DOI 10.37882/2223–2966.2022.08.28

УЛУЧШЕНИЕ ХАРАКТЕРИСТИК ГЕНЕРАТОРОВ ПСЕВДОСЛУЧАЙНЫХ ЧИСЕЛ НА КОРОТКИХ ВЫБОРКАХ

IMPROVING THE CHARACTERISTICS OF PSEUDORANDOM NUMBER GENERATORS ON SHORT SAMPLES

P. Novikov

Summary. The paper proposes an approach that improves the quality of pseudorandom numerical sequences on short samples. A discrete approximation of the continuous distribution function is made. Various rules for constructing histograms have been considered. An algorithm has been created that forms the best (according to the criterion of agreement) uniform distribution on short samples. The ways of applying the algorithm to generate other distributions are demonstrated.

Keywords: pseudorandom number generator, discrete approximation of the distribution function, agreement criterion, histogram, rule for choosing the number of intervals, inverse function method, frequency spectrum, the central limit theorem, the Box-Muller transformation.

Новиков Павел Владимирович
К.т.н., доцент, Московский авиационный институт (национальный исследовательский университет)
novikov.mai@mail.ru

Аннотация. В статье предложен подход, улучшающий на коротких выборках качество псевдослучайных числовых последовательностей. Проведена дискретная аппроксимация непрерывной функции распределения. Изучены различные правила построения гистограмм. Создан алгоритм, формирующий наилучшее (по критерию согласия) равномерное распределение на коротких выборках. Показаны способы применения алгоритма для генерации других распределений.

Ключевые слова: генератор псевдослучайных чисел, дискретная аппроксимация функции распределения, критерий согласия, гистограмма, правило выбора числа интервалов, метод обратной функции, частотный спектр, центральная предельная теорема, преобразование Бокса-Мюллера.

Генераторы псевдослучайных чисел (ГПСЧ) встроены практически во все алгоритмические языки программирования высокого уровня. Они активно используются программистами для создания псевдослучайных числовых последовательностей. Однако эти генераторы по своим заявленным характеристикам часто не соответствуют реальным результатам, полученным при их исследовании. В результате созданные псевдослучайные последовательности оказываются лишь приблизительно похожими на те, что необходимы для моделирования и последующего использования случайных процессов в различных приложениях, например, при проектировании и исследовании цифровых систем обработки информации и управления. Достоверность вероятностных исследований с помощью таких генераторов из-за неточностей характеристик ГПСЧ оказывается существенно ниже, чем если бы характеристики совпадали с заявленными.

Типичным примером, подтверждающим это тезис, является широко распространённый генератор псевдослучайных чисел, распределённых в ограниченном интервале от 0 до верхней границы. Обычно этот ГПСЧ называется rand [1–3]. Программисты часто используют этот генератор как генератор равномерно распределённых псевдослучайных чисел, хотя известно, что распределение чисел, вырабатываемых этим генератором, можно считать равномерным лишь условно, с некоторой степенью приближения. Однако в тех случаях, когда программист сам моделирует псевдослучайную последовательность с распределением желаемого типа (нормального, экспоненциального, и т.п.) на основе равномерного распределения, он часто использует такой ГПСЧ вместе с известным алгоритмом, преобразующим равномерное распределение [4] в какое-нибудь иное распределение [5, 6]. В результате характеристики полученных распределений сильно расходятся с теоретически ожидаемыми.

Интересно заметить, что, даже в некоторых учебниках по языкам программирования, распределения, генерируемые такими ГПСЧ, называют равномерными. На самом деле, эти распределения не являются равномерными, а оказываются лишь ограниченными в заданных интервалах. Примером служат методы класса Random библиотеки классов языка программирования Java [2], называемые методами, генерирующими равномерно распределённые числа.

Большинство ГПСЧ вырабатывают детерминированную псевдослучайную последовательность: при одинаковых стартовых условиях получаются одинаковые последовательности чисел [7–9]. Современные детерминированные ГПСЧ имеют структуру П. Лекуера,

предложенную в 1994 году [8]: конечный набор состояний, вероятностное распределение в пространстве состояний, используемое для выбора начального состояния, функция перехода из одного состояния в другое, а также пространство выходных значений. Текущее состояние задаётся рекуррентной формулой. Числовые последовательности на выходе называют *псевдослучайные числа*. Из анализа структуры следует, что генерируемая последовательность ещё и периодическая, Период ГПСЧ есть наименьший порядковый номер генерируемого числа, после которого псевдослучайная последовательность начинает повторяться.

С позиции практического использования каждый ГПСЧ должен отвечать широко известным требованиям, а именно:

А) Каждое новое генерируемое число создаваемой псевдослучайной последовательности не должно быть очевидным и предсказуемым**.**

Б) Все числа ГПСЧ должен вырабатывать с одной и той же вероятностью**,** если последовательность генерируемых псевдослучайных чисел в идеале должна иметь равномерную *плотность* распределения вероятностей**.** Если некоторые числа генерируются чаще, чем другие, то это снижает степень их случайности и повышает уровень их предсказуемости.

В) Последовательность псевдослучайных чисел должна быть хорошо распределена по своему диапазону**.** А именно: маленькие, средние и большие числа должны вырабатываться настолько случайно, насколько это возможно. ГПСЧ, который вырабатывает, к примеру, сперва все маленькие числа, а затем все большие, генерирует весьма предсказуемую последовательность.

Г) Так как все ГПСЧ на практике являются циклическими, период этого циклического повторения чисел псевдослучайной последовательности в качественном ГПСЧ должен быть максимально большим.

Задача создания ГПСЧ, который вырабатывал бы строго равномерно распределённые числа, на сегодняшний день является весьма сложной.

Среди современных ГПСЧ широко известен генератор, именуемый *вихрь Мерсенна*. Алгоритм *вихрь Мерсенна* генерирует числовую последовательность с большим периодом (2^{19937}–1), и обладает рядом достоинств в сравнении с ГПСЧ, использующими линейный конгруэнтный метод [9].

Важно отметить, что все широко известные подходы к генерации равномерно распределённых псевдослучайных чисел позволяют сформировать близкое к равномерному распределение только при достаточно большой длине выборки. Хорошие результаты с позиции соответствия равномерному закону распределения бывают, например, у выборки длиной десять тысяч (10000) шагов и более. Однако имеется много прикладных задач обработки информации и управления, когда длина выборки составляет менее тысячи шагов (<1000) (две-три сотни, например, или немногим более). Это так называемые терминальные задачи [10, 11].

Терминальными системами управления называют динамические системы, переходящие из некоторого начального состояния в заданное конечное состояние. Время перехода из одного состояния в другое конечно (заранее известно или определяется в реальном времени из каких-либо дополнительных соображений) [10, 11]. Команды управления, как правило, формируются из текущих неполных и неточных измерений. Так как эти измерения подлежат обработке, то, в результате, рассматривают терминальные системы обработки информации и управления.

К терминальным системам обработки информации и управления могут быть отнесены автоматические системы разгона и торможения транспортных объектов, системы сближения, парковки и маневрирования движущихся объектов, и т.п. Работа подобных систем имеет чётко выраженные начало и конец, в отличие, например, от систем автоматического регулирования. Время функционирования терминальных систем часто бывает очень коротким. Так как в цифровых системах дискретное время измеряется в целых шагах, то количество шагов по времени может ограничиваться всего несколькими сотнями. Точность характеристик тех псевдослучайных процессов, которые моделируют случайные воздействия на систему и случайные помехи измерений, в этом случае может оказаться недостаточной, если используются общеизвестные встроенные ГПСЧ.

Наиболее наглядно невысокое качество общеизвестных и широко распространённых ГПСЧ в терминальных задачах можно увидеть на гистограммах генерируемых ими распределений псевдослучайных чисел. Согласно стандарту ГОСТ Р 50779.10–2000 (Статистические методы. Вероятность и основы статистики. Термины и определения) **гистограмма** «**…** есть графическое представление распределения частот для дискретной случайной величины, образуемое набором столбцов равной ширины, высоты которых пропорциональны частотам появления чисел в заданных интервалах…». Сравнение разных гистограмм реальных псевдослучайных числовых последовательностей с их идеальными аналогами позволяет сделать вывод о качестве функционирования изучаемых ГПСЧ. Это сравнение

Таблица 1. Расчёт количества интервалов у гистограмм по формулам (1) — (5)

N	64	128	256	512	1024	2048	4096	8192	16384	32768	65536	131072
n_1	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
$[N/n_1]$	9	16	28	51	93	170	315	585	1092	2048	3855	7281
n_2	**8**	**11**	**16**	**22**	**32**	**45**	**64**	**90**	**128**	**181**	**256**	**362**
$[N/n_2]$	8	11	16	23	32	45	64	91	128	181	256	362
n_3	**5**	**10**	**10**	**10**	**15**	**15**	**15**	**15**	**20**	**20**	**20**	**25**
$[N/n_3]$	12	12	25	51	68	136	273	546	819	1638	3276	5242
n_4	**4**	**8**	**8**	**8**	**12**	**12**	**12**	**12**	**16**	**16**	**16**	**20**
$[N/n_4]$	16	16	32	64	85	170	341	682	1024	2048	4096	6553
n_5	-	**5**	**5**	**5**	**10**	**10**	**10**	**10**	**15**	**15**	**15**	**20**
$[N/n_5]$	**-**	25	51	102	102	204	409	819	1092	2184	4369	6553

фактически является одним из возможных критериев согласия, используемых для качественной проверки того, что генератор работает правильно.

Критерий согласия, как известно, есть такое статистическое правило, по которому на основе имеющейся выборки некоей случайной величины принимается гипотеза о том, что эта исследуемая случайная величина подчиняется заданному закону распределения. Этот метод позволяет оценить степень различий между полученным при статистических испытаниях фактическим количеством попаданий числа в заданный интервал и их теоретически ожидаемым количеством [12, 13].

Так как построенная в результате статистических испытаний гистограмма есть лишь приближённое отражение плотности распределения, то для построения гистограмм важен выбор наилучшего разбиения: при увеличении ширины интервалов разбиения вдоль оси абсцисс снижается детализация определения плотности распределения, а при уменьшении ширины этих интервалов падает точность вычисленного значения плотности распределения.

Существуют разные правила выбора количества интервалов n:

в [9, 14, 15] применяется правило Стёрджеса

$$n_1 = 1 + [log_2 N], \quad (1)$$

в [14, 16] рекомендуют формулу «квадратного корня»

$$n_2 = [\sqrt{N}] \quad (2)$$

в [14, 17] рекомендуется формула Брукса и Каррузера

$$n_3 = 5 \cdot [lg\ N], \quad (3)$$

в [14, 18] Таушанов З., Тонева Е., Пенова Р. предлагают

$$n_4 = 4 \cdot [lg\ N], \quad (4)$$

в [14, 19] Тонева Е. рассматривает формулу

$$n_5 = 5 \cdot [lg\ N] - 5, \quad (5)$$

где — N общее число наблюдений величины, а $[x]$ — целая часть числа x.

Таблица 1 построена по формулам (1)-(5). В ней показано, на какое число интервалов n_i следует делить гистограмму, и сколько (в среднем) попаданий $[N/n_i]$ псевдослучайных чисел в каждый интервал, где i — номер формулы. Здесь число N задаётся как натуральная степень числа два последовательно от 2^6 до 2^{17}:

Общей для всех правил является тенденция увеличения среднего количества попаданий в интервал $[N/n_i]$ с ростом размера выборки. Видно, что формулы (4) и (5) плохо подходят для самых коротких выборок. Правило «квадратного корня» всегда предлагает самое большое количество интервалов и, по этой причине, лучше подходит для самых коротких выборок (N=100 или N=225), а при большом числе N не очень удобно. При увеличении N до 500 или до 1000 удобно использовать правило Стёрджеса или формулу Брукса и Каррузера. Для N свыше 1000 и до 10000 уже подходят формулы (4) и (5). Ближе к 100000 все формулы, кроме «квадратного корня» дают очень близкие результаты.

Общеизвестно и очевидно: увеличение количества испытаний N оказывает сглаживающее влияние на характер гистограммы. В этом случае уже не столь важно, какое было применено правило выбора количества интервалов. Поэтому в научной литературе часто приводят примеры гистограмм без указания того, какое правило выбора количества интервалов использовалось. Однако в терминальных задачах длина выборки невелика и правило выбора числа интервалов весьма важно.

Таблица 2. Предлагаемое количество интервалов у гистограмм

N	40 - 100	100 - 500	500 - 1000	1000 - 10000
n	7 - 9	8 - 12	10 - 16	12 - 22

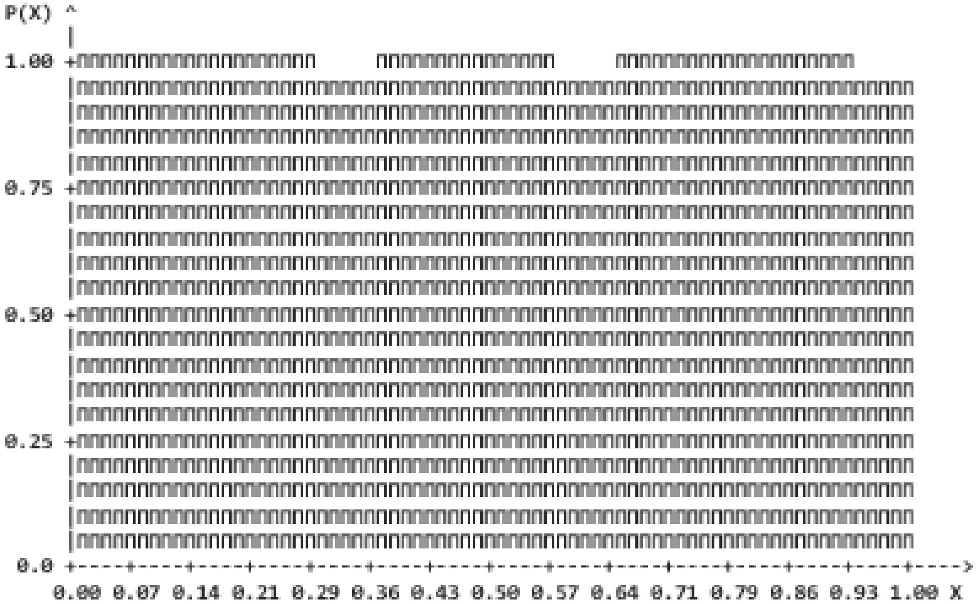

Рис. 1. N=16384, ГПСЧ rand — правило Стёрджеса

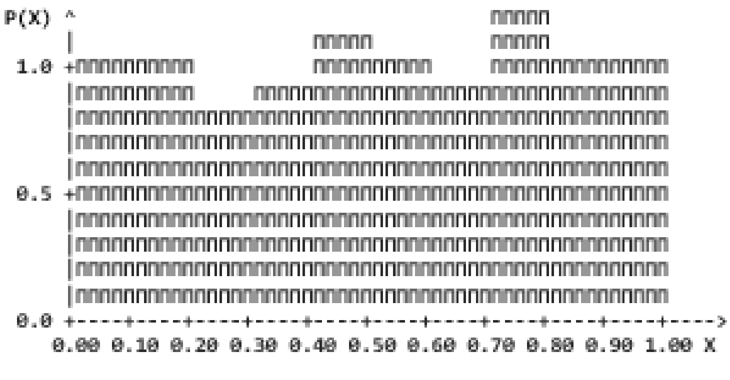

Рис. 2. N=1024, ГПСЧ mersenne — правило Стёрджеса

Поскольку нет однозначного критерия для выбора количества интервалов у гистограммы, вычисленная величина n на практике должна быть принята лишь как приблизительная, неточная, оценочная. Крайние значения, представленные первой таблицей, демонстрируют возможный разброс характеристик гистограмм у одной и той же выборки. Ведь точно определить можно только порядки размеров таких интервалов. Так, в [14] представлены иные соотношения для определения числа n:

Детальный анализ выборок требует строить гистограммы по тем правилам, результаты применения которых более наглядны, учитывая их приближённость. Чаще это правило Стёрджеса (1) и правило «квадратного корня» (2).

Общеизвестные ГПСЧ, использующие линейный конгруэнтный метод, ниже будут называться rand, а ГПСЧ «вихрь Мерсенна», соответственно, mersenne.

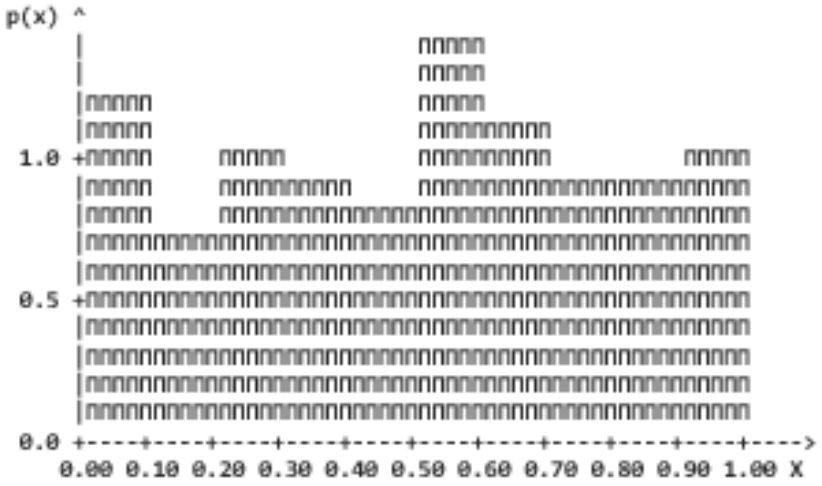

Рис. 3. N=512, ГПСЧ rand — правило Стёрджеса

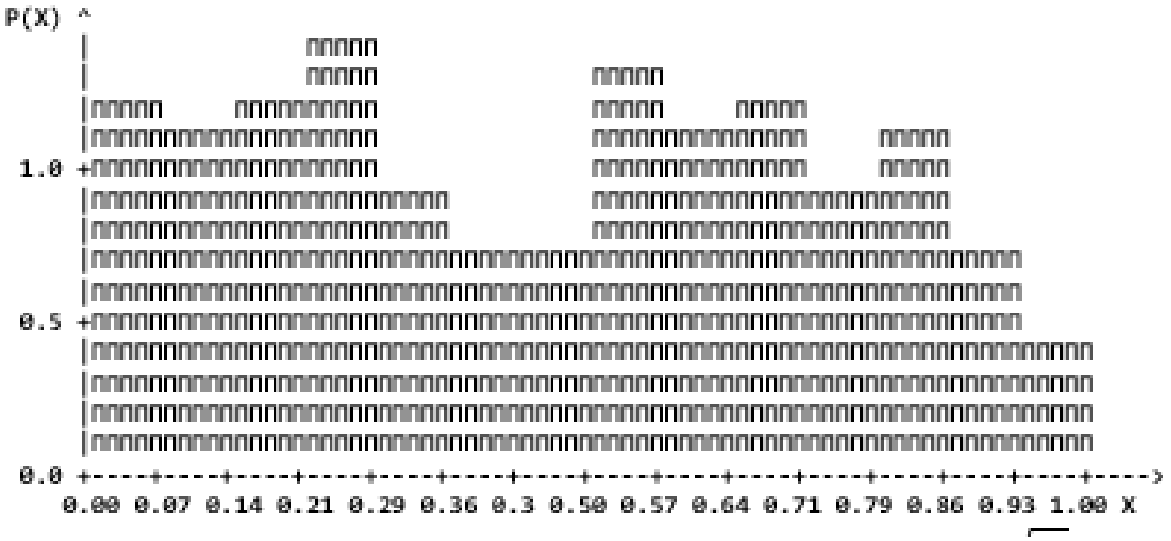

Рис. 4. N=225, ГПСЧ rand. Разбиение на √N

Гистограммы, построенные по реализациям последовательностей чисел, выработанным генераторами rand и mersenne, демонстрируют хорошее качество генерации равномерно распределенных чисел при N более 1000. Гистограмма на Рис. 1 демонстрирует равномерное распределение высокого качества при N= 16384 (2^{14}). Однако, при уменьшении числа N до 1000 и ниже, гистограммы существенно меняются. Распределения постепенно перестают быть похожими на равномерные и, фактически, становятся хоть и ограниченными, но совсем не равномерными. Последовательное изменение характера распределений можно увидеть на Рис. 1–6, переходя от 1-го к 6-му. Если N>500, применяется правило Стёрджеса. Если же N мало, то используется правило «квадратного корня», обозначаемое $\sqrt{N}$.

Гистограммы на Рис. 1–6 однозначно демонстрируют, что тип ГПСЧ, rand или mersenne, не влияет на качество распределений при малых выборках.

Из этих равномерных распределений создают гауссовские на основе ЦПТ и преобразования Бокса-Мюллера [4, 6]. Для ЦПТ берут 12 ГПСЧ rand или mersenne. Результирующее число есть сумма чисел, вырабатываемых каждым из 12-и ГПСЧ.

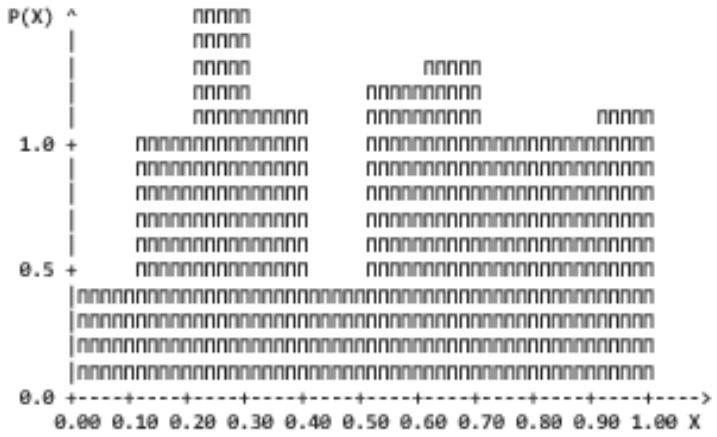

Рис. 5. N=100, ГПСЧ rand. Разбиение на √N

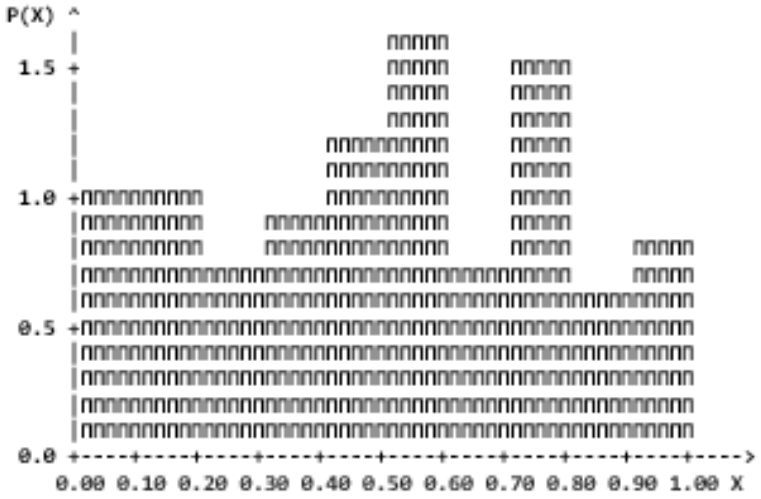

Рис. 6. N=100, ГПСЧ mersenne. Разбиение на √N.

Преобразование Бокса-Мюллера (ПБМ) использует два ГПСЧ типа rand или mersenne, создающие равномерно распределённые числа φ и R в интервале [0; 1). Результирующее число x вычисляют по формулам: $x = \sin(2\pi\varphi)\sqrt{-2\ln R}$ (синусное ПБМ) и $x = \cos(2\pi\varphi)\sqrt{-2\ln R}$ (косинусное ПБМ).

Переходя от Рис. 7–9 к Рис. 10–12 можно видеть, как ухудшается качество гистограмм нормального распределения с уменьшением количества чисел N.

Приведённые на Рис. 13 и на Рис. 14 гистограммы экспоненциального распределения, полученные из равномерного распределения методом обратной функции [5], также качественно ухудшаются с уменьшением количества чисел N.

Качественный анализ гистограмм может сопровождаться количественным с помощью критерия согласия. В качестве критерия согласия может быть использован критерий χ^2 (критерий Пирсона). Формула критерия согласия χ^2 Пирсона является взвешенной суммой квадратов уклонений реальных частот

$\frac{m_i}{m}$ от их предполагаемых значений p_i [14]:

$\chi 2 = m \cdot \sum_{i=1}^{n} \frac{\left(\frac{m_i}{m} - p_i\right)^2}{p_i}$,где m = [N/n].

Чем меньше χ^2, тем точнее выборка согласуется с предполагаемой.

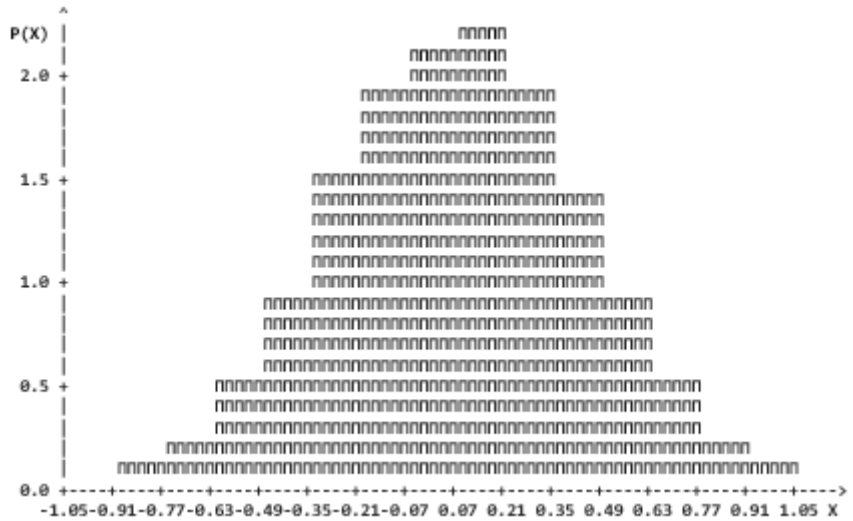

Рис. 7. N=16384. ЦПТ. Правило Стёрджеса

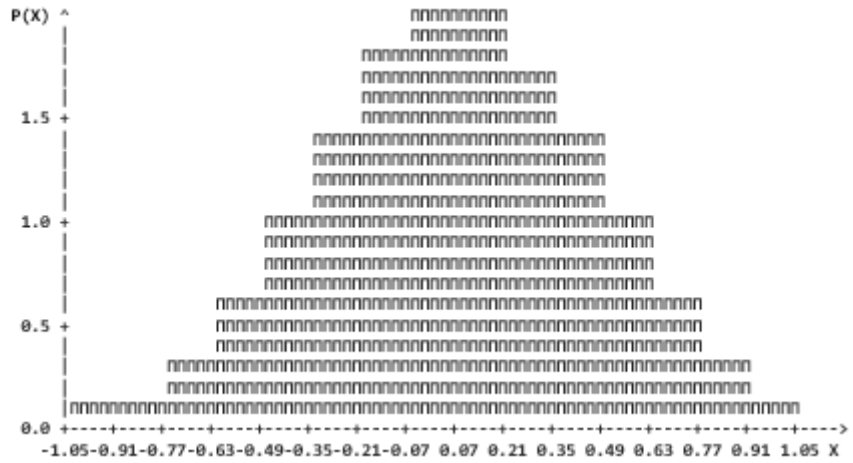

Рис. 8. N=16384. Синусное ПБМ. Правило Стёрджеса

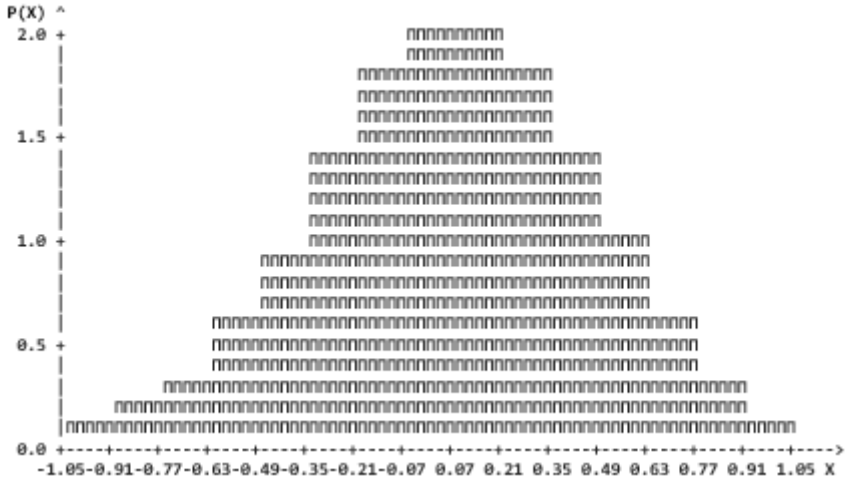

Рис. 9. N=16384. Косинусное ПБМ. Правило Стёрджеса

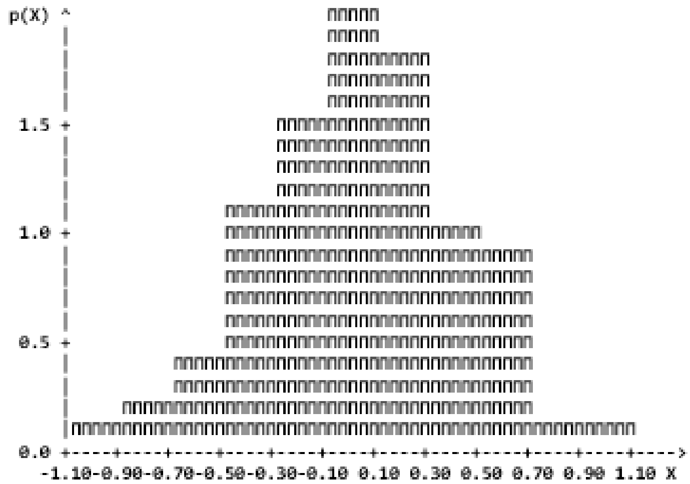

Рис. 10. N=100. ГПСЧ rand. ЦПТ. Разбиение на √N.

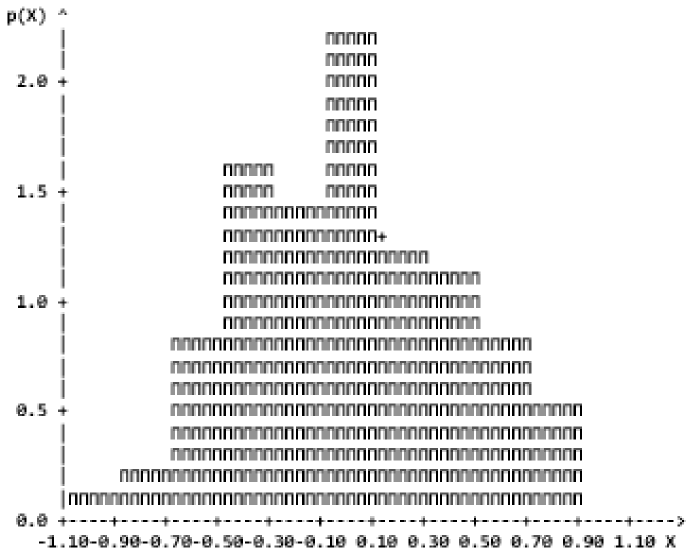

Рис. 11. N=100. ГПСЧ rand. Синусное ПБМ. Разбиение на √N.

Для улучшения совсем не идеальных (на коротких выборках) характеристик генераторов типа rand или mersenne предлагается следующий подход.

I) Непрерывная функция распределения F(x) случайного числа x, заданная на непрерывном аргументе x, заменяется на дискретную функцию M(k), заданную на целочисленном аргументе k. Функция M(k) аппроксимирует функцию F(x): все значения чисел M(k) для каждого k совпадают со значениями функции F(x) в каждой точке k·Δx, где Δx — интервал дискретизации непрерывной области определения функции F(x), а k — целое число с заданным конечным диапазоном.

II) Конечное дискретное множество значений M(k) сохраняется в программе в виде массива, списка или какой-либо иной структуры.

После этого стартует следующий алгоритм:

1. На вход алгоритма поступает псевдослучайное число, вырабатываемое каким-либо встроенным ГПСЧ.
2. Поступившее на вход алгоритма псевдослучайное число последовательно сравнивается со всеми числами из M(k), пока не будет найдено ближайшее к поступившему число из M(k).
3. После этого найденное число из M(k) поступает на выход алгоритма.

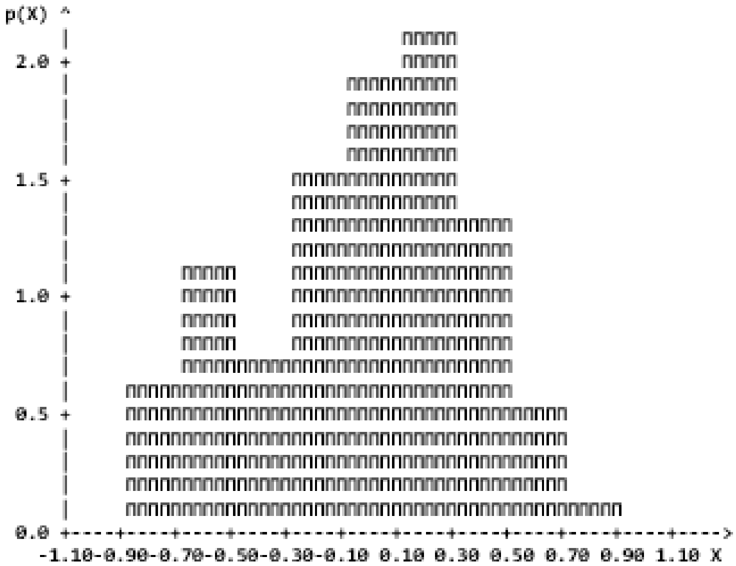

Рис. 12. N=100. ГПСЧ rand. Косинусное ПБМ. Разбиение на √N

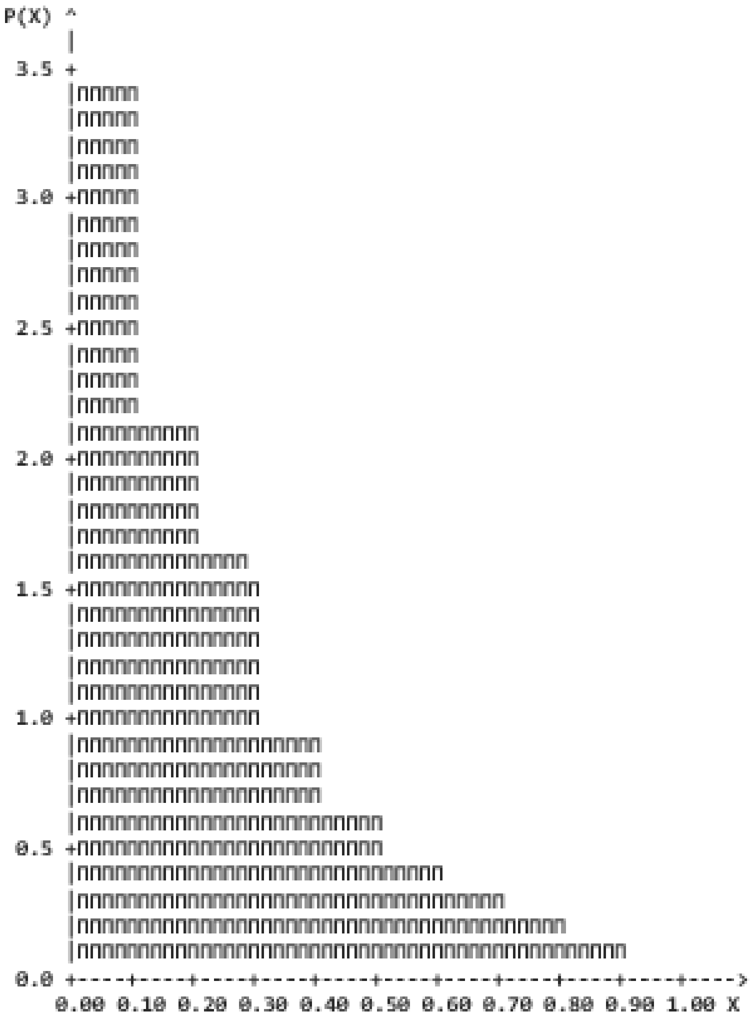

Рис. 13. N=1000 rand. Экспоненциальное распределение. Правило Стёрджеса

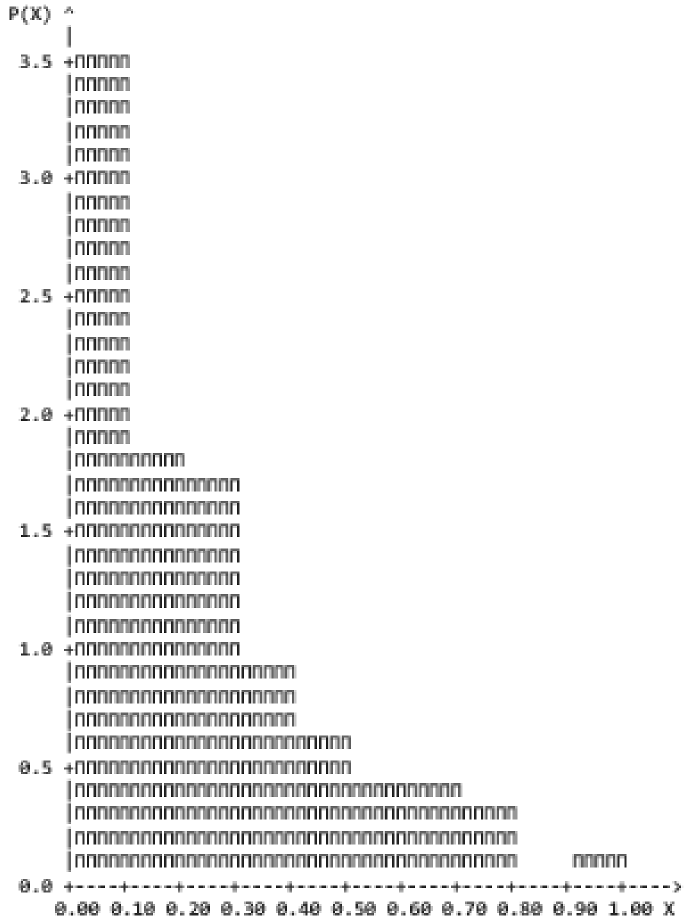

Рис. 14. N=100 rand. Разбиение на √N. Экспоненциальное распределение.

4. Из дискретной функции M(k) это найденное число удаляется.
5. Если множество M(k) после удаления элемента не стало пустым, то выполняется переход к шагу (1).
6. Иначе, если множество M(k) стало пустым, то выполняется завершение этого алгоритма.

Представленный подход требует детальных комментариев.

В (I) непрерывная функция распределения F(x) с непрерывным аргументом x соответствует базовым положениям теории вероятностей, но, фактически, в цифровом компьютере как сама функция распределения F(x), так и её аргумент x, заданы с конечной точностью. Как область определения, так и множество значений функции распределения фактически дискретны и ограничены. Таким образом, дискретная функция M(k) аппроксимирует, на самом деле, дискретную же функцию F(x). Только мощность множества, на котором определятся аргумент k, на порядки меньше мощности множества, на котором определятся аргумент x.

Псевдослучайное число, поступающее на вход алгоритма на шаге (1), создаётся встроенными ГПСЧ типа rand, mersenne и т.п.

Каждое поступившее на вход псевдослучайное число «выбивает» из дискретного множества M(k) ближайшее к нему число. Множество M(k) последовательно уменьшается с каждым новым поступившим на вход алгоритма, псевдослучайным числом. Ни одно число на выходе не повторяется дважды, и ни одно число из M(k) не бывает пропущено, что приводит к строго предсказуемому результату: в момент окончания работы алгоритма все числа на выходе строго соответствуют заданной в M(k) функции распределения. Эта гистограмма распределения, построенная на основе всех чисел, полученных на выходе алгоритма, будет точно соответствовать гистограмме требуемого распределения при любом заданном правиле выбора количества интервалов.

Дискретная функция M(k) фактически есть маска, накладываемая на входную псевдослучайную последова-

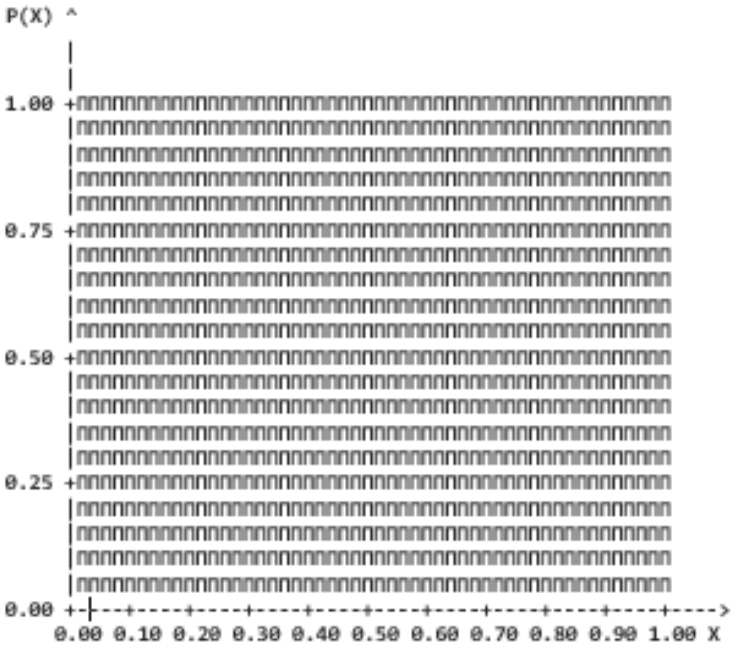

Рис. 15. N=100. Улучшенный ГПСЧ. Разбиение на √N

тельность с целью получить заданное распределение. Входная последовательность, считающаяся «равномерной», формируется традиционно, как во всех известных ГПСЧ, когда множество значений числа на каждом шаге по времени неизменно и каждое новое псевдослучайное число вырабатывается с одной и той же вероятностью, как в требовании (Б). Этот принцип формирования числовой последовательности можно сравнить с выпадением чисел в игре «Рулетка», когда каждое новое выпадающее число никак не связано с предыдущим, а множество значений чисел постоянно на каждом шаге. Предложенный же автором подход можно сравнить с игрой «Лото», когда после каждого нового выпавшего числа множество оставшихся значений сокращается, а все оставшиеся числа на следующем шаге остаются равновероятными. При игре в «Лото» каждое число за один цикл встречается ровно один раз, хотя порядок выпадения чисел случаен. Предложенный приём позволяет создавать последовательность чисел, реализующую любую желаемую функцию случайного распределения.

Важно рассмотреть одно возможное нарушение содержания пункта (А) требований к ГПСЧ о том, что «…каждое новое генерируемое число псевдослучайной последовательности не должно быть очевидным и предсказуемым…». Действительно, на последнем шаге работы алгоритма возвращается последнее оставшееся число из множества M(k). После предпоследнего шага работы алгоритма можно однозначно сказать, какое число будет последним, так как каждое число встречается ровно один раз. Но эта определённость наступает только на предпоследнем шаге. Исключив, для достижения «теоретической строгости», последний шаг из работы алгоритма, нельзя обнаружить заметного изменения гистограммы распределения.

Также нуждается комментариях другое возможное нарушение пункта (Б) требований к ГПСЧ «…*все числа* должны *вырабатываться* с одной и той же вероятностью…», так как вероятность каждого нового числа возрастает в сравнении с вероятностью каждого предыдущего числа последовательности из-за пошагового сокращения множества M(k). Но важно заметить, что эта возрастающая вероятность не привязана к каждому конкретному числу, а на каждом новом шаге все оставшиеся числа равновероятны. Строго говоря, при терминальной постановке задачи вероятность появления каждого нового числа возрастает и при работе ГПСЧ по принципу игры «Рулетка».

Удаление найденного числа из дискретной функции M(k), указанное в пункте (4), может быть осуществлено разными способами. Это может быть реальное удаление с перенумерованием массива M(k) и сокращением размера M(k). Либо удаляемое число заменяется на какое-нибудь запрещённое значение, пропускаемое при поиске. Размер M(k) не уменьшится, но упростится работа алгоритма.

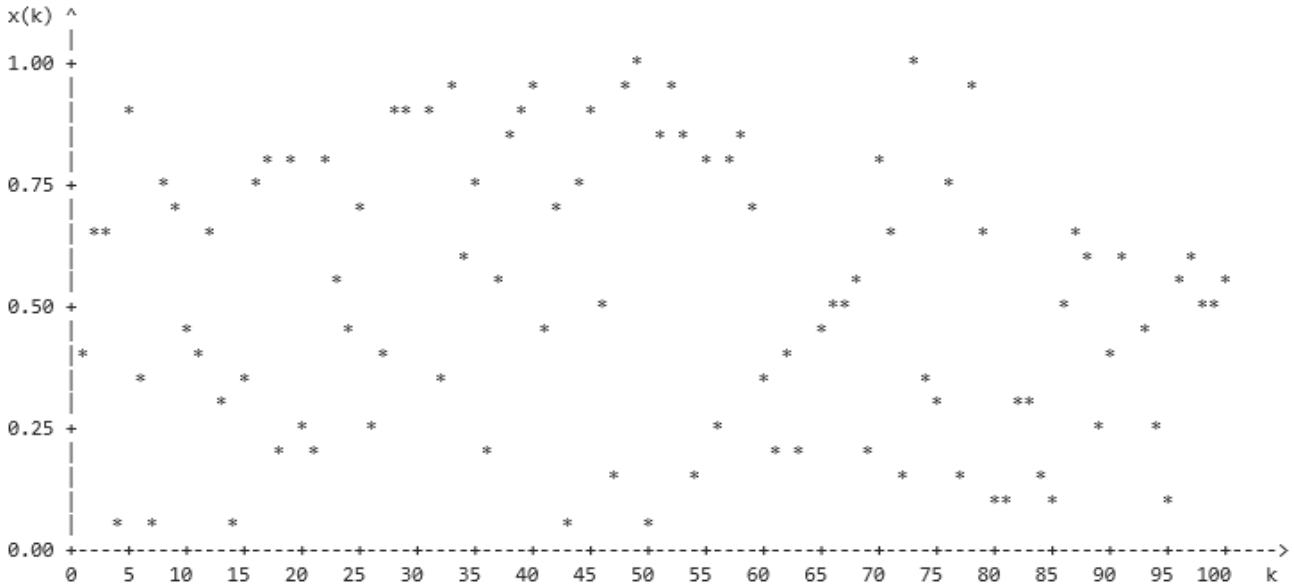

Рис. 16. N=100. Псевдослучайная последовательность улучшенного ГПСЧ.

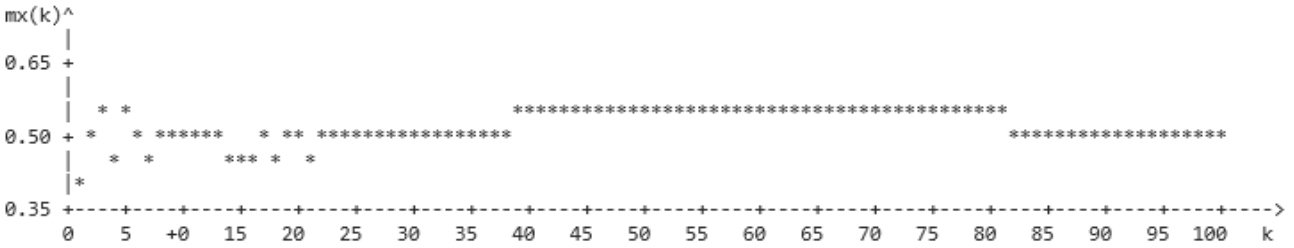

Рис. 17. N=100. Псевдослучайная последовательность улучшенного ГПСЧ.

Вместо проверки в (5) и (6) того, что множество M(k) не стало пустым, можно провести проверку того, что достигнут последний элемент массива, то есть, что счётчик по k достиг максимальной заданной величины.

Строго говоря, не важно, какая именно псевдослучайная последовательность поступает на вход алгоритма (1)-(6). Важно лишь, чтобы она не была монотонной. Тогда и формируемая алгоритмом числовая последовательность будет иметь «случайный» характер. Но необходимо следить, чтобы выполнялся пункт (В) требований к ГПСЧ, то есть чтобы псевдослучайная последовательность чисел на выходе алгоритма была бы хорошо распределена по своему диапазону. Для проверки этого полезно изучить характер графика сгенерированного псевдослучайного процесса. Если на последних шагах график остаётся «случайным», то условие (В) выполнено. Если же на последних шагах график становится монотонным, то следует каким-либо образом скорректировать входную «случайную последовательность, чтобы выполнить условие (В).

Например, можно вычислять на каждом шаге текущее математическое ожидание псевдослучайной числа на выходе алгоритма и корректировать с его учётом очередное входное значение, предупреждая возможное стремление псевдослучайной последовательности в сторону монотонного изменения, например, монотонного движения к одной из границ интервала распределения.

Пункт (Г) требований к ГПСЧ (о цикличности) в данном случае не имеет существенного значения, так как для терминальных задач достаточно, чтобы период ГПСЧ не был меньше всего времени функционирования терминальной системы. Если период ГПСЧ превышает длину выборки, то цикличность не успеет проявиться и не повлияет на результат работы алгоритма.

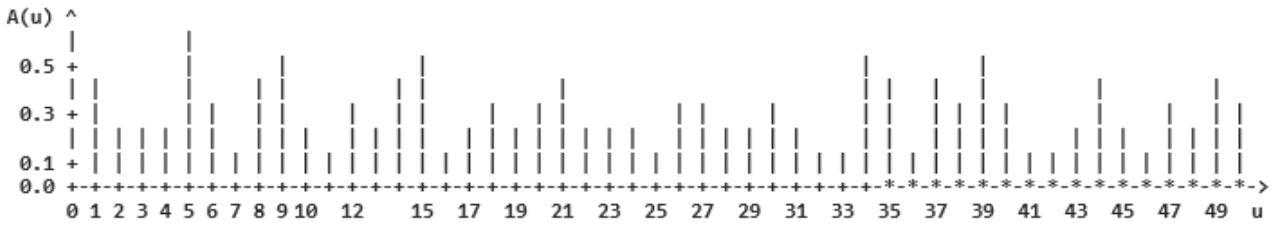

Рис. 18. N=100. Частотный спектр последовательности улучшенного ГПСЧ.

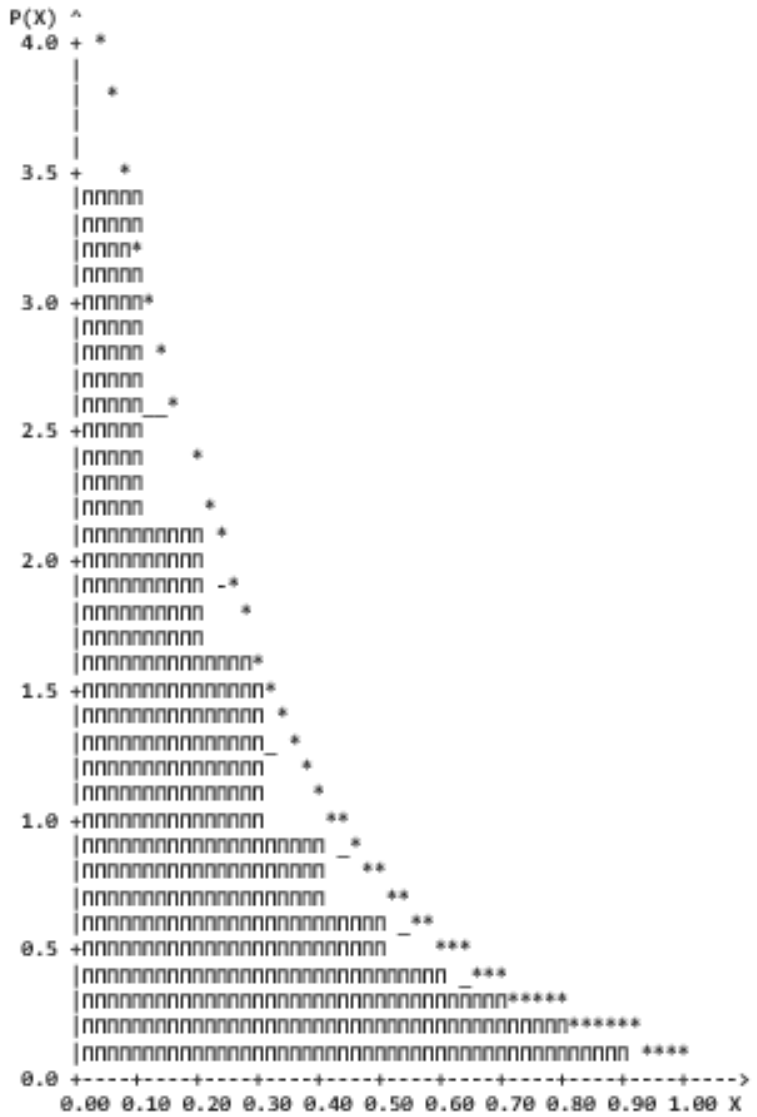

Рис. 19. N=100. ГПСЧ улучшен. Экспоненциальное распределение. Правило √N.

Время работы моделируемой терминальной системы (количество шагов по времени) должно совпадать с максимальным количеством шагов работы предлагаемого алгоритма. Это наилучшее условие, реализующее этот подход. Если же время работы терминальной системы дано приближённо, возможно сделать так, чтобы время работы системы было больше, чем время работы алгоритма. Так как алгоритм может перезапускаться и продолжать успешное функционирование, неточно заданное время работы терминальной системы нужно

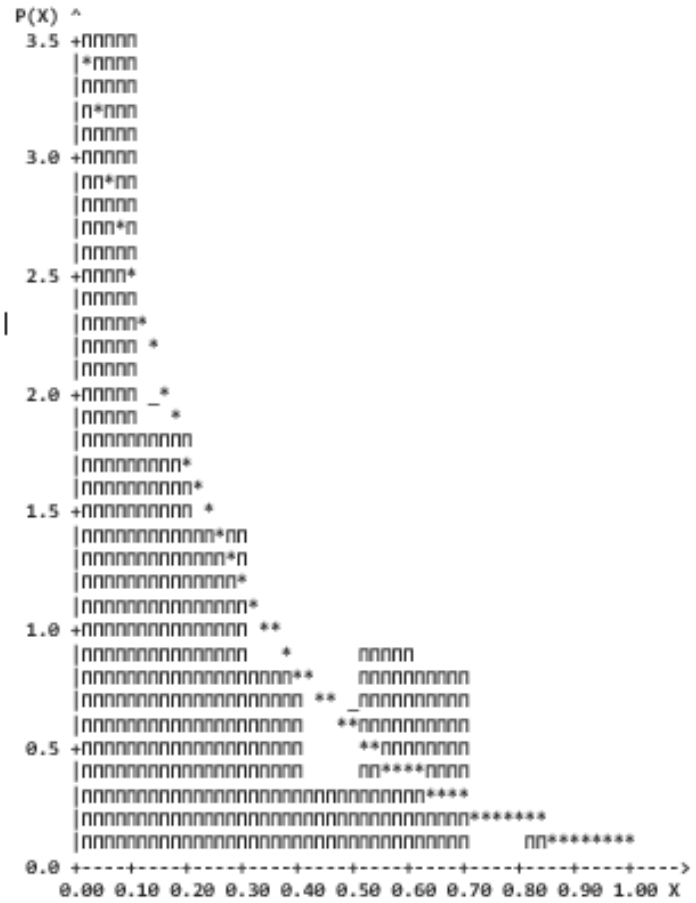

Рис. 20. N=100. ГПСЧ rand. Экспоненциальное распределение. Разбиение на √N

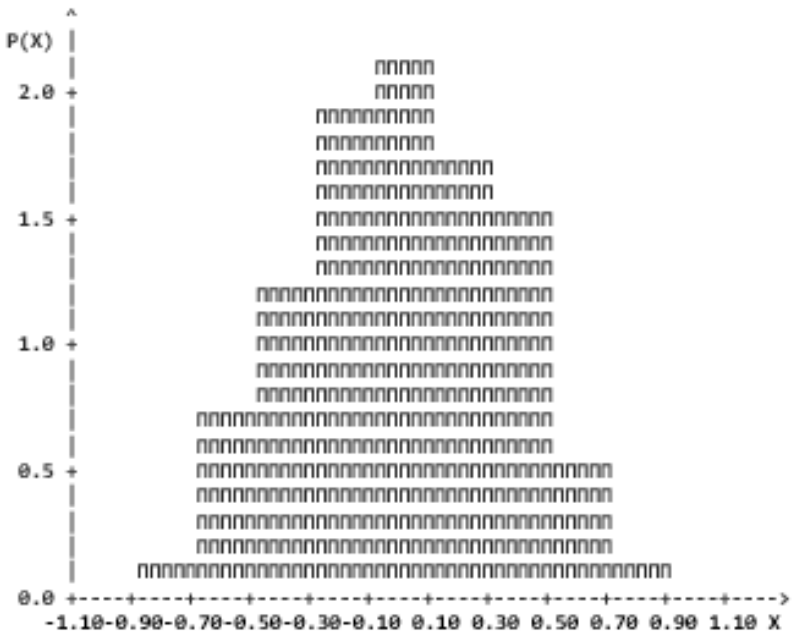

Рис. 21. N=100. Улучшенный ГПСЧ. ЦПТ. Разбиение на √N

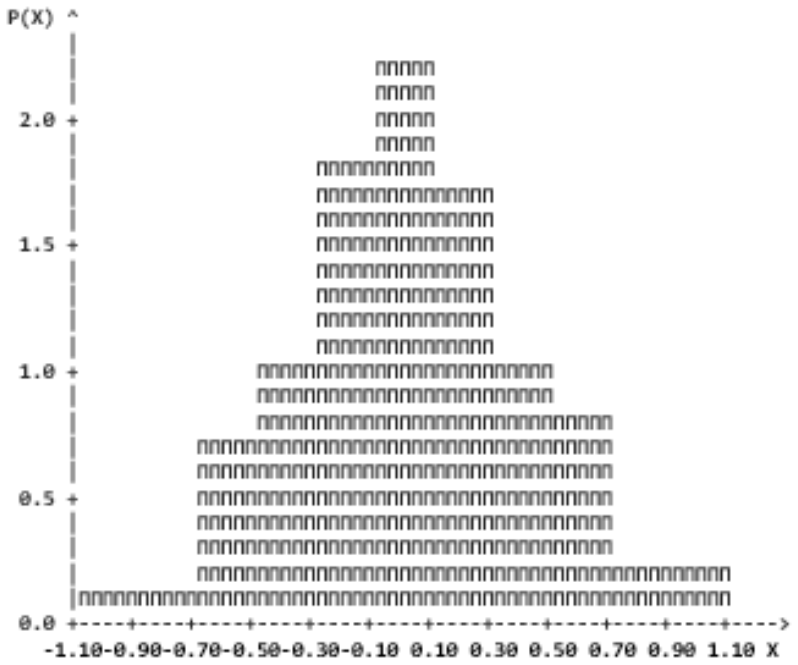

Рис. 22. N=100. Улучшенный ГПСЧ. Синусное ПБМ. Разбиение на √N

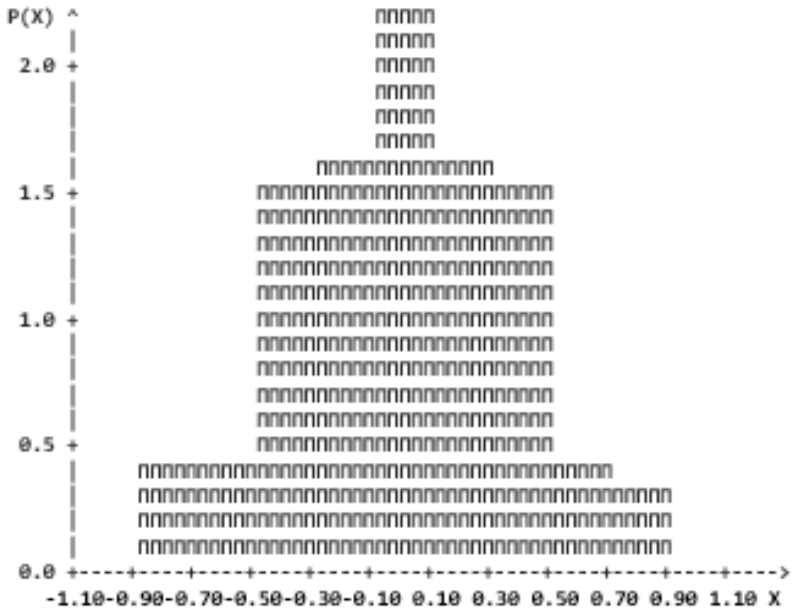

Рис. 23. N=100. Улучшенный ГПСЧ. Косинусное ПБМ. Разбиение на √N

подобрать так, чтобы оно в несколько раз превышало время работы алгоритма, пусть даже приближённо. Если время работы системы кратно одному циклу алгоритма, то характеристики ГПСЧ не изменятся. Однако и в случае, когда строгой кратности добиться не удаётся, характеристики ГПСЧ не будут существенно искажаться, ведь основной вклад в гистограмму внесут несколько целых интервалов (в крайнем случае один целый интервал).

Гистограмма на Рис. 15 отражает точность подхода (см. Рис. 1 и Рис. 5–6).

На Рис. 16 показан график псевдослучайной последовательности, которая создана улучшенным ГПСЧ и соответствует гистограмме на Рис. 15. Видно, что график x(k) имеет случайный характер несмотря на идеально ровную гистограмму этой последовательности. Внешний вид графика на Рис. 16 никак не указывает на то, что каждое число x(k) в исследуемой выборке встречается ровно один раз. Замечательно также то, что идеальный характер гистограммы на Рис. 15 никак не изменится при использовании любого из пяти вышеназванных правил разбиения.

Исследование характеристик улучшенного ГПСЧ с идеальным равномерным распределением показывает, что можно добиться качественного улучшения распределений на коротких выборках, если эти распределения будут созданы из равномерных вышеназванным методом обратной функции [5]. На Рис. 17 график текущего математического ожидания последовательности с Рис. 16, возможность использования которого описана выше в комментариях к предложенному подходу.

Для более полного представления о псевдослучайной последовательности

следует выполнить её частотный анализ. Формула амплитудного спектра A(u):

$$A(u) = \sqrt{Re(F(u))^2 + Im(F(u))^2},$$

где

$$Re(F(u)) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x(k) \cos\left(\frac{2\pi k u}{N}\right),$$

$$Im(F(u)) = -\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x(k) \sin\left(\frac{2\pi k u}{N}\right),$$

u — частота от 0 до N–1, x(k) — числовая последовательность, а k — шаг по времени.

На Рис. 18 показан амплитудный спектр улучшенной последовательности, график которой приведён на Рис. 16. Качественный анализ спектра позволяет сделать вывод, что, хотя этот спектр и не является спектром дискретного белого шума, таким спектром обладают большинство равномерных ГПСЧ. В спектре на Рис. 18 нет преимущественных участков. Он не является ни красным, ни синим, ни зелёным, ни серым, ни каким-либо иным типовым цветным спектром.

Улучшенный генератор позволяет повысить качество создаваемых на его основе распределений. Гистограмма на Рис. 19 демонстрирует высокое качество распределения, полученного вычислением экспоненциально распределённого числа y из равномерно распределённого числа x по формуле

$$y = -\frac{ln|1-x|}{\lambda}.$$

Поверх гистограммы построен график функции плотности распределения $p(x) = \lambda e^{-\lambda x}$.

Если сравнить Рис. 19 с Рис. 20, где представлена гистограмма такого же распределения, созданного тем же методом обратного преобразования, но из чисел, генерируемых ГПСЧ rand или mersenne, показанных на Рис. 5 и Рис. 6, то можно видеть, что на малой выборке эта гистограмма не отражает экспоненциальное распределение (не удовлетворяет критерию согласия).

Гистограмма нормального (гауссовского) распределения, созданного при использовании ЦПТ теории вероятностей [4] на основе улучшенного ГПСЧ (с идеальным равномерным распределением) показана на Рис. 21. В сравнении с гистограммой на Рис. 10 качество несколько лучше. Также существенно лучше, чем гистограммы на Рис. 11 и Рис. 12 (созданные на основе ГПСЧ rand и mersenne) гистограммы на Рис. 22 и Рис. 23, полученные на основе улучшенных ГПСЧ с помощью ПБМ. Во всех таких случаях применение улучшенных ГПСЧ приводит к повышению качества гистограмм, приближающемуся к идеальному. Гораздо лучший результат (почти идеальный) даёт использование вместе с улучшенным ГПСЧ метода обратного преобразования [5]. Гистограммы же, отражающие результаты использования ПБМ и ЦПТ вместе с улучшенным ГПСЧ, хоть и становятся несколько лучше, но остаются далёкими от идеала. Объяснить это можно тем, что метод обратного преобразования [5] использует только один псевдослучайный процесс, плотность распределения которого преобразуется из равномерной согласно заданной функции распределения. При применении же ЦПТ и ПБМ используются несколько псевдослучайных последовательностей в одной формуле. Дискретная аппроксимация непрерывной функции, указанная в (I), ухудшает результат применения ПБМ и ЦПТ, рассчитанных на взаимодействие двух и более случайных величин с непрерывными распределениями.

Недостаточно высокое качество полученных гауссовских распределений, генерируемых на основе улучшенных ГПСЧ (в сравнении с высоким качеством полученных гистограмм равномерного и экспоненциального распределений) не является фатальным, так как предлагаемый подход допускает создание любого желаемого распределения на основе ГПСЧ типа rand и mersenne путём наложения маски M(k)

(дискретной функции распределения), как было сказано выше в (I).

В итоге следует сказать, что представленный алгоритм (1) — (6) в рамках предлагаемого подхода (I) — (II) не является новым генератором псевдослучайных чисел. Это подход, улучшающий характеристики известных ГПСЧ в конкретных задачах. Его нельзя применять, например, для защиты информации в алгоритмах шифрования. Но предложенный подход весьма эффективен для очень точного моделирования коротких выборок случайных возмущений динамических систем и случайных помех измерений в терминальных задачах.

Ограничением подхода является требование дополнительной памяти для хранения значений дискретной аппроксимирующей функции M(k). Но, при малых выборках, это требования сколько-нибудь существенного влияния не оказывает.

Конструктивной является общая идея формирования и улучшения ГПСЧ не с общетеоретических позиций, а с точки зрения специфики конкретной задачи. Развивая эту идею, можно генерировать псевдослучайные числа с идеальным частотным спектром, аналогично числам с идеальной плотностью распределения.

ЛИТЕРАТУРА

1. Керниган Б., Ритчи Д. Язык программирования С.— М.: Вильямс, 2015, 304 с.
2. Гослинг Дж., Арнольд К. Язык программирования Java.— СПб.: Питер, 1997, 304 с.
3. Павловская Т.А. C#. Программирование на языке высокого уровня. Учебник для вузов.— СПб.: Питер, 2007, 432 с.
4. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и её инженерные приложения. 3-е изд.— М.: Издательский центр «Академия», 2003.— 464 с.
5. Сергиенко А.Б. Цифровая обработка сигналов. 3-е изд., перераб. и доп.— Издательство БХВ-Петербург, 2011.— 768 с.
6. Новиков П.В. Системы цифровой обработки сигналов: Учеб. пособие к лабораторным работам / Под ред. профессора, д-р. техн. наук О.М. Брехова.— М.: Изд-во МАИ, 2019.— 75 с.
7. Корн Т., Корн. Г. Справочник по математике.— М.: Наука, 1973.— 831 с.
8. L'Ecuyer P. Random Number Generation // Springer Handbooks of Computational Statistics: 2007.— С. 93–137.
9. M. Matsumoto, T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator// ACM Trans. on Modeling and Computer Simulations: journal. 1998.— Vol. 8, no. 1.— P. 3–30.
10. Фельдбаум А.А. О распределении корней характеристического уравнения системы регулирования. // Автоматика и Телемеханика, 1948, № 4, С. 253–279.
11. Lee E.B., Markus L. Foundations of Optimal Control Theory. New York, London: John Wily & Sons, 1967, 631 с.
12. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим.— М.: Издательство стандартов, 2006.— 87 с.
13. Кендалл М., Стьюарт А. Статистические выводы и связи.— М.: Наука, 1973.
14. Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия типа χ^2 //Заводская лаборатория. Диагностика материалов. 2003. Т. 69, вып. 1.— С. 61–67.
15. Sturgess H.A. The choice of classic intervals // J. Am. Statist. Assoc.— march 1926.— 47 p.
16. Heinhold I., Gaede K.W. Ingeniur statistic.— München; Wien, Springler Verlag, 1964/ — 352 s.
17. Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества.— М.: Мир, 1970.— 368 с.
18. Таушанов З., Тонева Е., Пенова Р. Вычисление энтропийного коэффициента при малых выборках//Изобретательство, стандартизация и качество, 1973. № 5.
19. Тонева Е. Аппроксимация распределений погрешности средств измерений // Измерительная техника, 1981. № 6.— С. 15–16.

© Новиков Павел Владимирович (novikov.mai@mail.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»