1. **Цель работы**

Получить знания в области генерации случайных (псевдослучайных) чисел, изучить свойства линейного конгруэнтного генератора, подобрать правильные параметры генератора и провести проверку получившейся псевдослучайной последовательности.

1. **Постановка задачи**

В ходе выполнения лабораторной работы было необходимо:

1. Выбрать правильные параметры ЛКГ с длиной машинного слова 64 бит.
2. На практике проверить характеристики полученного генератора.
3. Проверить, что младшие биты псевдослучайной последовательности также образуют псевдослучайную последовательность.
4. Проверить полученный генератор методом "хи-квадрат"(*X2)*.
5. Проверить полученный генератор одним из тестов на случайность из книги Д.Э. Кнута "Искусство программирования".
6. **Теоретические исследования**

ЛКГ (линейный конгруэнтный генератор) — это генератор псевдослучайных чисел, построенный на линейном конгруэнтном методе. Суть метода заключается в вычислении последовательности случайных чисел Xn.



Рисунок 1 — Формула Xn+1 элемента последовательности

* *m* — модуль, натуральное число, относительно которого вычисляет остаток от деления; (*m* >= 2); (в ходе выполнения данной работы, часть выражения “mod *m*” была опущена ввиду того, что переменная типа unsigned long long int имеет максимальное значение как раз 264-1, следовательно можно считать, что все числа и так берутся по модулю *m*);
* *a* — множитель (0 <= *a* <= *m*);
* *c* — приращение (0 <= *c* <= *m*);
* X0 — начальное значение (0 <= X0 <= m);

Данная последовательность называется линейной конгруэнтной последовательностью.

Максимальный период последовательности равен *m*, если выполняются следующие условия:

1) Число *a*-1 делится на все простые делители числа m;

2) Если m делится на 4, то и *a*-1 делится на 4;

3) Число *c* взаимно просто с *m*;

Из чего следует:

4) *a* mod 8 = 5

5) *c* – нечётное, простое;

6) *m* = 2^*k* (*k*=16, 32, 64); (в работе значение *k* принято за 64)

Основными характеристиками ЛКГ являются:

* Период – это наименьшее число *k*, при котором для любого возможного *n*;
* Разброс: считается, что генератор имеет хороший разброс, если все числа из сгенерированной псевдослучайной числовой последовательности появляются единожды и с одинаковой вероятностью.
* Мощность — это значение *s*, такое что (*a*-1)*s*mod *m* = 0. Считается что мощность правильного ЛКГ должна быть больше пяти.

Далее в ходе выполнения работы были изучены так называемые “статические гипотезы”. Проверка статистических гипотез является содержанием одного из обширных классов задач математической статистики

Статистическая гипотеза — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки.

Одной из самых известных статистических проверок гипотезы является критерий “хи-квадрат” — любая статистическая проверка гипотезы, в которой выборочное распределение критерия имеет распределение хи-квадрат при условии верности нулевой гипотезы.Проще говоря, необходимо проверить вероятность тестовых данных (данных выборки), чтобы выяснить, является ли наблюдаемое распределение набора данных статистической случайностью (случайно) или нет. Статистика теста хи-квадрат измеряет, насколько хорошо наблюдаемое распределение данных соответствует распределению, которое ожидается, если переменные являются независимыми. Если полученные значения будут достаточно близки к ожидаемым, то результаты нельзя рассматривать как случайные.

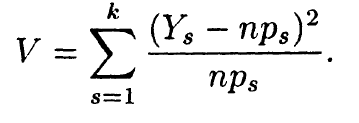


Рисунок 2 – Формула вычисления Хи-квадрата.

Где *Ys* –полученное значение на промежутке, *nps* – ожидаемое число, *k*- количество проверок.

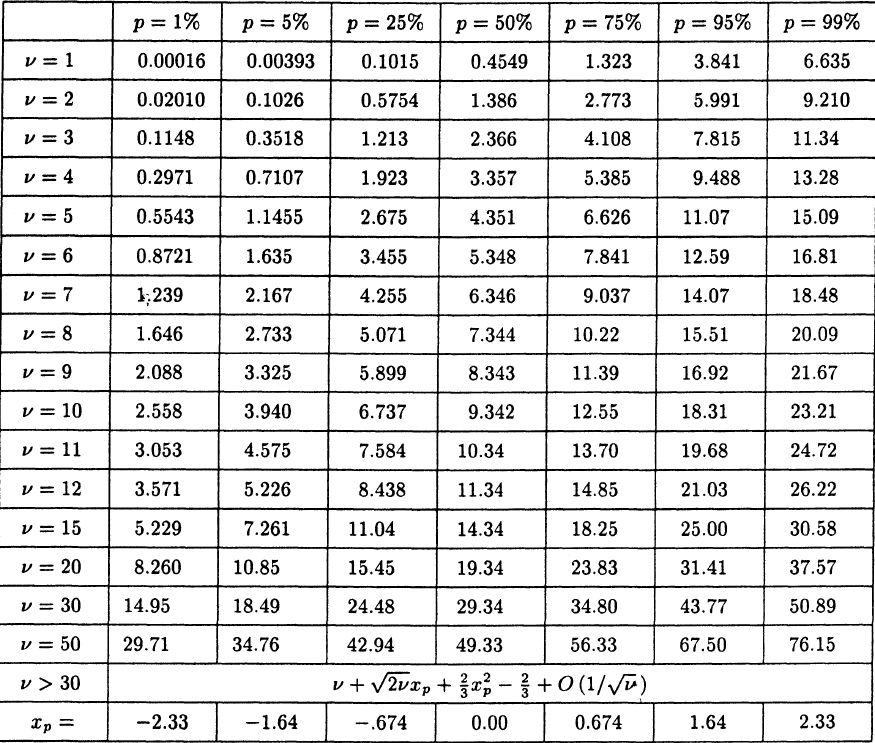


Рисунок 3 — Зависимость P-value от степеней свободы

Количество степеней свободы — это количество значений в итоговом вычислении статистики, способных варьироваться. Иными словами, количество степеней свободы показывает размерность вектора из случайных величин, количество «свободных» величин, необходимых для того, чтобы полностью определить вектор.

В формуле вычисления хи-квадрата наблюдается так называемое “p-значение”(P-value) — вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики (среднего арифметического, медианы и др.), по сравнению с ранее наблюдаемым, при условии, что нулевая гипотеза верна.

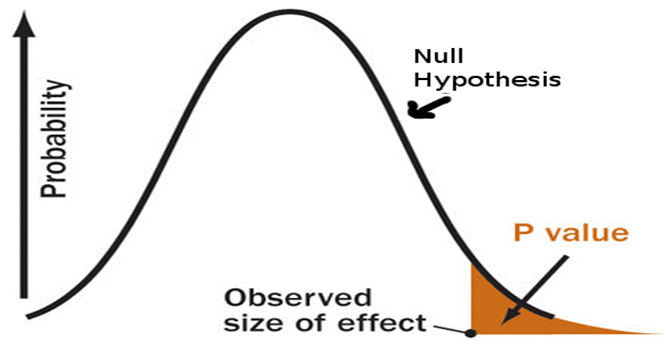


Рисунок 4 — Наглядный вид P-value

Предположим, что существует некое утверждение (нулевая гипотеза), которую мы считаем верной. Далее используется некий пример данных, чтобы проверить, действительно ли она является верной. Если это не так, то мы просто берём альтернативную ей гипотезу.

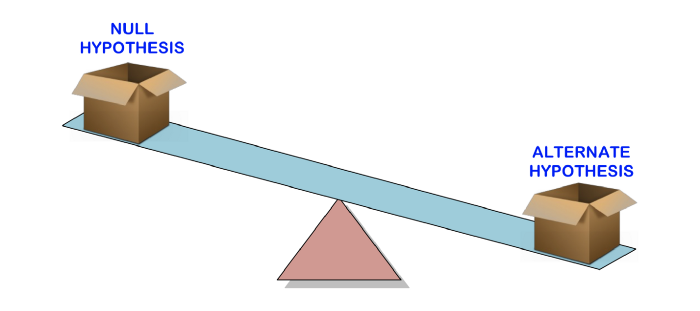


Рисунок 5 — Нулевая и альтернативная гипотезы

Чтобы узнать, является ли утверждение обоснованным или нет, будет использоваться p-значение для взвешивания силы доказательств, чтобы увидеть, является ли оно статистически значимым. Если доказательства подтверждают альтернативную гипотезу, нулевая гипотеза отбрасывается и принимается альтернативная гипотеза.

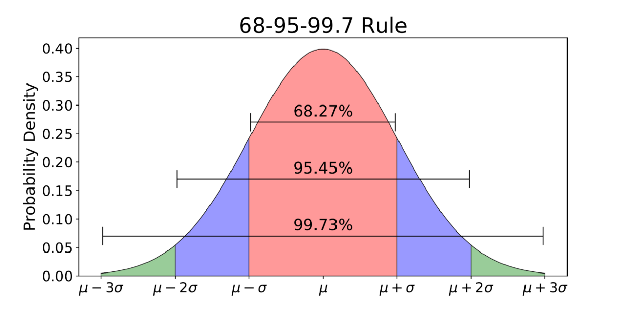


Рисунок 6 — Правило 68-95-99.7

Нормальное распределение — это функция плотности вероятности (функция, характеризующая сравнительную вероятность реализации тех или иных значений случайной переменной), используемая для просмотра распределения данных. Нормальное распределение имеет два параметра — среднее (μ) и стандартное отклонение, также называемое сигма (σ).

Нормальное распределение обычно связано с правилом 68-95-99.7

* 68% данных находятся в пределах 1 стандартного отклонения (σ) от среднего значения (μ)
* 95% данных находятся в пределах 2 стандартных отклонений (σ) от среднего значения (μ)
* 99,7% данных находятся в пределах 3 стандартных отклонений (σ) от среднего значения (μ)

Чем ниже значение p, тем более неожиданными являются доказательства, тем более странной выглядит нулевая гипотеза. Однако, Р-значение ничего не “доказывает”. Это просто способ использовать неожиданность в качестве основы для принятия разумного решения.

P-значения используются в качестве инструмента для оспаривания первоначального убеждения (нулевая гипотеза), когда результат является статистически значимым. В тот момент, когда собственное убеждение оказалось неверным (при условии, что р-значение показывает, что результат статистически значим), первоначальное убеждение отбрасывается (нулевая гипотеза отвергается) и принимается иное (разумное) решение.

Далее, для исследования последовательности на случайность, Дональдом Кнутом были предложены различные тесты, основанные на статистическом критерии *X2*(хи-квадрат). Вычисляемое значение статистики *X2* сравнивается с табличными результатами, и в зависимости от вероятности появления такой статистики делается вывод о её качестве. Среди достоинств этих тестов — небольшое их количество и существование быстрых алгоритмов выполнения. Недостаток — неопределенность в трактовке результатов. Ниже представлено краткое описание некоторых из этих тестов:

* *Проверка несцеплённых серий*. Последовательность разбивается на m непересекающихся серий и строится *X2* для частот появления каждой возможной серии.
* *Проверка корреляции*. Данный тест проверяет взаимонезависимость элементов последовательности.
* *Проверка интервалов.* Данный тест проверяет равномерность распределения символов в исследуемой последовательности, анализируя длины подпоследовательностей, все элементы которых принадлежат определённому числовому интервалу.

И так далее.

В ходе выполнения работы был взят первый, из предложенных тестов — тест несцеплённых серий.

Суть теста заключается в проверке двоичной последовательности, полученной из сгенерированной числовой последовательности. Выбирается число *k* – длина серии. Далее подсчитывается число появлений всевозможных непересекающихся серий длиной *k* (лишние биты при этом отбрасываются) и вычисляется статистика по следующей формуле:

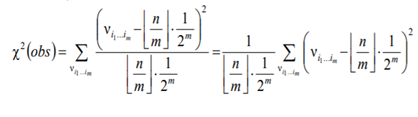


Рисунок 7 – Формула вычисления *Х2(obs)*

1. **Описание решения**

В ходе выполнения работы были взяты следующие табличные параметры:

* m=18446744073709551616; (264)
* a=6364136223846793005;
* c=1442695040888963407;

Как уже было описано ранее, значение *m* не было описано в программе ввиду того, что unsigned long long int имеет максимальное значение 264-1.

Для проверки полученных числовых последовательностей на случайность можно провести проверку младших битов чисел. Каждое число нужно перевести в двоичный вид и подсчитать число нулей и единиц в нем. В, правильно работающем, ЛКГ количество нулей и единиц в числах должны примерно совпадать.

Таблица 1 – Проверка младших битов для табличных параметров.

|  |  |  |  |
| --- | --- | --- | --- |
| Табличные параметры |  |  | |
| Числа | Нули | Единицы | Числа |
| 43766384568270 | 7 | 9 | 43766384568270 |
| 55549807145605 | 9 | 7 | 55549807145605 |
| 50237656849840 | 7 | 9 | 50237656849840 |
| 273959340351 | 6 | 10 | 273959340351 |
| 37804878996066 | 11 | 5 | 37804878996066 |
| 11584240429705 | 10 | 6 | 11584240429705 |
| 62888830366308 | 8 | 8 | 62888830366308 |
| 25304754272483 | 7 | 9 | 25304754272483 |
| 75001373302326 | 7 | 9 | 75001373302326 |

Из показаний таблицы можно заметить, что количество единиц и нулей в числах, при суммировании, примерно совпадает; так же отсутствуют большие последовательности идущих подряд единиц или нулей. При отдельном рассмотрении числа, можно отметить, что лишь одно из девяти чисел имеет количество нулей, более чем 2 раза превышающее количество единиц. Также можно заметить, что младшие биты не имеют закономерностей (постоянных чередований 1010), и, соответственно, младшие биты так же образуют псевдослучайную последовательность.

Характеристики ЛКГ: так как период равен 18446744073709551616, то идеальный разброс означает, что каждое число от 0 до 18446744073709551616 выпадает с одинаковой вероятностью, равной 1/18446744073709551616. В последовательности ЛКГ числа начинают повторяться только с периодом *Xn=Xn+k,* соответственно, можно прийти к выводу, что при *n*=0 и *k*=18446744073709551616 все числа внутри этого промежутка встретятся только один раз с вероятностью 1/18446744073709551616, и при *k*=18446744073709551616 начнется повторение.

P-value для метода *X2* и теста несцеплённых серий:

1. Метод *X2*

Для выбранных параметров получены следующие значения: *X2*=9,158001 и P=0,54884340163. Исходя из того, что P>0,01, можно точно сказать, что последовательность являются случайной на 99%.

1. Тест несцепленных серий

В случае проверки тестом несцепленных серий *X2* = 8,5. Подставляем полученный результат и находим *P* со степенью свободы равной 8. После подсчетов получаем *P*=0,57870370369. Это приемлемый результат (>0,01), который говорит о том, что получившаяся последовательность случайна с вероятностью 99%. В случае с табличными значениями полученный *X2* = 10,44445. Далее таким же образом находится P-value*.* Он равен 0,76647137963 и это так же приемлемый результат.

1. **Тестирование и результаты работы программы**

В результате работы алгоритма были получены требуемые результаты, они представлены в следующей таблице.

Таблица 2 – Наглядный вид полученных значений

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Период | Разброс | P-value (тест методом XI2) | P-value (тест серий) |
| Табличные значения | 2^64 | 1 / (2^64) | 0,54884340163 (> 0,01) | 0, 0,76647137963 (> 0,01) |

В таблице можно увидеть, что ЛКГ с данными параметрами генерирует псевдослучайную последовательность. Однако не стоит забывать, что полученная последовательность никогда не будет случайной, так как зависит от первого элемента.

1. **Выводы**

В результате работы алгоритма были получены требуемые результаты, а именно: ЛКГ сгенерировал последовательность длиной *m* = 2^64 с равновероятным появлением всех чисел из диапазона. Алгоритм прошел проверку младших битов, а также критерия «хи-квадрат» и теста несцепленных серий псевдослучайных последовательностей Дональда Кнута. Во время проведения работы были изучены принципы работы ЛКГ и сделаны вывод о том, что ЛКГ генерирует именно псевдослучайную числовую последовательность, из чего следует, что последовательность не является случайной.