СТАТИСТИЧЕСКИЙ АНАЛИЗ ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ПРЕОБРАЗОВАННЫХ НА ОСНОВЕ ЭМПИРИЧЕСКОЙ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Е. Е. Жук

Белорусский государственный университет
Минск, Беларусь
E-mail: zhukee@mail.ru

Рассматривается проблема улучшения статистических свойств псевдослучайных последовательностей. Предлагается специальное преобразование значений последовательности, основанное на эмпирической функции распределения, которое обеспечивает прохождение основных статистических тестов на стандартное равномерное распределение.

Ключевые слова: базовая случайная величина, псевдослучайная последовательность, эмпирическая функция распределения, статистические тесты.

ВВЕДЕНИЕ. ПРЕОБРАЗОВАНИЕ ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И ПРОБЛЕМА ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ

Как известно [1, 2], имитационное статистическое моделирование случайной величины ξ с заданной функцией распределения вероятностей $F_\xi(z) = \mathrm{P}\{\xi \le z\}$, $z \in R$, сводится к моделированию так называемой базовой случайной величины (БСВ) α, имеющей стандартное равномерное распределение на отрезке [0,1] с функцией распределения: $F_\alpha(z) = z$, $z \in [0,1]$, математическим ожиданием $E\{\alpha\} = 1/2$ и дисперсией $D\{\alpha\} = 1/12$.

Например, если существует $F_\xi^{-1}(\cdot)$, то случайную величину ξ можно смоделировать методом обратной функции [2]: $\xi = F_\xi^{-1}(\alpha)$.

Наиболее распространенным, в силу удобства использования, методом моделирования БСВ является программный или алгоритмический [1, 2]. Однако он обладает существенным недостатком [1, 2]: генерируемые последовательности значений БСВ порождаются детерминированным алгоритмом, являются "псевдослучайными", зависимыми (коррелированными) между собой, а их статистические свойства могут существенно отличаться от свойств БСВ с функцией распределения $F_\alpha(\cdot)$.

Для улучшения статистических свойств датчиков БСВ здесь предлагается подвергать порождаемые ими псевдослучайные последовательности функциональному преобразованию на основе эмпирической функции распределения (ЭФР) [2, 3].

Пусть ξ – некоторая непрерывная случайная величина с функцией распределения $F_\xi(\cdot)$, имеющей обратную $F_\xi^{-1}(\cdot)$, тогда легко установить [3], что случайная величина $\alpha = F_\xi(\xi)$ имеет функцию распределения $F_\alpha(\cdot)$ и является БСВ. Данный факт позволяет не только улучшать статистические свойства имеющегося датчика БСВ, но и получать БСВ из непрерывной случайной величины с произвольным законом распределения вероятностей.

Пусть имеется некоторый датчик БСВ. В результате n-кратного обращения к нему получаем псевдослучайную последовательность $X = \{x_1, \ldots, x_n\}$ объема n. По ней строим ЭФР [2, 3]:

$$\hat{F}_x(z) = \frac{1}{n}\sum_{i=1}^{n} I(z - x_i),\ z \in R; \qquad I(w) = \begin{cases} 1, & w \geq 0; \\ 0, & w < 0, \end{cases} \tag{1}$$

в качестве статистической оценки функции распределения $F_x(z) = \mathrm{P}\{x_i \leq z\}$, $z \in R$, которая, вообще говоря, может существенно отличаться от $F_\alpha(\cdot)$.

Затем элементы исходной последовательности $X = \{x_1, \ldots, x_n\}$ подвергаем функциональному преобразованию на основе ЭФР из (1):

$$y_l = \hat{F}_x(x_l),\ l = \overline{1,n}, \tag{2}$$

и получаем реализацию $Y = \{y_1, \ldots, y_n\}$ "улучшенной" БСВ.

Проблема заключается в том, что преобразование (1), (2) вместо неизвестной функции распределения $F_x(\cdot)$ использует ее статистическую оценку $\hat{F}_x(\cdot)$, и необходимо исследовать эффективность такой замены.

Будем предполагать, что выполняются следующие условия на элементы исходной последовательности: $X = \{x_1, \ldots, x_n\}$ образована независимыми в совокупности (У1), непрерывными случайными величинами (У2), одинаково распределенными с функцией распределения $F_x(\cdot)$ (У3).

Очевидно, что в условиях У1-У3 после преобразования (1), (2) элементы новой последовательности $Y = \{y_1, \ldots, y_n\}$ будут иметь одну и ту же функцию распределения $F_y(z) = \mathrm{P}\{y_l \leq z\}$, $z \in R$, но являться, вообще говоря, зависимыми между собой.

Исследуем эффективность преобразования (1), (2) на основе стандартного набора тестов Д. Кнута [1, 2]: критериев согласия с функцией распределения $F_\alpha(\cdot)$ (критерии Колмогорова и Пирсона), тестов "совпадение моментов" и "ковариация". В каждом из этих тестов относительно последовательности $Y = \{y_1, \ldots, y_n\}$ проверяется какая-либо гипотеза H_0 против альтернативы общего вида $H_1 = \overline{H_0}$ (отрицание H_0). Сами тесты строятся при наперед заданном малом уровне значимости $\varepsilon = P\{\textit{принять}\ H_1 \mid H_0\} \in (0,1)$ и имеют следующий общий вид:

$$\textit{принимается} \begin{cases} H_0, \textit{если}\ P(Y,n) \geq \varepsilon; \\ H_1, \textit{если}\ P(Y,n) < \varepsilon, \end{cases} \tag{3}$$

где $P = P(Y,n) \in [0,1]$ – так называемое P-значение [3].

ПРОВЕРКА ГИПОТЕЗ СОГЛАСИЯ СО СТАНДАРТНЫМ РАВНОМЕРНЫМ РАСПРЕДЕЛЕНИЕМ

Проверим относительно преобразованной последовательности $Y=\{y_1,\ldots,y_n\}$ гипотезы согласия со стандартным равномерным распределением:

$$\begin{aligned} H_0:&\quad F_y(\cdot)\equiv F_\alpha(\cdot);\\ H_1:&\quad \exists z\in R: F_y(z)\neq F_\alpha(z). \end{aligned} \tag{4}$$

Широко известный критерий Колмогорова для проверки гипотез (4) имеет вид (3) и основан на следующей статистике P-значения [2, 3] ($n\to+\infty$):

$$P=P(Y,n)=1-K\left(\sqrt{n}D_n\right),\qquad D_n=\sup_{z\in R}|\hat{F}_y(z)-F_\alpha(z)|, \tag{5}$$

где D_n – так называемое расстояние Колмогорова между гипотетической функцией распределения $F_\alpha(\cdot)$ и ЭФР, вычисленной по $Y=\{y_1,\ldots,y_n\}$:

$$\hat{F}_y(z)=\frac{1}{n}\sum_{l=1}^{n}I(z-y_l),\ z\in R, \tag{6}$$

а $K(z)=1-2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2z^2}$, $z\geq 0$, – функция распределения Колмогорова.

Теорема 1. Пусть выполнено условие У2, тогда расстояние Колмогорова $D_n=1/n$, а соответствующее P-значение из (5) имеет вид:

$$P(Y,n)=P(n)=1-K\left(1/\sqrt{n}\right). \tag{7}$$

Доказательство. В силу условия У2, элементы $\{x_l\}_{l=1}^n$ исходной последовательности имеют абсолютно непрерывные распределения вероятностей, поэтому $P\{x_t=x_l\}=0$, $t\neq l$, и после преобразования (1), (2) последовательность $Y=\{y_1,\ldots,y_n\}$ с точностью до порядка следования будет содержать значения $\{j/n\}_{j=1}^n$. Но тогда для ЭФР из (6) имеем:

$$\hat{F}_y(z)=\frac{1}{n}\sum_{j=1}^{n}I\left(z-\frac{j}{n}\right)=\left\{0,\ z\in\left(-\infty,\frac{1}{n}\right);\ \frac{j}{n},\ z\in\left[\frac{j}{n},\frac{j+1}{n}\right)\ (j=\overline{1,n-1});\ 1,\ z\in[1,+\infty)\right\},$$

и, с учетом вида гипотетической функции распределения: $F_\alpha(z)=z$, $z\in[0,1]$, и $F_\alpha(z)=0$, $z\notin[0,1]$, для расстояния Колмогорова из (5) получаем: $D_n=1/n$, что приводит к P-значению из (7) и доказывает теорему.

Из (7) видно, что $P(n)\to 1$, $n\to+\infty$, и увеличением объема реализации n для заданного уровня значимости $\varepsilon\in(0,1)$ всегда можно добиться того, что $P(n)\geq\varepsilon$, и гипотеза H_0 из (4) принимается.

Рассмотрим теперь χ^2-критерий Пирсона для проверки гипотез (4). Этот тест также имеет вид (3), но его P-значение [2, 3] ($n\to+\infty$):

$$P=P(Y,n)=1-F_{\chi^2_{k-1}}(\chi^2),\qquad \chi^2=\sum_{l=1}^{k}\frac{(n_l-n/k)^2}{n/k},\qquad n_l=\sum_{t=1}^{n}\begin{cases}1, & y_t\in\Gamma_l;\\ 0, & y_t\notin\Gamma_l,\end{cases} \tag{8}$$

где $k \geq 2$ – число ячеек, на которые разбивается отрезок $[0,1]$: $\Gamma_i = \big((i-1)/k,\, i/k\big]$, $i = \overline{1,k}$, а $F_{\chi^2_{k-1}}(\cdot)$ – функция χ^2-распределения с $k-1$ степенью свободы.

Теорема 2. Пусть выполнено условие У2, а объем реализации n кратен числу ячеек k, тогда χ^2-статистика из (8) равна нулю: $\chi^2 = 0$, а соответствующее P-значение: $P = 1$. В случае, когда n не кратно k, для P-значения из (8) выполняется:

$$P(Y,n) \geq P_+(n,k), \qquad P_+(n,k) = 1 - F_{\chi^2_{k-1}}\left(k^2/n\right) \tag{9}$$

Из результата теоремы 2 следует, что если объем реализации n кратен числу ячеек k, то гипотеза H_0 из (4) критерием Пирсона принимается при любом уровне значимости $\varepsilon \in (0,1)$. Но даже если n не делится на k без остатка, то для $P_+(n,k)$ из (9) выполняется: $P_+(n,k) \to 1$, $n \to +\infty$, и увеличением n для любого наперед заданного уровня значимости $\varepsilon \in (0,1)$ всегда можно добиться принятия гипотезы H_0: $P(Y,n) \geq P_+(n,k) \geq \varepsilon$.

ТЕСТЫ «СОВПАДЕНИЕ МОМЕНТОВ» И «КОВАРИАЦИЯ»

В условиях У1-У3 введем следующие обозначения для моментов: $\mu_y = E\{y_t\}$, $\sigma_y^2 = E\{(y_t - \mu_y)^2\}$ – математическое ожидание и дисперсия элементов преобразованной последовательности $Y = \{y_1, \ldots, y_n\}$.

Согласно критериям "совпадение моментов" [2], гипотезы относительно математического ожидания:

$$H_0: \ \mu_y = 1/2; \qquad H_1: \ \mu_y \neq 1/2, \tag{10}$$

проверяются тестом (3) с P-значением [2, 3] ($n \to +\infty$):

$$P = P(Y,n) = 2\left(1 - \Phi\left(\sqrt{12n}\left|\bar{y} - 1/2\right|\right)\right), \tag{11}$$

где $\bar{y} = \dfrac{1}{n}\sum_{t=1}^{n} y_t$ – оценка для μ_y по Y объема n, а $\Phi(z) = \dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{\omega^2}{2}}\, d\omega$ – функция распределения вероятностей стандартного нормального закона.

А гипотезы относительно дисперсии:

$$H_0: \ \sigma_y^2 = 1/12; \qquad H_1: \ \sigma_y^2 \neq 1/12, \tag{12}$$

проверяются в рамках теста "ковариация" (см. далее) тестом (3) с P-значением [2, 3] ($n \to +\infty$):

$$P = P(Y,n) = 2\left(1 - \Phi\left(6\sqrt{2(n-1)}\left|s_y^2 - 1/12\right|\right)\right), \tag{13}$$

где $s_y^2 = \dfrac{1}{n-1}\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2$ – оценка дисперсии σ_y^2 по реализации Y.

Как уже отмечалось выше, в силу условия У2, последовательность $Y = \{y_1, \ldots, y_n\}$ образована из элементов $\{j/n\}_{j=1}^{n}$, и легко получить:

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t = \frac{1}{n^2}\sum_{j=1}^{n} j = \frac{n(n+1)}{2n^2} = \frac{1}{2} + \frac{1}{2n}; \quad s_y^2 = \frac{1}{n-1}\sum_{t=1}^{n} y_t^2 - \frac{n}{n-1}\left(\bar{y}\right)^2 = \frac{1}{12} + \frac{1}{12n},$$

что при подстановке в (11), (13) приводит к следующим P-значениям для проверки гипотез относительно математического ожидания (10) и дисперсии (12) соответственно:

$$P(Y,n)=P(n)=2\left(1-\Phi\left(\sqrt{\frac{3}{n}}\right)\right);\qquad P(Y,n)=P(n)=2\left(1-\Phi\left(\sqrt{\frac{n-1}{2n^2}}\right)\right). \tag{14}$$

И в обоих случаях: $P(n)\to 1$, $n\to+\infty$.

Отметим также, что здесь и в теоремах 1 и 2 вместо выполнения условия У2 достаточно потребовать, чтобы исходная последовательность $X=\{x_1,\ldots,x_n\}$ не содержала совпадающих значений: $x_t\neq x_l$, $t\neq l$, $t,l=\overline{1,n}$.

Однако даже если элементы $\{x_t\}_{t=1}^n$ исходной последовательности удовлетворяют условиям У1-У3 с некоторой функцией распределения $F_x(\cdot)$, то элементы $\{y_t\}_{t=1}^n$ преобразованной последовательности, по построению (1), (2), будут зависимыми. Оценим эту зависимость, вычислив ковариации: $Cov\{y_t,y_l\}=E\{(y_t-E\{y_t\})(y_l-E\{y_l\})\}$, $t,l=\overline{1,n}$.

Теорема 3. Пусть исходная последовательность $X=\{x_1,\ldots,x_n\}$ удовлетворяет условиям У1-У3, тогда элементы $\{y_t\}_{t=1}^n$ преобразованной, согласно (1), (2), последовательности $Y=\{y_1,\ldots,y_n\}$ имеют следующие математическое ожидание, дисперсию и ковариации ($t\neq l$, $t,l=\overline{1,n}$):

$$E\{y_t\}=\frac{1}{2}+\frac{1}{2n},\qquad D\{y_t\}=\frac{1}{12}-\frac{1}{12n^2},\qquad Cov\{y_t,y_l\}=\frac{1}{12n}-\frac{1}{12n^2}. \tag{15}$$

Из теоремы 3 видно ($t,l=\overline{1,n}$): $Cov\{y_t,y_l\}\to\{0,\ t\neq l;\ 1/12,\ t=l\}$, $n\to+\infty$, что соответствует последовательности БСВ.

На практике для проверки гипотезы о некоррелированности $\{y_t\}_{t=1}^n$:

$$H_0:\ \ Cov\{y_t,y_l\}=1/12,\ t=l;\ \ Cov\{y_t,y_l\}=0,\ t\neq l\ \ (t,l=\overline{1,n}), \tag{16}$$

против альтернативы $H_1=\overline{H_0}$, используется тест "ковариация" [2, 3], согласно которому гипотеза H_0 принимается при заданном уровне значимости $\varepsilon\in(0,1)$, если одновременно выполняются следующие неравенства:

$$P_0(n)=P(n)=2\left(1-\Phi\left(\sqrt{\frac{n-1}{2n^2}}\right)\right)\geq\varepsilon;\ \ P_\tau(Y,n)=2\left(1-\Phi\left(12\sqrt{n-1}\,|c_\tau|\right)\right)\geq\varepsilon,\ \ \tau=\overline{1,\tau_+}, \tag{17}$$

где $c_\tau=\dfrac{1}{n-\tau-1}\sum\limits_{j=1}^{n-\tau}\left(y_j-\bar{y}\right)\left(y_{j+\tau}-\bar{y}\right)$ – оценки ковариаций между y_t и y_l, $\tau=|t-l|$ ($\tau_+/n\to 0$, $\tau_+\to+\infty$, $n\to+\infty$), по реализации Y объема n (при $\tau=0$: $c_0=s_y^2$ – оценка дисперсии из (13), а $P_0(n)=P(n)$ – соответствующее, второе, P-значение из (14)).

При $n\to+\infty$: $P_0(n)=P(n)\to 1$, а, в силу свойства состоятельности [3], оценки $c_\tau=c_{|t-l|}$ с ростом n приближаются к своим истинным значениям $Cov\{y_t,y_l\}$ из (15) ($t\neq l$), и из (17) видно, что P-значения $P_\tau=P_\tau(Y,n)$, $\tau=\overline{1,\tau_+}$, с ростом n также приближаются к 1, и гипотеза H_0 из (16) критерием (17) будет принята.

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ

Согласно полученным теоретическим результатам, в случае отсутствия в исходной последовательности совпадающих значений, при большом объеме n ($n \to +\infty$) полученная после преобразования (1), (2) последовательность будет "заведомо проходить" тесты Колмогорова, Пирсона и "совпадение моментов": увеличением n для любого заданного уровня значимости $\varepsilon \in (0,1)$ всегда можно добиться того, что соответствующее P-значение $P(n) \geq \varepsilon$ ($P(n) \to 1$, $n \to +\infty$). Для теста "ковариация" при условии, что элементы исходной последовательности независимы и одинаково распределены, данное свойство также имеет место, но не удается строго получить аналитическое соотношение для определения достаточного объема реализации n.

С целью экспериментального подтверждения полученных результатов преобразованию (1), (2) подвергались последовательности, порожденные различными датчиками БСВ. Результаты приведены в таблице, где для каждого случая в первой строке приводятся результаты для исходной последовательности, а для преобразованной – во второй (символ "z" в обозначениях вспомогательных величин для исходной последовательности понимается как "x", а для преобразованной – как "y").

Таблица

Результаты тестирования

Тип датчика БСВ	Объем, n	Критерий Пирсона		Критерий Колмогорова,	Величины		
		k	P	P	$\lvert\bar{z}-1/2\rvert$	$\lvert s_z^2-1/12\rvert$	$\lvert c_1\rvert/s_z^2$
Встроенный в язык C++	300	10	0,428	0,183	0,0224	0,007306	0,091
			1,000	1,000	0,0017	0,000278	0,101
Из пакета Statistica	1000	20	0,972	0,748	0,0129	0,005147	0,018
			1,000	1,000	0,0005	0,000083	0,037
Встроенный в язык Java	1000	20	0,364	0,107	0,0084	0,004332	0,081
			1,000	1,000	0,0005	0,000083	0,093

Видно, что предложенное преобразование значительно улучшает свойства исходных датчиков (соответствующие P-значения существенно увеличиваются). Отметим, что полные результаты тестов "совпадение моментов" и "ковариация" в целях экономии места не помещены в таблицу. Однако все преобразованные последовательности прошли и эти тесты (например, с уровнем значимости $\varepsilon = 0{,}05$).

ЛИТЕРАТУРА

1. *Кнут, Д.* Искусство программирования для ЭВМ. Т. 2: Получисленные алгоритмы / Д. Кнут. М. : Мир, 1977. 728 с.

2. *Харин, Ю. С.* Практикум на ЭВМ по математической статистике / Ю. С. Харин, М. Д. Степанова. Мн. : Университетское, 1987. 304 с.

3. *Харин, Ю. С.* Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук. Мн. : БГУ, 2005. 279 с.