4. Статистическое моделирование систем на ЭВМ

4.1 Сущность метода.

Формирование квазиравномерно распределенной на (0, 1) случайной величины

Статистическое моделирование –
метод получения с помощью ЭВМ статистических данных о процессах, происходящих в моделируемой системе.

Для получения оценок характеристик моделируемой системы ***S*** с учетом воздействий внешней среды ***E*** статистические данные обрабатываются и классифицируются с использованием методов математической статистики.

Теоретическая база – предельные теоремы теории вероятностей.

Сущность метода статистического моделирования

Сущность метода –

построение некоторого моделирующего алгоритма, имитирующего процесс функционирования исследуемой системы с учетом случайных входных воздействий и воздействий внешней среды,

и реализация этого алгоритма с использованием программно-технических средств ЭВМ.

Результат статистического моделирования системы:

серия значений искомых величин или функций

статистическая обработка

сведения о поведении реального объекта или процесса в произвольные моменты времени

При достаточно большом количестве реализаций ***N*** результаты моделирования могут быть приняты в качестве оценок искомых характеристик процесса функционирования системы.

Две области применения метода статистического моделирования:

1) изучение стохастических систем;

2) решение детерминированных задач.

Основная идея – замена детерминированной задачи эквивалентной схемой некоторой стохастической системы, выходные характеристики которой совпадают с результатом решения детерминированной задачи.
При такой замене – приближенное решение; погрешность уменьшается с увеличением числа испытаний (реализаций моделирующего алгоритма) ***N***.

Примеры.

1. Функционирование стохастической системы $\boldsymbol{S_R}$ описывается соотношениями:

   $x = 1 - e^{-\lambda}$ – входное воздействие,

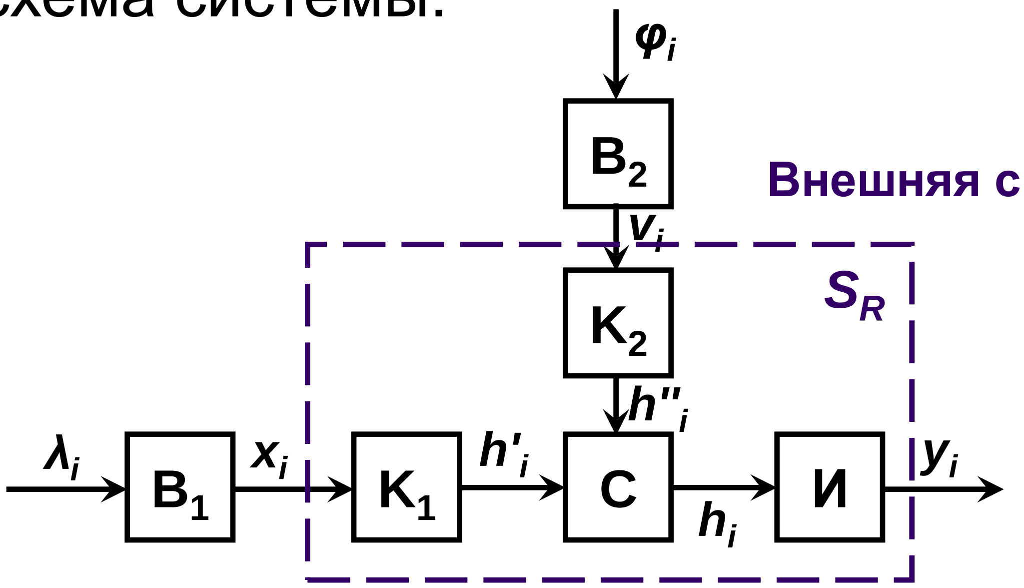
   $v = 1 - e^{-\varphi}$ – воздействие внешней среды,

   λ и φ – СВ с известными функциями распределения.

   Величина y определяется как $y = \sqrt{x^2 + v^2}$.

Требуется оценить среднее значение (мат. ожидание) $M[y]$.

Структурная схема системы:



В$_1$, **В$_2$** – вычисления $x_i = 1 - e^{-\lambda_i}$, $v_i = 1 - e^{-\varphi_i}$;

К$_1$, **К$_2$** – возведение в квадрат:

$$h'_i = (1 - e^{-\lambda_i})^2, \quad h''_i = (1 - e^{-\varphi_i})^2;$$

С – суммирование $h_i = (1 - e^{-\lambda_i})^2 + (1 - e^{-\varphi_i})^2$;

И – извлечение корня

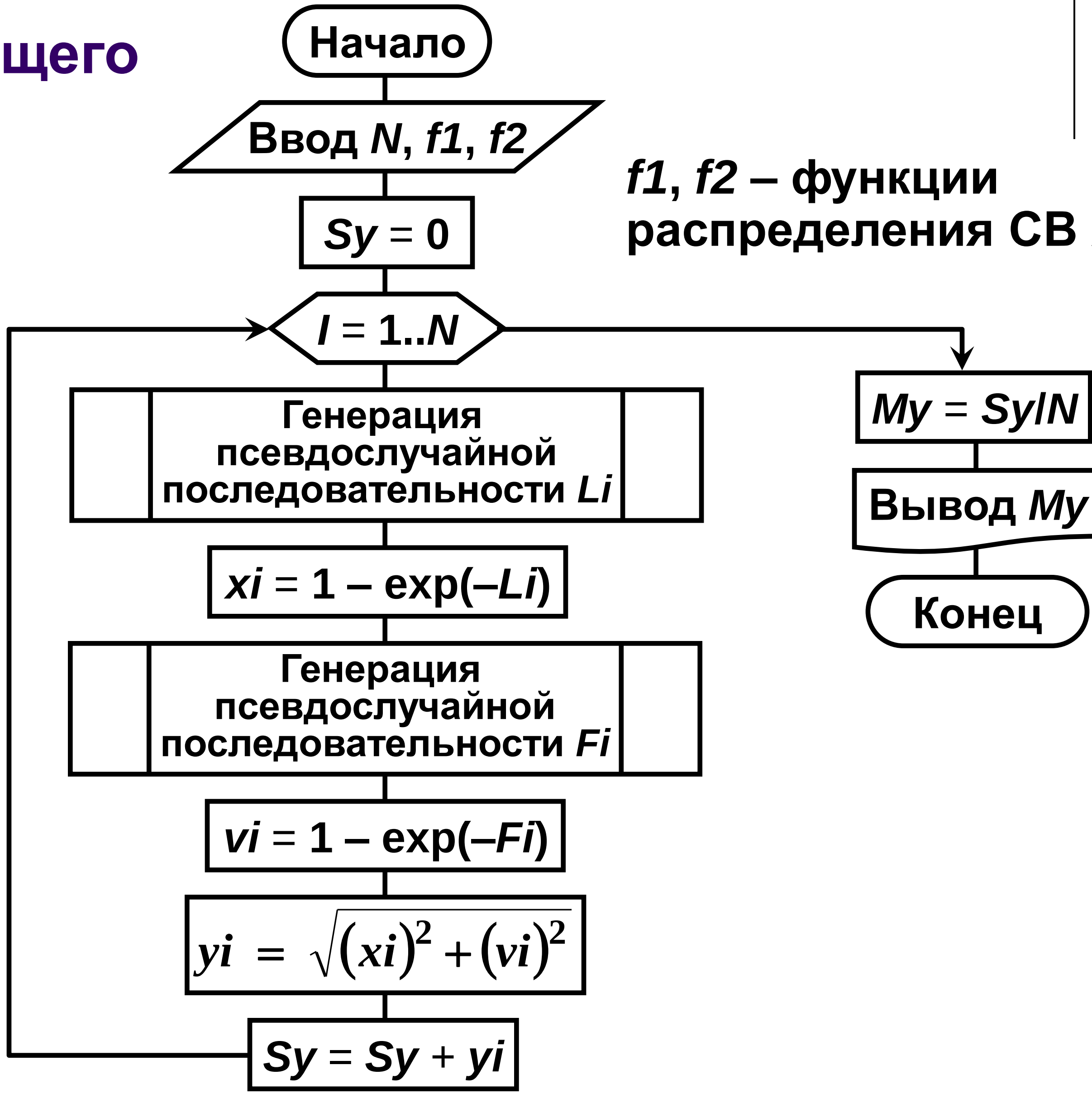
$$y_i = \sqrt{\left(1 - e^{-\lambda_i}\right)^2 + \left(1 - e^{-\varphi_i}\right)^2}\,.$$

Оценка мат. ожидания:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \ ,$$

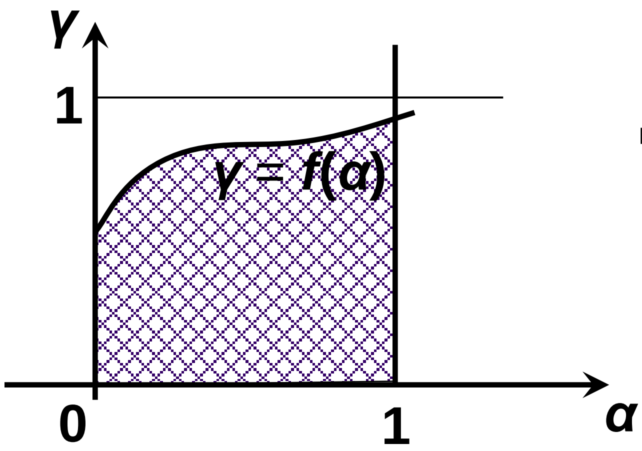
N – число реализаций, необходимое для статистической устойчивости результатов.

Блок-схема моделирующего алгоритма:



f1, ***f2*** **– функции распределения СВ *λ* и *φ***

2. Найти оценку площади фигуры, изображенной на чертеже.



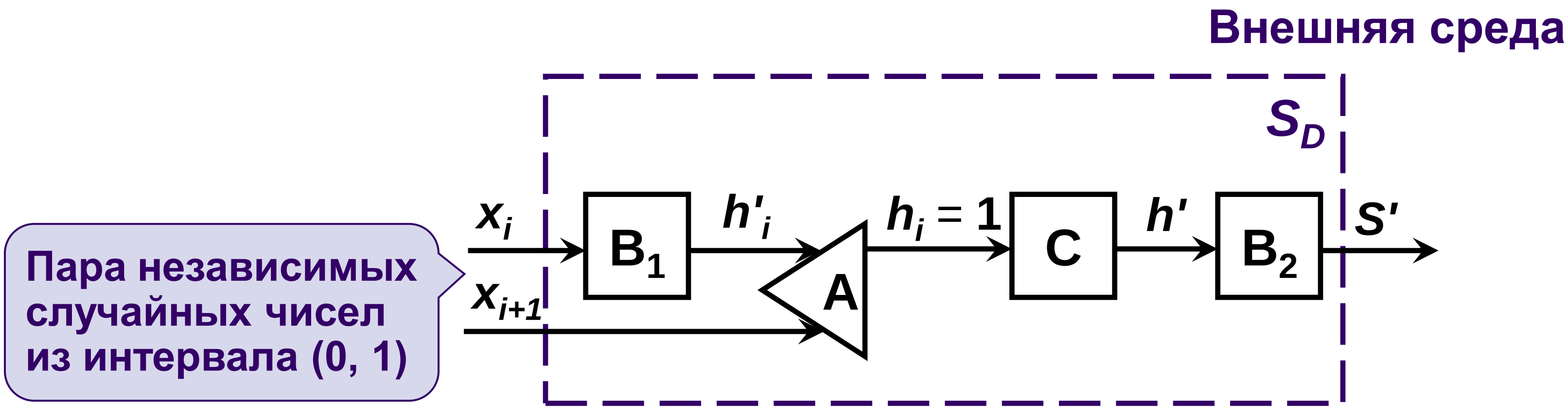
Для определенности предполагается, что $0 \le \boldsymbol{f}(\boldsymbol{\alpha}) \le 1$ при $0 \le \boldsymbol{\alpha} \le 1$.

Задача чисто детерминированная.

Аналитическое решение:

искомая площадь равна $S = \int_0^1 f(\alpha)d\alpha$.

Структурная схема адекватной по выходным характеристикам стохастической системы S_D :



B_1 – вычисление $h'_i = f(x_i)$,

A – анализ

$$h_i = \begin{cases} 1, & \text{если } x_{i+1} \le f(x_i), \\ 0 & \text{в противном случае,} \end{cases}$$

Точка с координатами (x_i, x_{i+1}) лежит ниже графика?

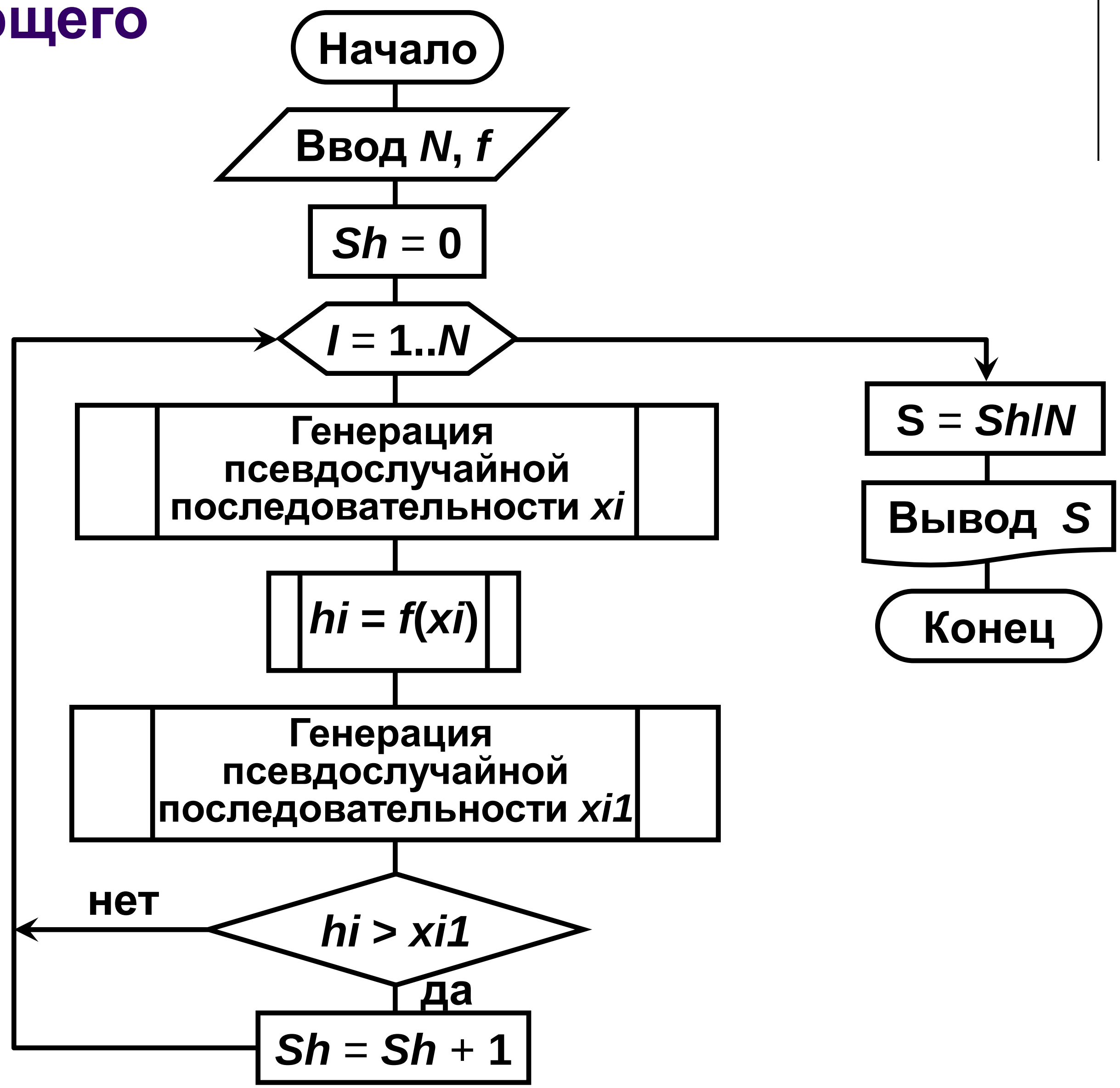
C – суммирование $h' = \sum_{i=1}^{N} h_i$,

Число испытаний, в которых $x_{i+1} \le f(x_i)$

B_2 – вычисление $S' = h'/N$.

Оценка искомой площади

Блок-схема моделирующего алгоритма:



При статистическом моделировании

- большое число операций → большая доля машинного времени расходуются на действия со случайными числами;
- результаты моделирования существенно зависят от качества исходных (базовых) последовательностей случайных чисел.

Наличие простых и экономных способов формирования последовательностей случайных чисел требуемого качества во многом определяет возможность практического использования этого метода.

Доказано:

СВ, обладающие практически любым законом распределения, можно конструировать *с помощью равномерно распределенных на* (0, 1) *случайных чисел*.

Для равномерного в интервале (0, 1) распределения:

$$f(x) = \begin{cases} 1, & \text{при } 0 < x < 1, \\ 0 & \text{при } x \notin (0, 1), \end{cases}$$

$$F(x) = \begin{cases} 0, & \text{при } x \le 0, \\ x, & \text{при } 0 < x < 1, \\ 1, & \text{при } x \ge 1. \end{cases}$$

$$M(X) = \frac{1}{2}, \quad \sigma(X) = \frac{1}{2\sqrt{3}}.$$

$= \sigma_X$

Формирование СВ, квазиравномерно распределенной на (0, 1)

Пусть Z – дискретная СВ, принимающая только два возможных значения:

Z	0	1
p	1/2	1/2

Бесконечную последовательность Z_1, Z_2, … можно рассматривать как двоичные знаки некоторого числа ξ, равного

$$\xi = z_1 \cdot 2^{-1} + z_2 \cdot 2^{-2} + \ldots + z_n \cdot 2^{-n} + \ldots \qquad (*)$$

Ясно, что

$$\xi < 1\cdot 2^{-1} + 1\cdot 2^{-2} + \ \dots \ + 1\cdot 2^{-n} + \ \dots \ = \ 1,$$

т. е. случайное число ξ принадлежит интервалу (0, 1).

При этом:

$$\mathbf{P}(0 < \xi < 1/2) = \mathbf{P}(z_1 = 0) = 1/2;$$

$$\mathbf{P}(0 < \xi < 1/4) = \mathbf{P}(z_1 = 0) \cdot \mathbf{P}(z_2 = 0) = 1/2 \cdot 1/2 = 1/4;$$

....

$$P\left(\xi \in \left(0, \frac{1}{2^n}\right)\right) = P(z_1 = 0)\cdot \ \dots \ \cdot P(z_n = 0) = \frac{1}{2^n}.$$

$$\xi < 1\cdot\frac{1}{2^{n+1}} + 1\cdot\frac{1}{2^{n+2}} + \ \cdots \ = \frac{\frac{1}{2^{n+1}}}{1-\frac{1}{2}} = \frac{1}{2^n}$$

Обобщим:

$$P\left(\xi \in \left(0,\ \frac{k}{2^n}\right)\right) = \frac{1}{2^n}\cdot k\,.$$

Комбинация первых n двоичных знаков в сумме (*) совпадает с одним из k возможных вариантов (соответствуют двоичной записи чисел 0, 1, 2, … , $k-1$)

Тогда

$$P\left(\xi \in \left(\frac{k}{2^n},\ \frac{k+1}{2^n}\right)\right) = P\left(\xi \in \left(0,\ \frac{k+1}{2^n}\right)\right) - P\left(\xi \in \left(0,\ \frac{k}{2^n}\right)\right) =$$

$$= \frac{k+1}{2^n} - \frac{k}{2^n} = \frac{1}{2^n}$$

для любых n, $k = 0, 1, \ldots, n-1$.

Итог:

вероятность попадания ξ в любой интервал вида $\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right)$ совпадает с длиной этого интервала, равной **1/2** ^{n}

ξ – равномерно распределенная СВ.

***Способ формирования равномерно распределенной на (0, 1) СВ*:**

взять бесконечную последовательность независимых СВ $\boldsymbol{z_i}$ и считать их двоичными знаками некоторого числа $\boldsymbol{\xi}$.

На цифровой ВМ для представления числа – только ***k*** двоичных разрядов сумма **(*)** заменяется конечной суммой ***k*** слагаемых.

Количество различных чисел, получаемых таким способом, равно $\mathbf{2}^{k}$.

Вместо непрерывной совокупности случайных чисел с равномерным распределением – дискретная совокупность $\mathbf{2}^{k}$ чисел с одинаковой вероятностью появления любого из них.

Такое распределение называется ***квазиравномерным***.

СВ ξ, имеющая квазиравномерное распределение на (0, 1), принимает значения

$$x_i = \frac{i}{2^k - 1}, \qquad i = 0, 1, 2, \ldots, 2^k - 1$$

2^k–1, а не 2^k, чтобы в число значений x_i можно было включить и 0, и 1, а интервалы между ними были одинаковы

с вероятностями

$$p_k = 1/2^k .$$

Числовые характеристики:

$$M(\xi) = \sum_{i=0}^{2^k-1} \frac{i}{2^k-1}\cdot\frac{1}{2^k} = \frac{1}{2^k\left(2^k-1\right)}\sum_{i=0}^{2^k-1} i =$$

$$= \frac{\left(2^k-1\right)\cdot 2^k}{2\cdot 2^k\left(2^k-1\right)} = \frac{1}{2},$$

$$\sum_{i=0}^{n} i = \frac{n(n+1)}{2},$$

$$\sum_{i=0}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$D(\xi) = \sum_{i=0}^{2^k-1}\left(\frac{i}{2^k-1}\right)^2\cdot\frac{1}{2^k} - \frac{1}{4} = \frac{1}{2^k\left(2^k-1\right)^2}\sum_{i=0}^{2^k-1} i^2 - \frac{1}{4} =$$

$$= \frac{1}{2^k\left(2^k-1\right)^2}\cdot\frac{\left(2^k-1\right)\cdot 2^k\left(2^{k+1}-1\right)}{6} - \frac{1}{4} = \frac{1}{12}\cdot\frac{2^k+1}{2^k-1},$$

$$\sigma(\xi) = \frac{1}{2\sqrt{3}}\sqrt{\frac{2^k+1}{2^k-1}}.$$

При $k \to \infty$ $\qquad \sigma(\xi) \to \dfrac{1}{2\sqrt{3}} = \sigma_X \approx 0{,}28868;$

для малых k разница между $\sigma(\xi)$ и σ_X может оказаться существенной.

k	2	3	5	10	15
$\sigma(\xi)$	0,3727	0,3274	0,2979	0,2889	0,2887
$\dfrac{\sigma(\xi)}{\sigma_X}$	1,290	1,140	1,030	1,001	1,000

Основные способы генерации квазиравномерных случайных чисел:

- *аппаратный* (*физический*),
- *алгоритмический* (*программный*),
- *табличный* (*файловый*).

Аппаратный способ

Случайные числа вырабатываются специальной электронной приставкой – *генератором* (*датчиком*) случайных чисел (одно из внешних устройств ЭВМ).

Дополнительные вычислительные операции не требуются; необходима только операция обращения к датчику.

Физический эффект (источник «случайности») в основе таких генераторов –

- шумы в электронных и полупроводниковых приборах,
- явления распада радиоактивных элементов и т. д.

Для получения $\boldsymbol{k}$-разрядного двоичного случайного числа, имеющего квазиравномерный закон распределения, необходимо:

появление каждом из $\boldsymbol{k}$ разрядов числа $\boldsymbol{Z}$, принимающего значения $z_1 = 0$ и $z_2 = 1$ с вероятностями $p_1 = p_2 = 1/2$.

Пример: генератор шума.

На выходе – сигнал, амплитуда которого изменяется во времени случайным образом;

этот сигнал пропускается через ограничитель уровня; с помощью электронного счетчика подсчитывается количество импульсов сигнала, превышающих заданный уровень;

если в момент снятия показания счетчик дает четное число, то в соответствующем разряде – 0, если нечетное число, то 1.

Результаты исследований:
при достаточно большом количестве импульсов вероятности появления 0 и 1 близки к 1/2

Параллельное соединение ***k*** одноразрядных датчиков случайных чисел – ***k***-разрядный датчик.

Должен вырабатывать случайные числа с частотой, соответствующей быстродействию машины.

Недостатки аппаратного способа:

- использование электронных приборов для генерации случайных чисел замедляет процедуру имитационного моделирования;
- электронный прибор активизируется случайным образом ⇒ невозможно по желанию воспроизвести одну и ту же последовательность случайных чисел.

Для отладки имитационной модели часто требуется дублирование одной и той же последовательности

Алгоритмический способ

Псевдослучайные числа генерируются в ВМ по специальным программам.

Такие числа не являются истинно случайными, т. к. могут быть определены заранее.

Программы для генерации случайных чисел также называют *генераторами* (*датчиками*) случайных чисел.

Требования к генератору:

- формируемая последовательность чисел должна иметь заданную статистическую структуру (например, быть последовательностью независимых СВ с квазиравномерным распределением);
- количество машинных операций, затрачиваемых на формирование одного числа, должно быть небольшим.

Наибольшее применение нашли алгоритмы вида

$$X_{i+1} = \Phi(X_i)$$

(рекуррентные соотношения первого порядка), для которых начальное число X_0 и постоянные параметры заданы.

- ***Метод срединных квадратов***

  (исторически – одна из первых процедур).

Пусть X_0 – некоторое k-разрядное двоичное число – считается первым в последовательности псевдослучайных чисел;

возведем его в квадрат и рассмотрим k средних разрядов X_0^2 как новое k-разрядное двоичное число X_1;

возведем в квадрат X_1 и k средних разрядов X_1^2 будем считать числом X_2 и т. д.

Недостатки:

- возможность появления в последовательности $\boldsymbol{X_i}$ повторяющихся групп чисел,
- возможность вырождения процесса (во всех разрядах – нули).

Следствие: несмотря на небольшое количество машинных операций, данная процедура редко используется на практике.

- ***Линейные конгруэнтные датчики.***

Наиболее широко известный алгоритм.

Пусть заданы:

$\boldsymbol{m} > 0$ – модуль,

$\boldsymbol{X_0}$, $0 \le \boldsymbol{X_0} < \boldsymbol{m}$ – начальное значение,

$\boldsymbol{a}$, $\boldsymbol{c}$, $0 \le \boldsymbol{a} < \boldsymbol{m}$, $0 \le \boldsymbol{c} < \boldsymbol{m}$ – параметры.

Построим последовательность неотрицательных целых чисел, не превосходящих $\boldsymbol{m}$:

Остаток от деления на m

$$X_{i+1} = (aX_i + c) \bmod m, \quad i = 0, 1, 2, \dots$$

Линейная конгруэнтная последовательность

Последовательность псевдослучайных чисел, имеющих квазиравномерное на [0, 1] распределение:

$$U_i = X_i/m .$$

В силу детерминированности метода получаются воспроизводимые последовательности

Конгруэнтная последовательность всегда содержит циклы (периоды).

Пример:

$X_0 = 11,\ \ \boldsymbol{a} = 9,\ \ \boldsymbol{c} = 5,\ \ \boldsymbol{m} = 12.$

$X_1 = (11 \cdot 9 + 5)\ \boldsymbol{mod}\ 12 = 8,\quad U_1 = 8/12 = 0{,}66667,$

$X_2 = (8 \cdot 9 + 5)\ \boldsymbol{mod}\ 12 = 5,\quad U_2 = 5/12 = 0{,}41667,$

$X_3 = (5 \cdot 9 + 5)\ \boldsymbol{mod}\ 12 = 2,\quad U_3 = 2/12 = 0{,}16667,$

$X_4 = (2 \cdot 9 + 5)\ \boldsymbol{mod}\ 12 = 11,\quad U_4 = 11/12 = 0{,}91667,$

$X_5 = (11 \cdot 9 + 5)\ \boldsymbol{mod}\ 12 = 8,\quad U_5 = 8/12 = 0{,}66667,$

далее числа повторяются.

Длина периода не может быть больше *m*!

Конкретный выбор параметров $\boldsymbol{X_0}$, $\boldsymbol{a}$, $\boldsymbol{c}$ и $\boldsymbol{m}$ – решающий фактор, определяющий статистические качества генератора случайных чисел и длину цикла полученной последовательности.

Для практических целей неприемлемыми являются значения $\boldsymbol{a} = 1$ и $\boldsymbol{a} = 0$

далее предполагается $\boldsymbol{a} \geq 2$.

Факторы, определяющие выбор модуля:

- ***m*** должно быть достаточно большим;
- значения ($aX_i + c$) ***mod m*** должны вычисляться быстро.

Рекомендация.

Пусть ***w*** – длина машинного слова

$$w = 2^e, \quad e \text{ – разрядность процессора.}$$

Тогда целесообразен выбор

$$m = w \pm 1.$$

Выбор множителя a.

Основное (но не единственное) требование – обеспечение максимальной длины периода.

Теорема (о максимальном периоде линейного конгруэнтного датчика с $c \neq 0$).

Линейная конгруэнтная последовательность, определенная числами m, a, c и X_0, имеет период длиной m тогда и только тогда, когда

1) числа c и m взаимно простые;

Не имеют общих делителей (кроме 1)

2) число $b = a - 1$ кратно p для каждого простого p, являющегося делителем m;
3) число b кратно 4, если m кратно 4.

- ***Мультипликативные линейные конгруэнтные датчики.***

Частный случай линейных конгруэнтных датчиков при $\boldsymbol{c} = 0$. Широко используются на практике.

В этом случае последовательность $\boldsymbol{X_i}$ имеет вид

$$\boldsymbol{X_{i+1}} = (\boldsymbol{aX_i})\ \boldsymbol{mod\ m}\ , \qquad \boldsymbol{i} = 0, 1, 2, \dots$$

Характерно:

- процесс генерации происходит быстрее;
- значение, равное нулю, не может быть получено;
- максимальный период не может быть достигнут (следствие теоремы).

Пусть ***a*** и ***m*** – взаимно простые числа; для некоторого ***λ*** выполняется

$$a^{\lambda} \; \boldsymbol{mod} \; m = 1.$$

Наименьшее значение ***λ***, удовлетворяющее этому условию, называется *порядком числа **a** по модулю **m***.

Пример.

$3^1 \; \boldsymbol{mod} \; 5 = 3,$

$3^2 \; \boldsymbol{mod} \; 5 = 4,$

$3^3 \; \boldsymbol{mod} \; 5 = 2,$

$3^4 \; \boldsymbol{mod} \; 5 = 1.$

порядок числа 3 по модулю 5 равен 4

Все значения $\boldsymbol{a}$, имеющие одинаковый максимально возможный порядок $\boldsymbol{\lambda}(\boldsymbol{m})$, называются *примитивными элементами по модулю **m***.

Для больших значений $\boldsymbol{m} = \boldsymbol{p}^{\boldsymbol{e}}$,

где $\boldsymbol{p}$ – простое число, $\boldsymbol{e}$ – целое,

примитивные элементы должны определяться с помощью компьютерных программ на основании следующей теоремы.

Теорема (о примитивных элементах по модулю p^e).

Для каждого целого e и простого числа p:

число a является примитивным элементом по модулю p^e тогда и только тогда, когда

a нечетно, $p^e = 2$;

$a \bmod 4 = 3$, $p^e = 4$;

$a \bmod 8 = 3, 5, 7$, $p^e = 8$;

$a \bmod 8 = 3, 5$, $p = 2$, $e > 3$;

$a \bmod p \neq 0$, $a^{(p-1)/q} \bmod p \neq 1$, $p > 2$, $e = 1$, q – все простые делители числа $p - 1$;

$a \bmod p \neq 0$, $a^{p-1} \bmod p^2 \neq 1$, $a^{(p-1)/q} \bmod p \neq 1$, $p > 2$, $e > 1$, q – все простые делители числа $p - 1$.

Теорема (о максимальном периоде для мультипликативных линейных конгруэнтных датчиков).

Максимальный период мультипликативного линейного конгруэнтного датчика с параметрами $\boldsymbol{m}$, $\boldsymbol{a}$, $\boldsymbol{c} = 0$, $\boldsymbol{X_0}$ равен $\boldsymbol{\lambda}(\boldsymbol{m})$. Он достигается, если коэффициент $\boldsymbol{a}$ является примитивным элементом по модулю $\boldsymbol{m}$, а числа $\boldsymbol{X_0}$ и $\boldsymbol{m}$ являются взаимно простыми.

Прикладное значение имеют два случая выбора $\boldsymbol{m}$:

1) $\boldsymbol{m} = 2^{\boldsymbol{e}}$;

   в этом случае $\boldsymbol{m}$–1 – наибольшее целое число, представимое в компьютере;

   можно показать, что максимальная длина периода будет равна $\boldsymbol{m}/4$;

2) $\boldsymbol{m} = \boldsymbol{p}$ (простое число);

   может достигаться период, равный $\boldsymbol{m}$–1.

- ***Объединение нескольких мультипликативных линейных конгруэнтных датчиков***.

  Метод объединения дает возможность достигать очень длинных периодов.

  Базируется на двух теоремах.

Теорема (о сумме дискретных случайных величин, одна из которых имеет квазиравномерное распределение).

Пусть $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, … , $\boldsymbol{X_k}$ – независимые случайные величины, которые могут принимать только целочисленные значения;

$\boldsymbol{X_1}$ имеет квазиравномерное распределение вероятностей

$$\boldsymbol{P}(\boldsymbol{X_1} = \boldsymbol{n}) = 1/\boldsymbol{d}, \quad \boldsymbol{n} = 0,1, \dots , \boldsymbol{d} - 1.$$

Тогда случайная величина

$$X = \left(\sum_{j=1}^{k} X_j \right) \bmod d$$

имеет такое же распределение.

Теорема (о периоде семейства датчиков).

Пусть датчик $\boldsymbol{j}$, $\boldsymbol{j} = 1, 2, \ldots, \boldsymbol{k}$, с периодом $\boldsymbol{p_j}$ генерирует последовательность чисел

$$\boldsymbol{x^{(j)}_0, x^{(j)}_1, \ldots, x^{(j)}_{p_j - 1}}.$$

Рассмотрим последовательности

$$\boldsymbol{X_i} = \{ \boldsymbol{x^{(1)}_i, x^{(2)}_i, \ldots, x^{(k)}_i} \}, \quad \boldsymbol{i} = 0, 1, \ldots$$

Их период $\boldsymbol{p}$ равен наименьшему общему кратному чисел $\boldsymbol{p_1, p_2, \ldots, p_k}$.

Если модули отдельных мультипликативных линейных конгруэнтных датчиков $\boldsymbol{m_j}$ являются простыми числами, то $\boldsymbol{p_j} = \boldsymbol{m_j} - 1$, $\boldsymbol{j} = 1, 2, \dots, \boldsymbol{k}$.

$\Downarrow$

числа $\boldsymbol{p_j}$ являются четными.

Поэтому

$$p \le \frac{\prod_{j=1}^{k} (m_j - 1)}{2^{k-1}}. \qquad (**)$$

Равенство достигается, если величины $(\boldsymbol{m_j} - 1)/2$ не имеют общих делителей.

Теорему о сумме дискретных случайных величин можно использовать для построения последовательности случайных чисел с периодом (**):

$$z_i = \left(\sum_{j=1}^{k} (-1)^{j-1} x_i^{(j)} \right) \operatorname{mod}(m_1 - 1),$$

$$z_i \in \{0,\ 1,\ \ldots,\ m_1 - 2\},$$

$$u_i = \begin{cases} \dfrac{z_i}{m_1}, & z_i > 0, \\ \dfrac{(m_1 - 1)}{m_1}, & z_i = 0, \end{cases}$$

$$0 < u_i < 1.$$

Оценка качества сгенерированной последовательности

Достижение максимально возможного периода – не единственная цель при моделировании случайных чисел.

<u>Важно</u>: формируемая последовательность чисел должна быть последовательностью независимых СВ с квазиравномерным распределением.

Последовательность действий в случае использования мультипликативных датчиков:

1) выбор модуля $\boldsymbol{m}$;
2) выбор коэффициента $\boldsymbol{a}$, обеспечивающего максимальный период в соответствии с теоремой о примитивных элементах по модулю $\boldsymbol{p^e}$;
3) программирование датчиков, заданных параметрами $\boldsymbol{a}$, $\boldsymbol{m}$ и $\boldsymbol{c} = 0$;
4) исследование статистической структуры полученных последовательностей чисел с помощью статистических критериев.

Статистические проверки последовательностей псевдослучайных чисел

- Проверка сгенерированной последовательности на согласование с теоретическим законом распределения.

  Необходима при любом методе получения псевдослучайных чисел.

Для этого используются статистические критерии согласия.

Наиболее известный –

критерий χ^2 (*критерий Пирсона*).

Пусть имеется выборка из $\boldsymbol{n}$ независимых наблюдений над случайной величиной $\boldsymbol{X}$.

Составим группированный статистический ряд:

Интервалы	(x_0, x_1)	(x_1, x_2)	…	(x_{k-1}, x_k)
Частоты	n_1	n_2	…	n_k

n_i – число значений СВ $\boldsymbol{X}$, принадлежащих интервалу (x_{i-1}, x_i), $i = 1, 2, \ldots, k$ (эмпирические частоты),

$$\sum_{i=1}^{k} n_i = n.$$

Проверяется гипотеза:

СВ ***X*** имеет заданный («теоретический») закон распределения.

Обозначим:

$\boldsymbol{p_i}$ – вероятность попадания СВ ***X*** в интервал $(\boldsymbol{x_{i-1}}, \boldsymbol{x_i})$, вычисленная в соответствии с теоретическим законом распределения.

$\boldsymbol{np_i}$ – «теоретические» частоты попадания СВ ***X*** в интервал $(\boldsymbol{x_{i-1}}, \boldsymbol{x_i})$.

В соответствии с критерием Пирсона, степень расхождения между n_i и np_i (эмпирическими и теоретическими частотами) оценивается величиной

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

При неограниченном увеличении n закон распределения этой СВ приближается к распределению χ^2 с r степенями свободы,

где r определяется так: k минус число условий (связей), накладываемых на эмпирические частоты.

В число связей входят:

$$\sum_{i=1}^{k} \frac{n_i}{n} = 1,$$

Это условие накладывается всегда

а также (возможно) выражения для оценок параметров теоретического закона распределения, получаемых по данным выборки.

Для корректного применения критерия Пирсона необходимо

- иметь достаточно большое число наблюдений $\boldsymbol{n}$,
- обеспечить выполнение условия $\boldsymbol{n_i} > 5 \div 10$ для всех $\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{k}$.

В случае необходимости можно объединить некоторые интервалы

***Алгоритм проверки*.**

- По данным выборки определить значение $\chi^2_{набл}$.

  Для организации вычислений удобно использовать таблицу

Интервалы	(x_0, x_1)	(x_1, x_2)	…	(x_{k-1}, x_k)	$\chi^2_{набл}$
Эмпирические частоты n_i	n_1	n_2	…	n_k	
Теоретические частоты np_i	np_1	np_2	…	np_k	
$\frac{(n_i - np_i)^2}{np_i}$	$\frac{(n_1 - np_1)^2}{np_1}$	$\frac{(n_2 - np_2)^2}{np_2}$	…	$\frac{(n_k - np_k)^2}{np_k}$	$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

- Выбрать уровень значимости $\boldsymbol{\alpha}$ или доверительную вероятность $1-\boldsymbol{\alpha}$ и определить критическое значение $\chi^2_{кр}$ (по таблице или средствами *MS Excel*).
- Если

$$\chi^2_{набл} < \chi^2_{кр},$$

то считается, что данные наблюдений не противоречат гипотезе о предполагаемом законе распределения;

в противном случае следует отвергнуть гипотезу, как противоречащую данным выборки.

Пример.

При $\boldsymbol{m} = 97$, $\boldsymbol{a} = 29$, $\boldsymbol{X}_0 = 1$ сгенерирована последовательность из 500 псевдослучайных чисел.

Группированный статистический ряд:

(x_{i-1}, x_i)	(0; 0,1)	(0,1; 0,2)	(0,2; 0,3)	(0,3; 0,4)	(0,4; 0,5)	(0,5; 0,6)	(0,6; 0,7)	(0,7; 0,8)	(0,8; 0,9)	(0,9; 1)
n_i	48	54	52	48	51	53	46	52	51	45

Гипотеза:

СВ $\boldsymbol{X}$ имеет равномерное на (0, 1) распределение.

Для равномерного на (0, 1) распределения при любом ***i***, ***i*** = 1, 2, … , 10

$$p_i = P\left(X \in \left(x_{i-1}, x_i\right)\right) = 0{,}1 .$$

При ***n*** = 500 np_i = 50 для всех ***i*** = 1, 2, … , 10.

$$\chi^2_{набл} = \sum_{i=1}^{10} \frac{\left(n_i - np_i\right)^2}{np_i} = 1{,}68 .$$

Выберем уровень значимости $\alpha = 0{,}05$ (доверительная вероятность $1-\alpha = 0{,}95$).

Параметры теоретического распределения не оценивались по данным выборки ➔ количество связей равно 1,

число степеней свободы $r = 10 - 1 = 9$.

$$\chi^2_{кр}(9;\, 0{,}05) = 16{,}919\,.$$

Функция *MS Excel* ХИ2ОБР с аргументами 0,05 и 9

$\chi^2_{набл} < \chi^2_{кр}$

⇓

нет оснований отвергнуть гипотезу о равномерном законе распределения.

Замечание.

Слишком малое значение величины $\chi^2_{набл}$ может свидетельствовать о «неслучайности» рассматриваемой последовательности.

Так, вероятность появления $\chi^2_{набл}$, не превышающего значения, полученного в рассмотренном примере, составляет менее 0,1 (по таблицам критических точек распределения χ^2).

Это означает: в результате испытания такое значение может появиться менее чем в 1% случаев.

- Для выявления неслучайных зависимостей между соседними элементами последовательности разработан *спектральный критерий*.

Пусть мультипликативный линейный датчик генерирует последовательность псевдослучайных чисел $\boldsymbol{U_i}$, $\boldsymbol{i} = 1, 2, \dots$,

Пары ($\boldsymbol{U_i}$, $\boldsymbol{U_{i+1}}$), $\boldsymbol{i} = 1, 2, \dots , \boldsymbol{m}-1$, можно рассматривать как координаты точек плоскости. Существуют семейства прямых, проходящих через эти точки.

Максимальное расстояние между прямыми одного и того же семейства $\boldsymbol{d_2}$ – мера «равномерности» полученной решетки.

Если расстояния между соседними прямыми примерно равны для всех семейств, то полученную решетку можно считать равномерной.

В этом случае

$$d_2 \approx m^{-\frac{1}{2}}.$$

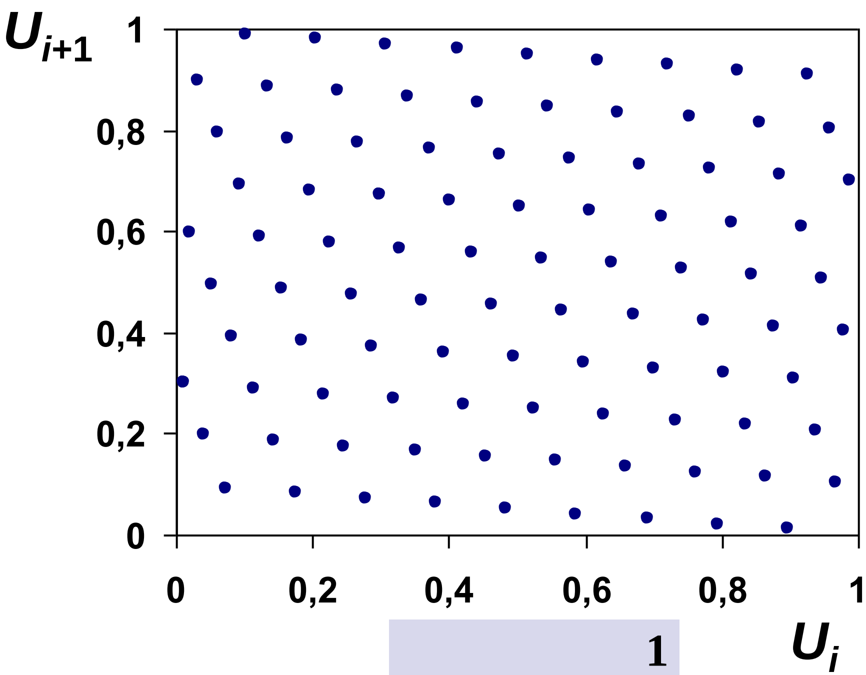
Для решетки с неравномерным распределением

$$d_2 >> m^{-\frac{1}{2}}.$$

Пример.

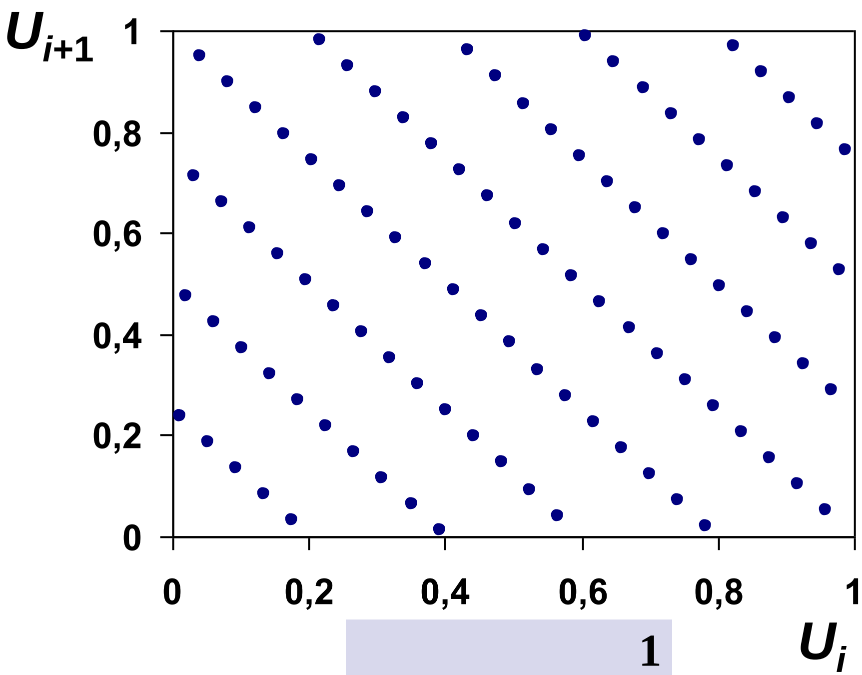
Рассеяние пар точек (U_i, U_{i+1}) при различных значениях m и a.

$m = 97$, $a = 29$



$$d_2 \approx m^{-\frac{1}{2}}$$

$m = 97$, $a = 23$



$$d_2 >> m^{-\frac{1}{2}}$$

Для наборов из t чисел $(U_i, U_{i+1}, \ldots, U_{i+t-1})$ соответствующие точки принадлежат семействам гиперплоскостей.

d_t – наибольшее расстояние между гиперплоскостями.

Для равномерно распределенных точек

$$d_t \approx m^{-\frac{1}{t}},$$

в случае неравномерной решетки

$$d_t >> m^{-\frac{1}{t}}.$$

Показано, что минимально возможный период решетки ограничен снизу:

$$d_t \geq d_t^* = c_t \cdot m^{-\frac{1}{t}},$$

$\boldsymbol{c_t}$ – константа для данного $\boldsymbol{t}$,

$\boldsymbol{c_2} = (4/3)^{-1/4}$, $\boldsymbol{c_3} = 2^{-1/6}$, $\boldsymbol{c_4} = 2^{-1/4}$,
$\boldsymbol{c_5} = 2^{-3/10}$, $\boldsymbol{c_6} = (64/3)^{-1/12}$, $\boldsymbol{c_7} = 2^{-3/7}$, $\boldsymbol{c_8} = 2^{-1/2}$.

Кнут Д. Искусство программирования. Т. 2.

Эффективный компьютерный алгоритм определения значений $d_t(m, a)$ для заданных параметров m и a мультипликативного линейного датчика приведен в книге

Д. Кнут. Искусство программирования. Т. 2.

Применение спектрального критерия:

- для заданных значений ***m*** и ***a*** мультипликативного линейного датчика определяются значения $\boldsymbol{d_t(m, a)}$ для $\boldsymbol{t}$ = 2, 3, … , 6;
- определяется значение

$$S_t(m,a) = \frac{d_t^*(m)}{d_t(m,a)};$$

- датчик считается приемлемым, если величина $\boldsymbol{S_t(m, a)}$ не превышает критического значения.

Примеры приемлемых значений модуля *m* и коэффициента *a* для датчиков, ориентированных на 32-битную целочисленную арифметику.

m	*a*
2 147 483 647	39 373
2 147 483 563	40 014
2 147 483 399	40 692
2 147 482 811	41 546
2 147 482 801	42 024
2 147 482 739	45 742

m – простые числа, близкие к максимальному целому, представимому 32 битами

Результаты исследований П. Лекюйе.

З. Брандт. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров.

Случайные числа с квазиравномерным на (0, 1) распределением – исходный материал для получения случайных объектов более сложной природы:

- случайных величин с произвольным законом распределения;
- многомерных случайных векторов;
- случайных процессов.