

10, октябрь 2015

УДК 519.212.2

О свойствах вероятностей f – рекуррентных серий и цепочек в последовательности независимых случайных величин

***Якашев Б.К.**, студент
Россия, 105005, г. Москва, МГТУ им Н.Э. Баумана,
кафедра «Прикладная математика»*

*Научный руководитель: Меженная Н.М., к.ф.-м.н, доцент
Россия, 105005, г. Москва, МГТУ им Н.Э. Баумана,
кафедра «Прикладная математика»
Mezhennaya@bmstu.ru*

Введение

Понятие и потенциальное использование цепочек могут быть объяснены даже в простых терминах. Цепочку можно определить, как непрерывную последовательность определенного типа элементов, которая граничит с другим типом элементов.

Для примера рассмотрим двоичную последовательность 11001110, эта последовательность состоит из двух единиц, двух нулей, трех единиц и одного нуля. Таким образом, последовательность имеет четыре цепочки из одинаковых знаков. Свойства дискретной случайной последовательности во многом определяются количественными величинами связанными с цепочками в последовательности. Для примера мы можем рассмотреть длину максимальную цепочки равную 3 или разницу между числом цепочек состоящих из 0 или 1, она равна $2 - 2 = 0$.

Одно из применений цепочек возникает в статистическом контроле качества или регулирование технологического процесса. Здесь мы предполагаем, что у нас имеется численная характеристика производимых деталей, которая имеет верхние и нижние пределы. Производственный процесс называется контролируемым, если он производит достаточное количество изделий, отвечающих техническим требованиям, и процесс объявляют неконтролируемым, если среди произведенных деталей количество не отвечающих техническим требованиям превышает максимальную длину цепочки.

В данной работе будет изучено точное и асимптотическое распределение чисел f – рекуррентных серий в последовательности независимых случайных величин.

Реккурентные формулы для вычисления вероятностей

Пусть $X_1, X_2, \dots, X_T, \dots$ – последовательность независимых случайных величин, принимающих значения из алфавита $A_M = \{0, 1, \dots, M-1\}$, с вероятностями $\{p_0, p_1, \dots, p_{M-1}\}$. Такие случайные величины принято называть случайными знаками, или просто знаками. Пусть $f: A_M^l \rightarrow A_M$ согласно работе [1], случайные знаки $X_i, \dots, X_{i+l+s-1}$ образуют f – рекуррентную серию длиной s ($s \geq l+1$), если выполнено

$$X_{i+l+k} = f(X_{i+k}, \dots, X_{i+l+k-1}), \quad k = 0, 1, 2, \dots, s-1.$$

Начнем с того, что найдем вероятность того, что появилась f – рекуррентная серия длины S , которая заканчивается на $a_1, \dots, a_l \in \{1, \dots, l\}$.

Пусть $p_{s,a_1,\dots,a_l} = P\{f \text{ – рекуррентная серия длины } s, \text{ заканчивается } a_1, \dots, a_l\}$. Тогда имеет место рекуррентная формула

$$p_{s+1,a_1,\dots,a_l} = \sum_{b=1}^l p_{s,b,a_1,\dots,a_{l-1}} \cdot P\{X_{s+1} = f(b, (a_1, \dots, a_l)) = a_l\}.$$

Нетрудно заметить, что поиск вероятности в общей постановке является достаточно трудоемкой задачей из-за своей вычислительной сложности. Действительно, получается система из M^l уравнений, а не менее M^l начальных данных.

Рассмотрим случай, когда $l = 2, M = 2, A_2 = \{0, 1\}$ тогда $f: \{0, 1\}^2 \rightarrow \{0, 1\}$.

X_1	X_2	$f_1(X_1, X_2)$	$f_2(X_1, X_2)$	...	$f_{16}(X_1, X_2)$
0	0	0	1	...	1
0	1	0	0	...	1
1	0	0	0	...	1
1	1	0	0	...	1

В этом случае получаем шестнадцать рекуррентных функций, каждая функция имеет по четыре начальных распределений.

Найдем вероятность появления f – рекуррентной серии длины s , которая заканчивается на ab , (см.[3]) зафиксируем в качестве рекуррентной функции $f(a, b) = a \oplus b$.

Пусть $p_{s,ab} = P\{f \text{ – рекуррентная цепочка длины } s, \text{ заканчивается } ab\}$

$$\underbrace{X_1, \dots, \overbrace{X_{s-1}}^a, \overbrace{X_s}^b, X_{s+1}}_{f\text{-рекуррентная цепочка}}.$$

Тогда получаем систему уравнений

$$\begin{cases} p_{s+1,00} = \sum_a p_{s,a0} \cdot P\{X_{s+1} = (a \oplus 0 = 0)\} = p_{s,a0} \cdot P\{a = 0\} = p_{s,00} \cdot p_0, \\ p_{s+1,10} = \sum_a p_{s,a1} \cdot P\{X_{s+1} = (a \oplus 1 = 0)\} = p_{s,a1} \cdot P\{a = 1\} = p_{s,11} \cdot p_1, \\ p_{s+1,01} = \sum_a p_{s,a0} \cdot P\{X_{s+1} = (a \oplus 0 = 1)\} = p_{s,a0} \cdot P\{a = 1\} = p_{s,10} \cdot p_1, \\ p_{s+1,11} = \sum_a p_{s,a1} \cdot P\{X_{s+1} = (a \oplus 1 = 1)\} = p_{s,a1} \cdot P\{a = 0\} = p_{s,01} \cdot p_0. \end{cases} \quad (1)$$

Вероятность искомого события состоит из суммы вероятностей нескольких событий, а именно:

$$p_s = p_{s,00} + p_{s,10} + p_{s,01} + p_{s,11}.$$

Таким образом, необходимо решить систему (1) относительно $p_{s,00}, p_{s,10}, p_{s,01}, p_{s,11}$, и начальные условия.

Для начала перепишем систему (1) в виде

$$\begin{cases} p_{s+1,00} = p_{s,00} \cdot p_0, \\ p_{s+1,10} = p_{s,11} \cdot p_1, \\ p_{s+1,01} = p_{s,10} \cdot p_1, \\ p_{s+1,11} = p_{s,01} \cdot p_0. \end{cases} \quad (2)$$

Рассмотрим 2-ое уравнение в системе (2), его можно представить в виде

$$p_{s+1,10} = p_0 \cdot p_1 \cdot p_{s-1,01} = p_0 \cdot p_1^2 \cdot p_{s-2,10}. \quad (3)$$

Начальные условия имеют вид

$$p_{1,10} = p_1^2 \cdot p_0, \quad p_{2,10} = p_1^2 \cdot p_0^2, \quad p_{3,10} = p_1^3 \cdot p_0^2. \quad (4)$$

Решение рекуррентного уравнения (3) будем искать в виде $p_{s,10} = z^s$. Подставив эту функцию в (3), получим $z = \frac{p_0 p_1^2}{z^2}$. Из последнего уравнения вытекает

$$z_1 = p_0^{1/3} p_1^{2/3}, \quad z_2 = -\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{1/3} p_1^{2/3}, \quad z_3 = \left(-\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{1/3} p_1^{2/3}.$$

Тогда

$$\begin{cases} p_{1,10} = C_1 p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} + C_2 \left[-\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right] + C_3 \left[\left(-\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right], \\ p_{2,10} = C_1 p_0^{\frac{2}{3}} p_1^{\frac{4}{3}} + C_2 \left[\left(\frac{1}{2} - \frac{i\sqrt{3}}{2}\right) p_0^{\frac{2}{3}} p_1^{\frac{4}{3}} \right] + C_3 \left[\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{2}{3}} p_1^{\frac{4}{3}} \right], \\ p_{3,10} = C_1 p_0 p_1^2 + C_2 p_0 p_1^2 + C_3 p_0 p_1^2. \end{cases}$$

Используя начальные условия (4), получаем выражения для C_1 , C_2 и C_3 .

$$\begin{aligned} C_1 &= \frac{1}{3} \left(p_0^{\frac{4}{3}} p_1^{\frac{2}{3}} + p_0^{\frac{2}{3}} p_1^{\frac{4}{3}} + p_0 p_1 \right), \\ C_2 &= - \frac{i p_0^{\frac{2}{3}} \left(p_0^{\frac{1}{3}} - p_1^{\frac{1}{3}} \right) \left(2\sqrt{3} p_0^{\frac{1}{3}} + (\sqrt{3} + 3i) p_1^{\frac{1}{3}} \right) p_1^{2/3}}{9 + i3\sqrt{3}}, \\ C_3 &= \left(\frac{1}{12} + \frac{i}{4\sqrt{3}} \right) p_0^{\frac{2}{3}} \left(p_0^{\frac{2}{3}} - p_1^{\frac{1}{3}} \right) \left((1 + i\sqrt{3}) p_0^{\frac{1}{3}} + (3i + \sqrt{3}) p_1^{\frac{1}{3}} \right) p_1^{\frac{2}{3}}. \end{aligned}$$

Возвращаясь к системе (2), выражаем остальные вероятности

$$\begin{aligned} p_{s,01} &= p_1^2 C_1 \left(p_1^{\frac{2}{3}} p_0^{\frac{1}{3}} \right)^s + p_1^2 C_2 \left[-\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s \\ &\quad + p_1^2 C_3 \left[\left(-\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s, \\ p_{s,11} &= p_0 p_1 C_1 \left(p_1^{\frac{2}{3}} p_0^{\frac{1}{3}} \right)^s + p_0 p_1 C_2 \left[-\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s \\ &\quad + p_0 p_1 C_3 \left[\left(-\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s, \\ p_{s,10} &= C_1 \left(p_1^{\frac{2}{3}} p_0^{\frac{1}{3}} \right)^s + C_2 \left[-\left(\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s + C_3 \left[\left(-\frac{1}{2} + \frac{i\sqrt{3}}{2}\right) p_0^{\frac{1}{3}} p_1^{\frac{2}{3}} \right]^s, \\ p_{s,00} &= p_0^s p_0^2. \end{aligned}$$

В рассмотренном случае для $l = 2$ и функции $f(a, b) = a \oplus b$ решение системы представляет собой сложность. Также можно заметить, что вид системы и её решение существенным образом зависит от вида функции f .

Для этой же функции $f(a, b) = a \oplus b$, ниже представлены графики, при различных вероятностях и различной длины цепочки S .

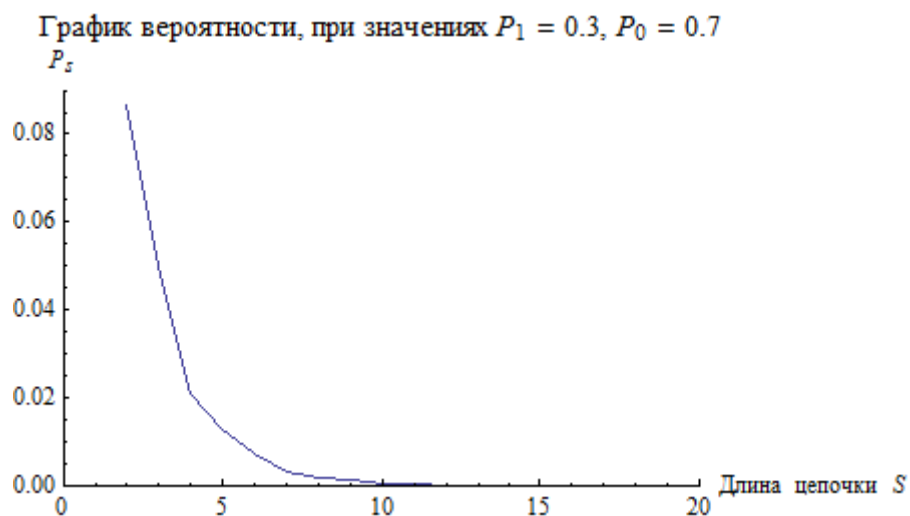


Рис. 1. График вероятности

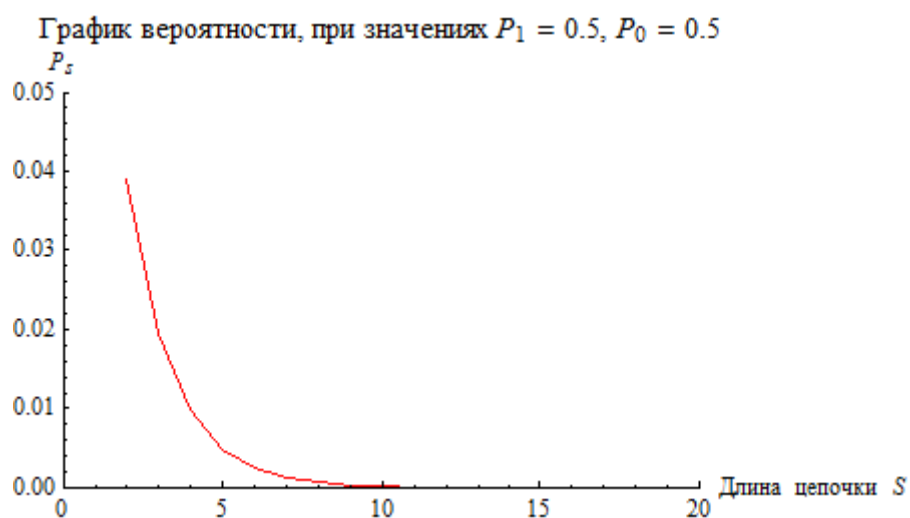


Рис. 2. График вероятности

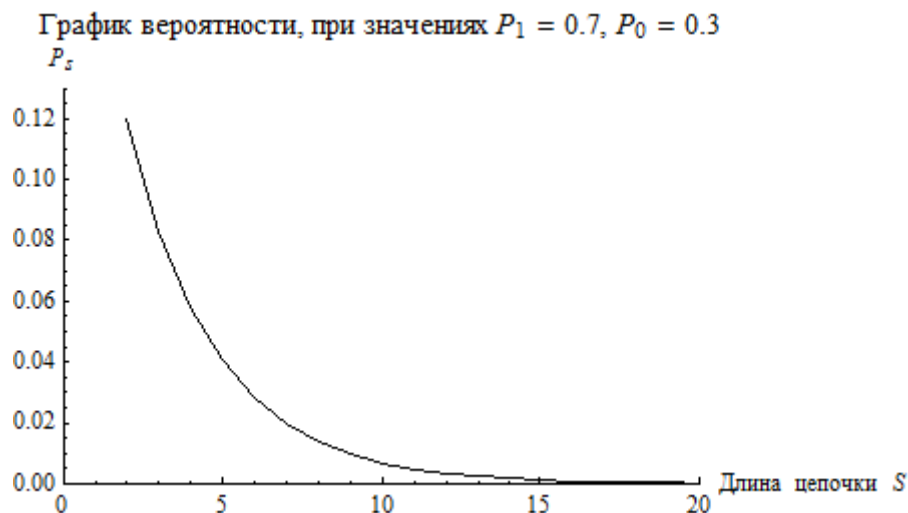


Рис. 3. График вероятности

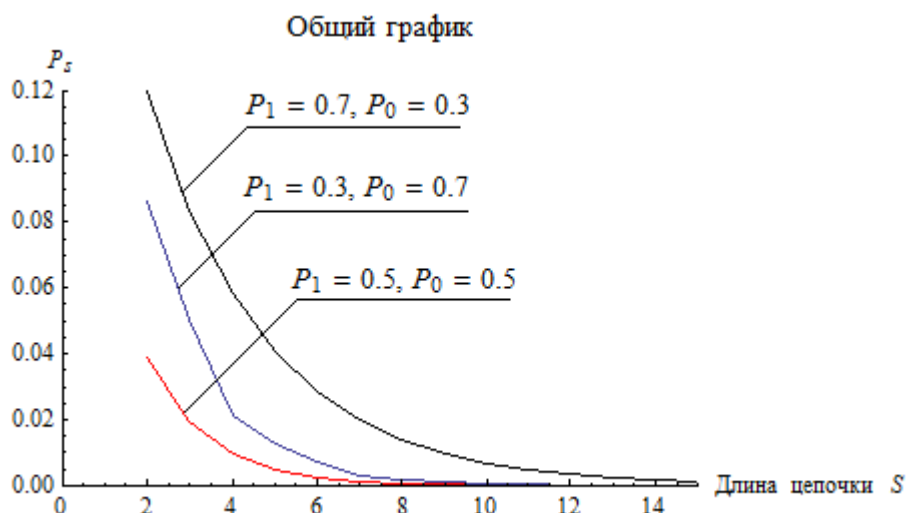


Рис. 4. Графики вероятностей

Вычисление точного распределения

Пусть ξ_s — число f — рекуррентных серий длины не менее s . Зафиксируем в качестве функции $f(a, b) = I_{\{a=1\}}$. Для того чтобы число f — рекуррентных серий длины не меньше s , сходилось к распределению Пуассона, необходимо чтобы $E\xi = N \cdot p_s$ была константой, но также известно, что у распределения Пуассона $E\xi = D\xi = \lambda$ (см. [2]). Для сходимости к нормальному распределению необходимо, чтобы $D\xi \rightarrow \infty$ тогда устремив $\lambda \rightarrow \infty$ получим сходимость к нормальному распределению.

Рассмотрим $f(a, b) = I_{\{a=1\}}$, $l = 2$ в равновероятном случае $p_1 = p_0 = 0.5$. Тогда имеем четыре начальных распределения, по которым будем строить функцию. В

зависимости от начального распределения, необходимо наложить дополнительные условия на f – рекуррентную цепочку, добавив такие элементы, чтобы было нарушено свойство f – рекуррентности и знаки образовывали f – рекуррентную серию.

На примере выбранной функции $f(a, b) = I_{\{a=1\}}$, подробно распишем эту процедуру.

$$\begin{cases} p_{s+1,00} = \sum_a p_{s,a0} \cdot P\{X_{s+1} = (f(a, 0) = 0)\} = p_{s,a0} \cdot P\{a = 0\} = p_{s,00} \cdot p_0, \\ p_{s+1,10} = \sum_a p_{s,a1} \cdot P\{X_{s+1} = (f(a, 1) = 0)\} = p_{s,a1} \cdot P\{a = 0\} = p_{s,01} \cdot p_0, \\ p_{s+1,01} = \sum_a p_{s,a0} \cdot P\{X_{s+1} = (f(a, 0) = 1)\} = p_{s,a0} \cdot P\{a = 1\} = p_{s,10} \cdot p_1, \\ p_{s+1,11} = \sum_a p_{s,a1} \cdot P\{X_{s+1} = (f(a, 1) = 1)\} = p_{s,a1} \cdot P\{a = 1\} = p_{s,11} \cdot p_1. \end{cases} \quad (5)$$

Вероятность наблюдаемой величины состоит из суммы вероятностей

$$p_s = p_{s,00} + p_{s,10} + p_{s,01} + p_{s,11}. \quad (6)$$

Таким образом, необходимо решить систему относительно $p_{s,00}, p_{s,10}, p_{s,01}, p_{s,11}$.

Перепишем систему (5) в виде

$$\begin{cases} p_{s+1,00} = p_{s,00} \cdot p_0, \\ p_{s+1,10} = p_{s,01} \cdot p_0, \\ p_{s+1,01} = p_{s,10} \cdot p_1, \\ p_{s+1,11} = p_{s,11} \cdot p_1. \end{cases}$$

Рассмотрим 2-ое уравнение в системе (5), его можно представить в виде

$$p_{s+1,10} = p_0 \cdot p_1 \cdot p_{s-1,10}. \quad (7)$$

Начальные условия имеют вид

$$p_{1,10} = p_0^2 \cdot p_1, \quad p_{2,10} = p_1^2 \cdot p_0^2.$$

Решение рекуррентного уравнения (7) будем искать в виде $p_{s,10} = z^s$. Подставив эту функцию в (5), получим $z = \frac{p_0 \cdot p_1}{z}$, решая которое получаем два значения

$$z_1 = +\sqrt{p_1 p_0}, \quad z_2 = -\sqrt{p_1 p_0}.$$

Решение будет искать в виде:

$$p_{s,10} = C_1 (\sqrt{p_1 p_0})^s + C_2 (-\sqrt{p_1 p_0})^s.$$

Используя начальные условия, составляем систему уравнений для неизвестных C_1 и C_2 .

$$\begin{cases} p_{1,10} = C_1 \sqrt{p_0 p_1} - C_2 \sqrt{p_0 p_1}, \\ p_{2,10} = C_1 p_0 p_1 + C_2 p_0 p_1. \end{cases}$$

Решая которую получаем,

$$\begin{aligned} C_1 &= \frac{1}{2} p_0 (p_1 + \sqrt{p_0 p_1}), \\ C_2 &= \frac{1}{2} p_0 (p_1 - \sqrt{p_0 p_1}). \end{aligned}$$

Возвращаясь к остальным уравнениям системы получаем,

$$\begin{aligned} p_{s,00} &= p_0^2 p_0^s, \\ p_{s,10} &= C_1 (\sqrt{p_1 p_0})^s + C_2 (-\sqrt{p_1 p_0})^s, \\ p_{s,01} &= C_1 p_1 (\sqrt{p_1 p_0})^{s-1} + C_2 p_1 (-\sqrt{p_1 p_0})^{s-1}, \\ p_{s,11} &= p_1^2 p_1^s. \end{aligned}$$

Процедуру подсчета вероятности можно выполнить и численно, для этого зафиксируем начальное распределение 10 и $s = 3$ в этом случае получаем f – рекуррентную цепочку вида: 10101 для нарушения свойства f – рекуррентности достаточно добавить 00 перед цепочкой, и после тогда получаем серию 0010101 и соответствующую ей вероятность $\widetilde{P}_3 = P_3 \cdot p_0^2$, где $P_3 = p_0^2 p_1^3$ – вероятность появления цепочки длины 3.

Аналогичным образом, продолжаем для остальных начальных распределений. В зависимости от начального распределения могут существенно отличаться вероятности. Даже в самом простом равновероятном случае получаются неравнозначные вклады слагаемых (7) в общую вероятность.

Используя равенство $N = \frac{E\xi}{\widetilde{P}_s}$, находим количество элементов в выборке, при котором распределение будет сходиться к нормальному. Далее приводятся несколько способов численной реализации вычисления вероятностей.

Способ 1

Генерируется 100 случайных последовательностей длины N , состоящих из независимых случайных величин, принимающих значения в алфавите $A_2 = \{0, 1\}$ с вероятностями p_0 и p_1 . Затем, в каждой из этих последовательностей ищется количество встретившихся цепочек.

На основании этих результатов выполняется проверка на нормальность по критерию Шапиро – Уилка, ($p - Value = 0.557841$). На рисунке 5 приведена гистограмма чисел появления серий.

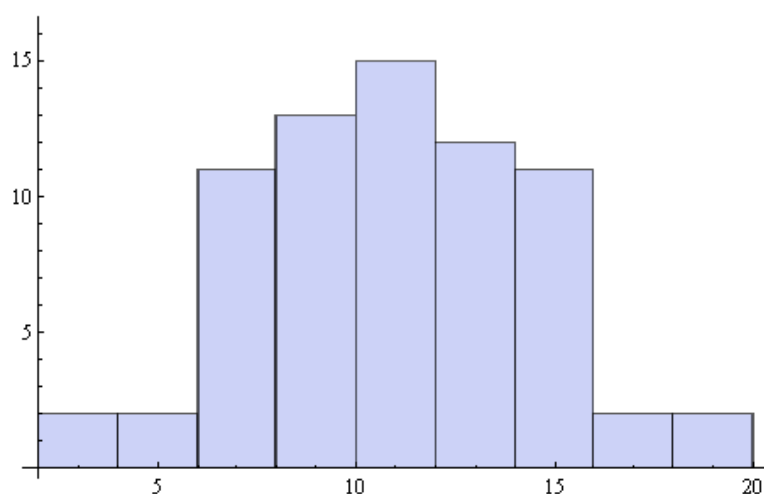


Рис. 5. Гистограмма

Способ 2

Перебираем все возможные различные последовательности длины N из знаков алфавита $A_2 = \{0, 1\}$. Для каждой из этих последовательностей известна вероятность её появления. В каждой последовательности ищем количество встретившихся цепочек. Затем вероятности чисел появления цепочек считаются как суммы вероятностей соответствующих последовательностей. Далее можно провести проверку на нормальность.

Основной недостаток этого способа заключается в большой вычислительной сложности и невозможности вычисления вероятностей для большой длины последовательности.

Вычисление точного распределения с помощью цепи Маркова

Вероятность появления f – рекуррентной цепочки длины не менее s можно найти, используя цепь Маркова.

Пусть f – фиксированная функция, необходимо найти вероятность появления цепочки длины не менее s . Введем состояния цепи Маркова, которые будут описывать характеристики цепочек:

$$(l, m, ab),$$

где

l – число серии длины не менее s , m – длина незавершенной серии, ab – последние знаки серии.

Пусть $l = 0, m = 0$, в этом случае возможны два перехода из этого состояния

$$1) (0, 0, ab) \xrightarrow{P_{f(ab)}} (0, 1, ab),$$

$$2) (0, 0, ab) \xrightarrow{1 - P_{f(ab)}} (0, 0, ab),$$

Пусть

$$1) (l, s, ab) \xrightarrow{P_{f(ab)}} (l, s + 1, dc),$$

$$2) (l, s, ab) \xrightarrow{1 - P_{f(ab)}} (l, 0, dc), \text{ при } l \geq s.$$

$$3) (l, s, ab) \xrightarrow{1 - P_{f(ab)}} (l + 1, 0, dc), \text{ при } l < s.$$

Нетрудно заметить, что максимальное число цепочек длины не менее s , в N элементах, не превосходит $\left\lfloor \frac{N}{s} \right\rfloor$.

Такому процессу можно поставить в соответствие матрицу переходных вероятностей A , которая будет иметь размерность $\left(\left\lfloor \frac{N}{s} \right\rfloor + 1 \right)^M + 1$, где M – алфавит.

Используя начальные данные, находим начальное распределение $\vec{f}_0$.

Для поиска вероятности используем формулу

$$\vec{f}_s = \vec{f}_{s-1} \cdot A.$$

В результате, вероятность искомого события, может быть найдена по следующей формуле

$$P\{\text{число серий} \geq s\} = \sum_{m,a,b} P\{l, m, ab\}.$$

Выводы

В данной статье исследованы свойства вероятностей f – рекуррентных серий и цепочек в последовательности независимых случайных величин.

В первом разделе были составлены рекуррентные формулы для вычисления вероятностей. Был рассмотрен частный случай, для которого были найдены аналитически выражения для вероятностей. Стоит отметить, что аналитическое решение в общей постановке имеет большую вычислительную сложность и существенным образом

зависит от выбираемой функции. В связи с этим целесообразно использовать численное решение данной задачи.

Во втором разделе предложен способ численной реализации алгоритма вычисления вероятностей появления чисел f –рекуррентных цепочек. Рассмотрено несколько различных алгоритмов.

На основании полученных результатов, была построена гистограмма и произведена проверка на нормальность.

В третьем разделе реализовали ещё один подход, который заключался в использовании цепи Маркова для вычисления аналогичных вероятностей.

Список литературы

1. Михайлов В.Г. О предельной теореме Б.А. Севастьянова для сумм зависимых случайных индикаторов // Обзорение прикладной и промышленной математики. 2003. Т. 10. Вып. 3. С. 571 - 578.
2. Михайлов В.Г. Явные оценки в предельных теоремах для сумм случайных индикаторов // Обозрения прикладной и промышленной математики. 1994. Т. 1, вып. 4. С. 580 - 617.
3. Меженная Н. М. Предельные теоремы для числа (a, d) – серий заданного веса в последовательности независимых случайных величин // Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки. Спец. вып. 2012. С. 20 - 28.