tau_i - \hat m_\tau)^2; \quad \hat m_\tau = \frac{1}{n}\sum_i \tau_i.$$

Зная эти величины, мы можем вычислить значение оценки $\hat R_\tau^2$ величины R_τ^2 по формуле (3):

$$\hat R_\tau^2 = \frac{\hat\sigma_\tau^2}{\hat m_\tau^2}.$$

Справедливо следующее представление:

$$\hat R_\tau^2 = \sum_{i=1}^n \frac{n\tau_i^2}{(S_n)^2} - 1 = \sum_{i=1}^n n w_i^2 - 1, \tag{4}$$

где

$$S_n = \sum_{i=1}^n \tau_i, \quad w_i = \frac{\tau_i}{S_n}.$$

Таким образом, значение статистики выражается через сумму одинаково распределенных зависимых слагаемых. В связи с этим полезным оказывается следующее

Утверждение I

Пусть $\xi_i,\ i = 1,\dots,n$ – независимые экспоненциально распределенные случайные величины, $S_n = \sum_1^n \xi_i$.

Тогда величины $w_i^2 = \frac{\xi_i^2}{(S_n)^2}$ одинаково распределены, имеют функцию распределения $F(x)$:

$$F(x) = 1 - (1 - \sqrt{x})^{n-1}, \quad x \in [0,1],$$

и плотность $f(x)$:

$$f(x) = \frac{n-1}{2} \cdot \frac{(1-\sqrt{x})^{n-2}}{\sqrt{x}}.$$

Математическое ожидание величин w_i из (4) в случае экспоненциально распределенных интервалов τ_i может быть непосредственно вычислено:

$$Ew_i^2 = \int x f(x)dx = \frac{n-1}{2}\int_0^1 \sqrt{x}(1-\sqrt{x})^{n-2}dx =$$

$$= (n-1)\cdot\frac{\Gamma(3)\Gamma(n-1)}{\Gamma(n+2)} = 2\cdot\frac{1}{n(n+1)}.$$

Отсюда имеем для статистики $\hat R_\tau^2$:

$$E\hat R_\tau^2 = E\left\{\sum n w_i^2 - 1\right\} = n^2\cdot\frac{2}{n(n+1)} - 1 = 1 - \frac{2}{n+1}. \tag{5}$$

Если для дисперсии в (3) использовать несмещенную оценку

$$\hat{\sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}\xi_i,$$

то аналогично получим

$$E\hat{R}_\tau^2 = 1 - \frac{1}{n+1}. \tag{6}$$

Из (5), (6) видим, что оценка $\hat{R}_\tau^2$ является смещенной. Ее среднее значение оказывается меньше 1, что надо помнить особенно при малых значениях n.

Отметим, что при нулевой гипотезе для величин w_i^2 в явном виде можно получить моменты любого порядка:

$$E\left[w_i^2\right]^k = \frac{(2k)!}{n(n+1)\dots(n+2k)}.$$

Зависимость величин w_i^2 не позволяет вычислить дисперсию оценки $\hat{R}_\tau^2$ аналогичным образом. Однако, моменты высших порядков для статистики $\hat{R}_\tau^2$ можно получить, используя теорему Дирихле (Kendall,[6]). Опуская промежуточные вычисления, приведем лишь окончательное выражение для дисперсии при нулевой гипотезе:

$$D\hat{R}_\tau^2 = 4\frac{n^2(n-1)}{(n+1)^2(n+2)(n+3)}. \tag{7}$$

Обратимся теперь к асимптотическому поведению оценки $\hat{R}_\tau^2$ при $n \to \infty$.

Утверждение II

Пусть $\tau_i, \quad i = 1, \dots, n$ *– независимые экспоненциально распределенные случайные величины.*

Тогда для статистики $\hat{R}_\tau^2$ *верно следующее асимптотическое представление:*

$$\hat{R}_\tau^2 = 1 + \frac{1}{\sqrt{n}}\left(2\sqrt{5}\eta_2 - 4\eta_1\right) + o\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty,$$

где (η_1, η_2) *– нормально распределенные случайные величины с нулевыми средними и матрицей ковариаций*

$$\Sigma_{\eta_1\eta_2} = \begin{pmatrix} 1 & 2/\sqrt{5} \\ 2/\sqrt{5} & 1 \end{pmatrix}.$$

Из *Утверждения II* непосредственно вытекает, что при нулевой гипотезе асимптотическая дисперсия оценки $\hat{R}_\tau^2$ равняется $4/n$, что согласуется с точным выражением (7).

Для построения критериев более естественно использовать статистику $\hat{R}_\tau$, а не ее квадрат, который был введен исключительно из технических соображений. Извлекая корень из $\hat{R}_\tau^2$ и пользуясь *Утверждением II*, получаем, что $\hat{R}_\tau$ может быть представлена при $n \to \infty$ как

$$\hat{R}_\tau = 1 + \frac{1}{\sqrt{n}}\eta + o\left(\frac{1}{\sqrt{n}}\right),$$

где η – стандартная нормальная величина.

Обратим внимание, что использование статистики $\hat{R}_\tau$ при проверке гипотезы H_0 ничем не ограничивает круг альтернативных гипотез H_A. Можно надеяться, что, рассматривая статистики, пригодные лишь для некоторого специального класса распределений, мы повысим надежность критерия.

В качестве такого класса может выступать, например, 2-параметрическое семейство распределений Вейбула:

$$f(x,\alpha,\lambda)=\begin{cases}\frac{\alpha}{x}\left(\frac{x}{\lambda}\right)^{\alpha} exp\left\{-\left(\frac{x}{\lambda}\right)^{\alpha}\right\}, & x\geq 0\\ 0, & x<0\end{cases} \tag{8}$$

Семейство (8) достаточно богато, чтобы описывать широкий спектр реальных процессов (см. например Kagan [2]). При этом интересующий нас феномен – кластеризация – в данном семействе определяется одним параметром α. Нулевая гипотеза, таким образом, может быть параметрически задана как

$$H_0 \ : \ \alpha=1.$$

Утверждение III

Статистика

$$\hat{\alpha}_n=\frac{2+\frac{1}{n}\sum_1^n \ln w_i-\frac{1}{n}\sum_1^n w_i\ln w_i+\frac{1}{n}\sum_1^n w_i\ln^2 w_i}{1+\frac{1}{n}\sum_1^n w_i\ln^2 w_i}, \tag{9}$$

где

$$w_i=\frac{x_i}{\bar{x}_n},\quad \bar{x}_n=\frac{1}{n}\sum_1^n x_i$$

при нулевой гипотезе H_0 допускает следующее асимптотическое представление:

$$\hat{\alpha}_n=1+\frac{1}{\sqrt{n}}\eta+o\left(\frac{1}{\sqrt{n}}\right),$$

где η – нормальная случайная величина с нулевым средним и дисперсией

$$\sigma^2=\frac{1}{(1-c)^2+\pi^2/6}=0.5483...,$$

$c=0.57721566490...$ *– константа Эйлера.*

Пользуясь асимптотическим выражением для моментов распределения Вейбула (Cramer [4]), можно показать, что производная по параметру α от асимптотического матожидания статистики $\hat{\alpha}_n$ в точке $\alpha=1$ приближенно равна 1.084, в то время как соответствующая производная коэффициента вариации равняется -1. Поэтому критерий проверки гипотезы H_0, основанный на статистике $\hat{\alpha}_n$ будет более мощным в окрестности точки $\alpha=1$, чем критерий, основанный на коэффициенте вариации.

Зная асимптотическое распределение статистики при нулевой гипотезе H_0, мы можем строить критерий значимости. Однако такой критерий бует надежен лишь при достаточно больших обьемах выборки, что не всегда выполняется на практике. Поэтому нам хотелось бы уметь строить подходящие критерии для любых значений n.

Ниже мы опишем способ вычисления распределения оценки квадрата коэффициента вариации $\hat{R}^2_\tau$ при нулевой гипотезе для произвольного n. Отметим, что точные выражения для этого распределения, описанные в литературе, известны лишь для $n<5$ (см. например Kendall [6]) и при непосредственной аналитической реализации описываемого ниже способа мы сталкиваемся с существенными трудностями уже при $n=3$. Однако, он может быть использован в качестве основы для численного оценивания.

Известно (см.Feller, [5],т.II,стр.95), что если $\tau_i,\ i=1,\ldots,n$ распределены экпоненциально и независимы, то строка $(w_1,w_2,\ldots,w_{n-1})$ (где обозначения взяты из (4)) распределена равномерно в области

$$\Omega=\left\{(w_1,w_2,\ldots,w_{n-1})\ \vdots\ w_1+\ldots+w_{n-1}\leq 1,\ w_i\geq 0\right\}.$$

Отсюда непосредственно вытекает, что при условии $w_n = w_0$ строка $(w_1, w_2, \ldots, w_{n-1})$ оказывается вырожденной, то есть принимает значения в $n-2$ мерном пространстве, причем ее распределение в области

$$\Omega_1 = \left\{ (w_1, w_2, \ldots, w_{n-1}) \vdots\, w_1 + \ldots + w_{n-1} = 1 - w_0,\ w_i \geq 0 \right\}$$

равномерно.

С учетом этого мы заключаем, что вероятность

$$P(r^2, w_0) \equiv \Pr\left\{ w_1^2 + w_2^2 + \ldots + w_{n-1}^2 \leq r^2 | w_1 + \ldots + w_{n-1} = 1 - w_0 \right\}$$

равна отношению обьемов областей Ω_2 и Ω_1:

$$P(r^2, w_0) = \frac{V(\Omega_2)}{V(\Omega_1)},$$

где

$$\Omega_2 = \Omega_1 \cap \left\{ w_i \vdots\, \sum_1^{n-1} w_i^2 \leq r^2,\ w_i \geq 0 \right\}.$$

Нас интересует распределение $F_\eta(x)$ величины $\eta = \sum_1^n w_i^2$. Полагая w_n фиксированным, мы можем записать

$$F_\eta(x|w_n) = \Pr\{\eta \leq x\} = \Pr\left\{ \sum_1^{n-1} w_i^2 \leq x - w_n^2 \right\} = P(x - w_n^2, w_n). \qquad (10)$$

Воспользовавшись *Утверждением I*, усредняем (10) по величине w_n:

$$F_\eta(x) = \int_0^1 w(u) P(x - u^2, u) du = \int_0^1 (n-1)(1-u)^{n-2} P(x - u^2, u) du,$$

и, дифференцируя по x, получаем плотность $f_\eta(x)$:

$$f_\eta(x) = \frac{dF_\eta(x)}{dx} = \int_0^1 w(u) p(x - u^2, u) du,$$

где $p(x,u) = \frac{dP(x,u)}{dx}$.

Обратим внимание, что $\hat{\alpha}_n$, рассматриваемая как функция от α, линейна лишь в некоторой окрестности точки $\alpha = 1$ с точностью до ϵ^2 и как оценка параметра α в общем случае применяться не может.

Работа выполнена при поддержке грантов РФФИ 97-05-64583 и МНТЦ 415-96.

Литература

[1] Ризниченко Ю.В. Проблемы сейсмологии. Избр.тр.//М.:Наука,1985.-408 с.

[2] Kagan,Y.Y. Jeckson,D.D. Long-term earthquake clustering.// Jeophys J.Int.,1991, vol.104, pp.117-133

[3] M.S.Bartlett, The Statistical Analysis of Spatial Pattern// Adv.Appl.Prob. 6, 1974, pp.336-358.

[4] Крамер Г. Математические методы статистики.//М.:МИР,1975.-648 с.

[5] В.Феллер, Введение в теорию вероятностей и ее приложения.//Т.2, М.:Мир, 1984.-746 с.

[6] М.Кендалл, П.Моран, Геометрические вероятности.//М.: Наука, 1972.-192 с.