УДК 519.254

ОБ ИНВАРИАНТАХ СТРУКТУРЫ СЕРИЙ И КРИТЕРИЯХ СЛУЧАЙНОСТИ ПОСЛЕДОВАТЕЛЬНОЙ ВЫБОРКИ

© 2006 А.Н. Плотников

Самарский государственный аэрокосмический университет

В статье рассмотрены закономерности формирования серий, образуемых в последовательной выборке непрерывной случайной величины положениями индивидуальных значений относительно центральной линии (медианы) и отношениями порядка между соседними значениями. Показана возможность обобщения результатов теории рекуррентных событий Мизеса-Феллера для серий, образуемых отношениями порядка. Получен общий вид производящей функции времени возвращения и асимптотические оценки числовых характеристик числа серий фиксированной длины для серий указанного типа. На основании установленных законов распределения выявлены инварианты структуры серий, сформулированы критерии случайности выборки и проведена их экспериментальная проверка путем статистического моделирования методом Монте-Карло.

Одним из основных исходных понятий в приложениях теории вероятностей является понятие случайная выборка (из генеральной совокупности). При этом категория случайности, строго определяемая, как равновероятность попадания в выборку каждого из значений генеральной совокупности при постоянстве последней, зачастую интерпретируется, как отсутствие какой бы то ни было закономерности в последовательности выборочных значений, то есть как хаотичная последовательность. Однако это не совсем так и при ближайшем рассмотрении случайность, а именно, равновероятность, обнаруживает признаки закономерности вполне детерминированного характера.

Одним из проявлений закономерностей в случайной последовательности является образование в ней по мере возрастания длины характерных структур (структурных инвариантов), которые представляются интересными с точки зрения возможных приложений и могут служить критериями случайности.

Известными элементами таких структур являются инверсии, циклы и серии [1,2].

В [3] был рассмотрен закон распределения длины максимальной серии. Далее речь пойдет о некоторых результатах, касающихся более детальной структуры серий. Прежде видимо следует в двух словах повторить преобразования, позволяющие установить связь между последовательной выборкой и классической теорией серий.

Рассмотрим последовательную выборку непрерывной случайной величины (С.В.). Для каждого выборочного значения очевидно существуют два и только два равновозможных и взаимоисключающих положения относительно медианы ((>>//<<)). Причем, положения всех значений (точек) независимы в совокупности, следовательно, закон их чередования идентичен закону чередования исходов опытов с симметричной монетой.

Для каждой пары соседних точек существуют также два равновозможных отношения порядка (два знака последовательной разности). Однако последовательные разности (П.Р.) уже не являются независимыми. Любые две соседние П.Р. коррелированы с коэффицентом $rr=-\frac{1}{2}$ [[4]]. Применительно к опытам с монетой можно представить дело таким образом, что монета “запоминает” предидущий исход и с вероятностью $\frac{1}{3}$ воспроизводит его в следующем опыте. Соответственно с вероятностью $\frac{2}{3}$ реализуется альтернативный результат [3,4].

Серией в двоичной последовательности, однозначно определяемой исходной выборкой, является группа последовательных то-

чек одного знака. Причем, следуя В. Феллеру [[11]], определим серию, как рекуррентное событие. Например, отрезок последовательности …0111110… одновременно содержит 5 серий “1” (успехов) длиной 1, 2 серии длиной 2 и по 1-ой серии длиной 3,4 и 5. Такое определение серий позволяет использовать аналитический аппарат теории рекуррентных событий, являющейся , в свою очередь, частным случаем теории восстановления [1].

Серии положений точек относительно медианы являются сериями успехов в последовательности испытаний Бернулли с вероятностью успеха $p=\frac{1}{2}$, для которых в литературе имеется исчерпывающие или почти исчерпывающие результаты (относительно рассматриваемой задачи). Их краткое изложение ниже приведено исключительно с целью сокращения последующих выкладок, касающихся серий отношений порядка, путем рассуждения по аналогии там, где это представится возможным.

Рассмотрим последовательность исходов испытаний Бернулли и введем в рассмотрение целочисленные С.В. T_l – длину последовательности, при которой образуется первая серия успехов длиной l – время возвращения серии длины l и $R_n^{(l)}$ – число серий длины l в последовательности длиной $n \geq l$. Как показано в [1] , серия успехов длины l является достоверным рекуррентным событием с конечным средним временем возвращения μ_{T_l}=M[[T_l]] и конечной дисперсией $\sigma^2_{T_l}$=D[[T_l]]. Для числа серий при больших n, как гласит теорема Мизеса-Феллера[1], справедлива асимптотическая нормальная оценка:

$$R_n^{(l)} \sim N\left(\frac{n}{\mu_{T_l}}, \sigma_{T_l}\sqrt{\frac{n}{\mu^3_{T_l}}}\right) .$$

Таким образом, задача установления закона распределения числа серий $R_n^{(l)}$ сводится к отысканию числовых характеристик времени возвращения T_l. Для вычисления числовых характеристик удобнее всего воспользоваться аппаратом производящих функций.

Пусть $u_n^{(l)}$ – вероятность того, что на шаге с номером n образуется очередная серия длиной l. Тогда для $u_n^{(l)}$ справедливо рекуррентное соотношение [1] :

$$\begin{aligned} &u_n^{(l)} + pu_{n-1}^{(l)} + p^2 u_{n-2}^{(l)} + \ldots + p^{l-1}u_{n-l+1}^{(l)} = p^l, \\ &u_0^{(l)} = 1,\ u_1^{(l)} = u_2^{(l)} = \ldots = u_{l-1}^{(l)} = 0 \end{aligned} \qquad (1)$$

Умножая первое соотношение (1) на s^k и суммируя по всем $k \geq l$, получаем производящую функцию последовательности $u_n^{(l)}$:

$$U_l(s) = \frac{1 - s + (1-p)p^l s^{l+1}}{(1-s)(1-p^l s^l)} . \qquad (2)$$

Переходя к производящей функции хвостов времени возвращения, используя соотношение $Q_l(s) = \frac{1}{(1-s)U_l(s)}$ [11], получаем :

$$Q_l(s) = \frac{1 - p^l s^l}{1 - s + (1-p)p^l s^{l+1}} . \qquad (3)$$

Числовые характеристики времени возвращения находим используя свойство функции Q (s) и полагая $p=\frac{1}{2}$:

$$\begin{aligned} &\mu_{T_l} = Q_l(1) = 2^{l+1} - 2, \\ &\sigma^2_{T_l} = 2Q_l^{'}(1) + Q_l(1) - Q_l^2(1) = 2^{2(l+1)} - (2l+1)2^{l+1} - 2 . \end{aligned} \qquad (4)$$

Далее, на основании теоремы Мизеса – Феллера получаем числовые характеристики числа серий успехов длины l :

$$\mu_{R_n^{(l)}} \approx \frac{n}{\mu_{T_l}} \quad , \quad \sigma^2_{R_n^{(l)}} \approx \frac{n\sigma^2_{T_l}}{\mu^3_{T_l}} . \qquad (5)$$

Для больших l очевидна асимптотическая оценка:

$$\mu_{R_n^{(l)}} \approx \sigma^2_{R_n^{(l)}} \approx \frac{n}{2^{l+1}} \quad . \qquad (6)$$

Таким образом число длинных серий успехов имеет Пуассоновское распределение с параметром $\lambda_n^{(l)} = \frac{n}{2^{l+1}}$.

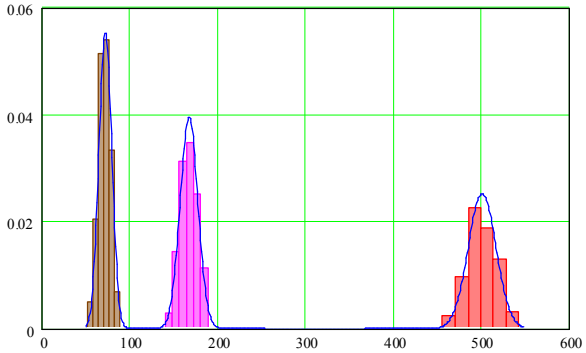

Рис. 1. Распределение числа серий успехов в выборке объема n=1000 в зависимости от длины серии ($l = 1 \div 3$)

На рис. 1 представлены результаты статистического моделирования. Гистограммы числа серий для значений $l = 1 \div 3$ (в порядке убывания средних) построены по 200 реализациям нормальной выборки объема $n = 1000$. Сглаживающие кривые представляют собой функции Гаусса с числовыми характеристиками, вычисленными в соответствии с (5).

Средние и дисперсии числа серий первого типа приведены в табл. 1.

Зная закон распределения числа рекуррентных (Феллеровых) серий, можно установить закон распределения "естественных" серий фиксированной длины $\tilde{R}_n^{(l)}$, то есть когда короткие серии поглощаются покрывающей их более длинной серией. Так, в ранее рассмотренном примере будет содержаться только одна серия успехов длиной $l = 5$.

Числа рекуррентных серий связаны с числами естественных серий системой линейных уравнений с матрицей следующего вида:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4\ldots \\ 0 & 1 & 1 & 2\ldots \\ 0 & 0 & 1 & 1\ldots \\ 0 & 0 & 0 & 1\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \end{bmatrix}.$$

Общая формула для элемента такой матрицы имеет вид: $a_{ln} = \left[\frac{n}{l}\right]$. Определитель A равен 1, следовательно существует A^{-1}. При этом среднее значение числа естественных серий можно вычислить и без решения системы уравнений. Они составляют $\mu_{\tilde{R}_n^{(l)}} = \frac{n}{2^{l+2}}$. Дисперсии представляют собой взвешенные суммы: $\sigma^2_{\tilde{R}_n^{(l)}} = \sum_{k=l}^{n} \left(a_{lk}^{-1}\right)^2 \sigma^2_{R_n^{(k)}}$, где a_{lk}^{-1} – элементы матрицы A^{-1}. Отсюда следует, что числа рекуррентных серий обладают минимальными дисперсиями, стало быть, критерии, построенные на законах их распределения обладают большей эффективностью. То же самое справедливо и для суммарного числа серий успехов и неудач фиксированной длины.

Производящую функцию времени возвращения трендовой серии найдем по аналогии. Пусть, как и раньше, $u_n^{(l)}$ – вероятность образования на шаге с номером n очередной восходящей серии длиной l. В пространстве П.Р. соответственно образуется рекуррентная серия успехов ("1") с параметрами: $n' = n - 1$, $l' = l - 1$. Как было показано в [33], вероятности серий в пространстве П.Р. инвариантны по отношению к закону распределения совокупности и определяются с помощью собственных функций: $u_n^{(l)} = \int_0^1 \varphi_n^{(l)}(x)dx$.

Для последних справедливы рекуррентные соотношения, аналогичные (1). Отличие заключается в том, что порядок рекуррентного соотношения на единицу меньше, а ум-

Таблица 1. Среднее значение и дисперсия числа серий в зависимости от длины серии

l	1	2	3	4	5	6	≥ 7
$\frac{\mu}{n}$	0,5	0,167	0,071	0,033	0,016	0,008	$\frac{1}{2^{l+1}}$
$\frac{\sigma^2}{n}$	0,25	0,102	0,052	0,027	0,014	0,007	$\frac{1}{2^{l+1}}$

ножению на вероятность успеха p соответствует операция $\int_0^x \bullet\, dx$. Например, для l=2÷5 рекуррентные соотношения имеют вид:

$$\varphi_n^{(2)}(x) = x,$$
$$\varphi_n^{(3)}(x) + \int_0^x \varphi_{n-1}^{(3)}(x)dx = \frac{1}{2}x^2,$$
$$\varphi_n^{(4)}(x) + \int_0^x \varphi_{n-1}^{(4)}(x)dx + \int_0^x\int_0^x \varphi_{n-2}^{(4)}(x)dx = \frac{1}{6}x^3,$$
$$\varphi_n^{(5)}(x) + \int_0^x \varphi_{n-1}^{(5)}(x)dx + \int_0^x\int_0^x \varphi_{n-2}^{(5)}(x)dx + \int_0^x\int_0^x\int_0^x \varphi_{n-3}^{(5)}(x)dx = \frac{1}{24}x^4.$$

В общем виде, при произвольном l, рекуррентное соотношение выглядит следующим образом:

$$\varphi_n^{(l)}(x) + \int_0^x \varphi_{n-1}^{(l)}(x)dx + \ldots + \int_0^x \ldots\ldots \int_0^x \varphi_{n-l+2}^{(l)}(x)dx = \frac{x^{l-1}}{(l-1)!},$$
$$\varphi_1^{(l)}(x) \equiv 1,\ \varphi_2^{(l)}(x) = \ldots = \varphi_{l-1}^{(l)}(x) \equiv 0. \quad (7)$$

По аналогии с сериями успехов в последовательных испытаниях Бернулли, производящую функцию

$$\mathrm{U}_l(s) = 1 + u_l^{(l)} s^{l-1} + u_{l+1}^{(l)} s^l + \ldots$$

можно получить непосредственно из рекуррентного соотношения (7):

$$U_l(s) = 1 + \int_0^1 \left\{ \frac{(sx)^{l-1}}{(l-1)!} + s^l \left[\frac{x^{l-1}}{(l-1)!} - \int_0^x \frac{x^{l-1}}{(l-1)!}dx \right] + \right.$$
$$\left. + s^{l+1} \left[\frac{x^{l-1}}{(l-1)!} - \int_0^x \left[\frac{x^{l-1}}{(l-1)!} - \int_0^x \frac{x^{l-1}}{(l-1)!}dx \right] dx \right] + \ldots \right\} dx, \quad (8)$$

Приводя подобные по степеням x и суммируя образующиеся геометрические прогрессии с показателем s, получим:

$$U_l(s) = 1 + \frac{1}{1-s}\int_0^1 \left\{ \frac{(sx)^{l-1}}{(l-1)!} - \frac{(sx)^l}{l!} + \frac{(sx)^{2l-2}}{(2l-2)!} - \frac{(sx)^{2l-1}}{(2l-1)!} + \ldots \right\} dx \quad (9)$$

Выражение под интегралом в (9) представляет собой ряд Макларена функции $\psi_l(sx)$, которая является решением рекурсивного уравнения, соответствующего (7):

$$\psi_l(x) + \int_0^x \psi_l(x)dx + \ldots + \int_0^x \ldots \int_0^x \psi_l(x)dx = \frac{x^{l-1}}{(l-1)!}$$

или

$$\frac{d^{l-2}\psi_l}{dx^{l-2}} + \frac{d^{l-3}\psi_l}{dx^{l-3}} + \ldots + \frac{d\psi_l}{dx} + \psi_l = x. \quad (10)$$

В компактной записи соотношение (9) примет вид:

$$U_l(s) = 1 + \frac{1}{1-s}\int_0^1 \psi_l(sx)dx = 1 + \frac{1}{s(1-s)}\int_0^s \psi_l(x)dx. \quad (11)$$

Переходя к производящей функции хвостов времени возвращения, получаем:

$$Q_l(s) = \frac{1}{(1-s)U_l(s)} = \frac{s}{s(1-s) + \int_0^S \psi_l(x)dx} \quad (12)$$

Рассмотрим уравнение (10) для случая $l > 2$. Данное уравнение представляет собой неоднородное линейное с постоянными коэффициентами (все равны 1). Его частный интеграл имеет вид:

$$\hat{\psi}(x) = x - 1.$$

Общий интеграл однородного уравнения будем искать с помощью преобразования Лапласа. Исходя из вида уравнения (10) получаем характеристическое уравнение:

$$1 + q + q^2 + \ldots + q^{l-2} = 0 \quad (13)$$

Корни уравнения (13) в Эйлеровом тригонометрическом виде образуют группу по умножению:

$q_k = e^{i\frac{2\pi k}{l-1}}$, $k = 0,1,\ldots,l-2$, за вычетом точки $q_0 = 1$.

Таким образом общий интеграл (10) получаем в виде:

$$\psi_l(x) = x - 1 + \sum_{k=1}^{l-2} b_k e^{xq_k}. \quad (14)$$

Неопределенные коэффициенты b_k, $k = 1,\ldots,l-2$ находим из однородных начальных условий в точке $x = 0$: $\frac{d^{k-1}\psi_l}{dx^{k-1}} = 0$.

Таблица 2. Средние и дисперсии числа трендовых серий

l	2	3	4	5	≥ 6
$\dfrac{\mu}{n-1}$	0,5	0,132	0,034	$6{,}9\cdot 10^{-3}$	$\dfrac{l}{(l+1)!}$
$\dfrac{\sigma^2}{n-1}$	0,074	0,060	0,026	$6{,}5\cdot 10^{-3}$	$\dfrac{l}{(l+1)!}$

Для $l = 2 \div 5$ получим:

$$\psi_2 = x\,, \quad \psi_3 = e^{-x} + x - 1\,, \quad \psi_4 = \frac{\left(\sqrt{3}+i\right)e^{-\frac{1-i\sqrt{3}}{2}x} + \left(\sqrt{3}-i\right)e^{-\frac{1+i\sqrt{3}}{2}x}}{2\sqrt{3}} + x - 1\,,$$

$$\psi_5 = \frac{1}{2}e^{-x} + \frac{1+i}{4}e^{ix} + \frac{1-i}{4}e^{-ix} + x - 1.$$

При больших l можно получить асимтотическую оценку $\psi_l(x)$, вполне удовлетворительную для поставленных целей, и, соответственно, оценки искомых числовых характеристик числа серий.

Рассмотрим более детально ряд Маклорена функции $\psi_l(x)$.

Исходя из вида уравнения (10) следует, что последовательность производных $\psi_l(x)$ в нуле имеет период $l-1$, а именно, отличны от нуля только члены, кратные $l-1$ и на "1" старше. Причем, все члены первой подпоследовательности равны "1", второй – "-1".

Поскольку члены ряда имеют факториальную скорость убывания, то на интервале $x \in [0;1]$ главным значением ряда будет сумма двух первых членов, а остаток можно оценить порядком третьего (первого отброшенного) члена:

$$\psi_l(x) = \frac{x^{l-1}}{(l-1)!} - \frac{x^l}{l!} + O\left(\frac{x^{2(l-1)}}{(2l-2)!}\right) \quad .$$

Откуда получаем асимтотические оценки:

$$\psi_l(1) \approx \frac{l-1}{l!} \quad , \quad \int_0^1 \psi_l(x)dx \approx \frac{l}{(l+1)!}. \quad (15)$$

Далее, на основании (8) и (4) находим среднее и дисперсию времени возвращения трендовой серии:

$$\mu_{T_l} = \left[\int_0^1 \psi_l(x)dx\right]^{-1} \quad , \quad \sigma^2_{T_l} = 3\mu_{T_l} + \left[1 - 2\psi_l(1)\right]\mu^2_{T_l} \quad (16)$$

При использовании теоремы Мизеса-Феллера следует учесть, что длина цепи последовательных разностей на 1 короче длины исходной выборки. По этому формулы для числовых характеристик числа серий будут несколько отличаться от (5):

$$\mu_{R_n^{(l)}} \approx \frac{n-1}{\mu_{T_l}} \quad , \quad \sigma^2_{R_n^{(l)}} \approx \frac{(n-1)\sigma^2_{T_l}}{\left(1+\mu_{T_l}\right)^3}\,. \quad (17)$$

Подставляя в (17) (16) и (15), получаем оценки числовых характеристик числа серий при больших l:

$$\frac{\mu_{R_n^{(l)}}}{n-1} \approx \frac{\sigma^2_{R_n^{(l)}}}{n-1} \approx \frac{l}{(l+1)!}. \quad (18)$$

Для практических целей полученными Пуассоновскими оценками (18) с достаточной точностью можно пользоваться уже начиная с l=6. Точные значения числовых характеристик для коротких серий приведены в табл. 2.

Результаты статистического моделирования представлены на рис. 2. Как и в предыдущем случае, гистограммы числа возрастающих трендовых серий построены по 200 реализациям. Сглаживающие кривые – функции Гаусса с числовыми характеристиками (17).

Как видно из представленных на рис.1, 2 графиков, отличие в структурах серий разных типов лишь количественное, заключающееся в различии рядов средних и дисперсий. Практически полное совпадение наблюдается лишь при $l = 4$ (табл.1, 2.) При этом число, а точнее поток серий имеет отчетливую спектральную структуру. Количество различимых спектральных полос и их контрастность возрастают пропорционально $\sqrt{n}$. Такое свойство структуры серий позволяет установить надежный критерий случайности –

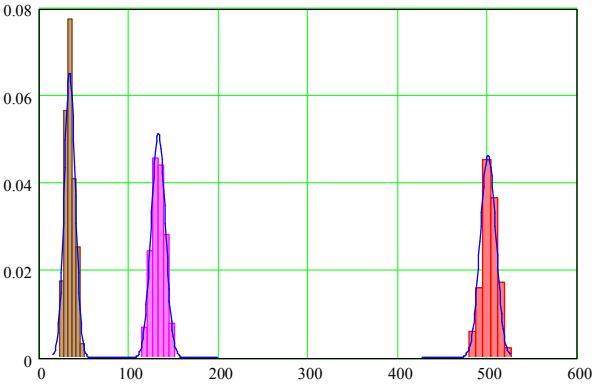

Рис. 2. Распределение числа восходящих трендовых серий в выборке объема n=1000 в зависимости от длины серии $l = 2 \div 4$

отсутствие инверсий среди контрастных спектральных полос. Или, другими словами, наличие хотя бы одной инверсии в спектре числа серий можно обоснованно интерпретировать, как искусственное упорядочение последовательной выборки.

В заключение следует указать на еще один интересный эффект в структуре серий случайной последовательности – наивероятнейшее появление серии в первом из возможных исходов. Это явление, природа которого, по видимому, имеет нечто общее с “Петербургским парадоксом” [1], обусловлено тем, что ряд распределения времени возвращения $f_n^{(l)} = \mathrm{P}\{\mathrm{T}_l = n\}$ монотонно убывает, и первый отличный от нуля член $(n = l)$ является существенно доминирующим [1, 3].

Другим, более значимым проявлением указанной закономерности является “притяжение” длинных серий (преимущественно разного знака). Дело в том, что в рекуррентной трактовке любая конечная последовательность рассматривается как отрезок бесконечной в обе стороны последовательности, и после точки, завершающей очередную серию, отсчет начинается заново. В связи с этим максимальным правдоподобием среди возможных расположений двух или более серий в отрезке последовательности обладает конгломерат, то есть имеет место “эффект притяжения” серий. И напротив, большое расстояние между сериями является маловероятным, а, если наблюдается, то может свидетельствовать о неслучайном характере последовательности.

СПИСОК ЛИТЕРАТУРЫ

1. *Феллер В.* Введение в теорию вероятностей и ее приложения. Т. 1 (Дискретные распределения). М.: МИР, 1984.
2. *Дунин-Барковский И.В., Смирнов Н.В.* Теория вероятностей и математическая статистика в технике (общая часть). М.: ГИТТЛ, 1955.
3. *Плотников А.Н.* Закон распределения длины максимальной серии и его статистические приложения / Известия СамНЦ РАН. 2006. Т 8. №4.
4. *Юнак Г.Л., Годлевский В.Е., Плотников А.Н.* Об интерпретации серий на контрольных картах // Методы менеджмента качества. 2005. №4.

ABOUT INVARIANTS STRUCTURES OF SERIES AND CRITERIA OF ACCIDENT OF CONSECUTIVE SAMPLE

© 2006 A.N. Plotnikov

Samara State Aerospace University

In article are considered laws of formation of series, in consecutive sample of a continuous random variable by positions of individual values concerning the central line and attitudes of the order between the next values. The opportunity of generalization of results of the theory of Mizes-Feller's recurrent events for series, is shown by attitudes of the order. The general view of making function of time of returning and asymptotic estimations of numerical characteristics of number of series of the fixed length for series of the specified type is received. On the basis of the established laws of distribution are revealed инварианты structures of series, criteria of accident of sample are formulated and their experimental check by statistical modelling by a method of Monte-Carlo is lead.